Consistency of nonparametric Bayesian methods for two statistical inverse problems arising from partial differential equations

Luke Kweku William Abraham

Supervisor: Prof. R. Nickl

Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St. John's College August 2019

To Granny and Grandad. Thanks for all the love and faith.
Also to Nanny and BD. I wish you were here to share in this.
You all four were academic pioneers in your families, and I am proud to join your legacy.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared below and specified in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

In Chapter 1 I review known results, from sources cited throughout.

Chapter 2 consists of original work, essentially the content of [1].

Chapter 3 consists of original work, produced in collaboration with Richard Nickl, and is essentially the content of [2].

Kweku Abraham
August 2019

Acknowledgements

Thanks to Richard, for your immense support throughout my PhD, but also for holding back at times to help me develop into an independent researcher.

Thanks to Alberto, for the CAKE seminar so many years ago which set me on the path to my research topic, and to Jakob for pushing me further on the road through my Part III essay. Thanks Andi and Antoni, my most proximal companions in this journey.

Thanks to my research group, and even you Sam, for being friendly faces to bounce ideas off, and making conferences so fun; similarly thanks to Kolyan and the others. Thanks Matthias for always organising, and Sven for making me feel organised myself.

Thanks to my CCA cohort. Mathematically speaking, I am particularly grateful to Fritz and Megan, for helping me through many battles with PDEs, and to Mo Dick, for always humouring my probability questions, even the fifth time I asked about Girsanov's theorem. Socially, thanks especially to Eardi, an ever present lunch companion, and to Lisa, my officemate, and companion through many dark hours when we forgot to put the light on. Thanks to everyone who supported CCA coffee, in particular Sam, for accepting being the butt of all the jokes with only moderate complaint, and Tessa, for her commitment to the coffee cause.

Thanks to my colleagues in outreach, for your commitment to something I care deeply about, and for all the opportunities to run taster sessions.

Thanks to the staff and students at St. John's. Particular thanks to Dr. Dörzzapf, an ever-present since my days as a mere interview candidate. Thanks to my supervisees, for bringing fresh enthusiasm whenever maths was getting me down.

Thanks, of course, to all my friends, so many of whom have been from St. John's. Approaching the end of this section, in something of a rush due to the negative 24 hours left to submit at my planned time, I wish to acknowledge those who I've neglected to explicitly mention, with thanks for your toleration.

Above all, thanks to Mum, Dad, Mel and Musa, and my wider family, on both the Galbraith and the Abraham sides. You have been a constant source of love and support and I am profoundly grateful.

Formal acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis. Two anonymous referees helped improve Chapter 2 with their constructive suggestions for the associated article. Tapio Helin contributed very helpful discussions at the outset of the project on which Chapter 3 is based. Richard Nickl was supported by ERC grant No. 647812.

Abstract

Partial differential equations (PDEs) govern many natural phenomena. When trying to understand the parameters driving these phenomena, we must be aware of the inevitable errors in our measurements; in statistical inverse problems these measurement errors are modelled by statistical noise. One approach to recovering the PDE coefficients governing such statistical inverse problems is through Bayesian methodology. This thesis investigates the theoretical performance of the Bayesian approach in two particular cases.

The first model considered is the *advection-diffusion* equation. Kolmogorov's equations link this partial differential equation to a corresponding (time-homogeneous) *stochastic* differential equation, in which a diffusion process flows according to a 'drift function' and is buffeted by a Brownian motion effect of spatially varying magnitude; this diffusion formulation forms the focus herein. Assuming the diffusion coefficient (the magnitude of the Brownian effect) is given, this thesis considers the problem of recovering the drift function from observations of the diffusion at discrete time intervals.

Chapter 2 gives explicit conditions on priors under which the corresponding Bayesian posteriors provably contract in L^2–distance, as data is collected, around the true drift function, at the frequentist minimax rate (up to logarithmic factors) over periodic Besov smoothness classes. These conditions are verified for some natural nonparametric priors, some of which are shown to adapt to an unknown smoothness parameter. The results are given in the high-frequency regime, where the diffusion is observed to a later time horizon and at ever closer intervals, but in fact the minimax rate (again up to logarithmic factors) is also attained in the low-frequency regime, where the intervals between samples remain fixed. This yields the first drift estimator robust to the sampling regime.

The second model considered is the Calderón problem. This is the mathematical formulation of electrical impedance tomography, in which electrodes are attached to a patient's skin and used to apply voltages and record the corresponding current fluxes. The current flux corresponds to the Neumann data for the solution to a PDE, governed by an interior 'conductivity parameter', in which the voltage gives the Dirichlet boundary values. Varying the applied voltage, we consider observing the 'Dirichlet-to-Neumann map', and attempt to recover the interior conductivity. The data considered in Chapter 3

consists of the Dirichlet-to-Neumann map corrupted by additive Gaussian noise. A prior is exhibited for which the posterior mean statistically converges to the true conductivity (as the noise level is taken to 0) at a near-optimal rate.

The introductory chapter outlines the minimax framework by which the posteriors are judged, and provides the background material relevant to this thesis. Of particular interest may be the included proof, in an general inverse problem setting, of natural conditions under which the consistency of the posterior mean can be guaranteed.

Table of contents

Chapter 1

Introduction

The goal of statistical estimation is to infer the value of a parameter θ governing the distribution of observed data X. In statistical inverse problems, the distribution of X does not depend on θ directly, but only via the value of $G(\theta)$ for some operator G. The *forward map* $\theta \mapsto G(\theta)$ is assumed to be "well-behaved" in some respect (say continuous with respect to some strong metrics, or easy to compute numerically) while its inverse is assumed to be much worse behaved (perhaps discontinuous except with respect to weak metrics, or numerically very unstable). It is generally easy to statistically estimate $G(\theta)$, but estimating the underlying parameter θ is harder because it involves inverting the map G.

Bayesian methods provide a natural statistical approach to inverse problems, as advocated in Stuart [77], but with some conceptual roots tracing further back to Diaconis [24] and even Poincaré [67, Chapter XV §216]. Placing a prior Π on the parameter θ induces a posterior via Bayes' rule:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

The likelihood can be calculated with calls only to the forward map, hence inverting the operator G is not a prerequisite to sampling from the posterior. One hopes that good estimators of θ can be built from these posterior samples, so that the Bayesian method "automatically" solves the inverse problem. The remit of this thesis is proving this hope is well-founded, for two particular statistical inverse problems where the parameter of interest, in each case a *function* so that the models are infinite dimensional ('nonparametric'), can be viewed as the coefficient of some partial differential equation (PDE). The sense in which posterior-based estimators are shown to have good statistical properties is a (frequentist) *minimax* sense, explained in Section 1.3.2.

This introductory chapter provides the core background material for this thesis. First, the two models to be studied are informally introduced.

1.1 Stochastic diffusions

Consider the problem of estimating the coefficients b and σ of the advection-diffusion PDE

$$\begin{aligned} \frac{\partial u}{\partial t} &= b(x)\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2}, \quad x \in \mathbb{R},\ t \in [0,T], \\ u(x,0) &= f(x), \quad x \in \mathbb{R}. \end{aligned} \tag{1.1}$$

Such a PDE arises as the macroscopic behaviour for a system of particles evolving according to a stochastic differential equation (SDE)[1]

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \tag{1.2}$$

where W_t is a standard Brownian motion: the precise relationship is that the solution to (1.1) is given by $u(x,t) = E[f(X_t)|X_0 = x]$ provided f is smooth enough (by the Kolmogorov forward equation or *Fokker–Planck* equations, or see e.g. [26, §7.2-7.3]). Since the observable behaviour of a large number of particles distributed according to some distribution ν is governed by integrals of the form $E_{X\sim\nu}f(X)$ for sufficiently smooth functions f, the PDE indeed captures the evolution of the macroscopic properties of the system.

Observing the solutions to the PDE for all initial data f in some class $\mathcal{F}$ of functions is equivalent to observing the action on $\mathcal{F}$ of the operator P_t, defined by $P_t[f](x) = E[f(X_t)|X_0 = x]$. Given non-noisy access to the semigroup $(P_t)_{t\geq 0}$ it is possible to exactly recover the parameters b and σ, for example via expressions for the eigenfunctions and the invariant density (see Gobet–Hoffmann–Reiss [39]). Rather than observing the semigroup $(P_t)_{t\geq 0}$ governing the law of solutions to the SDE, consider data consisting of discrete samples from the trajectory $(X_t)_{t\geq 0}$ of a single particle obeying the SDE. We assume the coefficient σ is given and attempt to deduce the coefficient b of the PDE, or equivalently of the associated SDE, yielding the statistical inverse problem

[1] The SDE itself governs for example the behaviour of pollen in a river: the pollen drifts with the river (whose velocity at a point x in space is $b(x)$) and is buffeted randomly by the water molecules, leading to the Brownian motion term (of spatially varying noise level $\sigma(x)$, which may be expected to grow as water temperature rises).

Recover b given a sample $(X_{j\Delta})_{j\leq n}$ for X satisfying $\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t$.

Note that while it may be initially surprising that a *single* trajectory of the process X allows any meaningful inference to be made, a periodicity assumption outlined in Chapter 2 will ensure the process is recurrent, so that we gain more information about b near each point $x \in \mathbb{R}$ as we watch for longer.

As well as arising as a measurement model for the PDE inverse problem (1.1), stochastic diffusions are a statistically interesting model in their own right, with uses classically in finance (e.g. [11]) and in life sciences (e.g. [29]), and more recently in evolutionary biology (e.g. [12]) and in climate science (e.g. [28]), to name a few examples, and in all these applications the coefficient b (the *drift function*) is of prime interest to estimate, as the trend we would see in the absence of any random effects.

The main contribution of Chapter 2 is to give examples of priors for which the posterior provably estimates the true drift function well in a sense to be made precise.

Bypassing the inverse problem

As advertised in the introductory paragraphs of this thesis, a chief attraction of Bayesian methods is their ability to bypass any need to calculate the inverse operator. In fact, for diffusions it will be seen in Chapter 2 that in the 'high-frequency' setting considered there, wherein we assume the sampling period Δ depends on n, and take asymptotics as $n\Delta_n \to \infty$, $\Delta_n \to 0$, the inverse nature of the problem can be avoided by using tools from stochastic calculus to view drift estimation as a regression problem. Why then do we not revert to frequentist regression methods to estimate b? An attraction in this setup is that Bayesian methods are *robust to the sampling regime*. As well as the high-frequency setting of primary consideration here, 'low-frequency' asymptotics, where $\Delta > 0$ is a fixed constant, are also widely studied. A practitioner is faced with the subjective choice, based on their fixed finite data set, as to which regime's methods are appropriate: if the time horizon $n\Delta_n$ is large and the sampling period Δ_n is small, estimators designed for high-frequency are appropriate, while if there are many data points, thinly spaced enough that the increments do not contain much local information about the drift function, more conservative low-frequency methods are appropriate. Making the wrong choice yields undesirable consequences: typical high-frequency estimators (e.g. as in Comte et al. [21]) are biased if used on low-frequency data, while typical low-frequency estimators are valid for high-frequency data but do not appear to achieve the fastest possible rate of convergence (see the discussion in Chorowski [18]).

The Bayesian method circumvents this subjective choice of asymptotics by allowing a uniform approach to the two regimes, as shown in Chapter 2.

1.2 The Calderón problem

For ∇ denoting the usual gradient operator and $\nabla\cdot$ the usual divergence operator, the *Calderón problem* consists of calculating the coefficient γ of the PDE

$$\begin{aligned} \nabla\cdot(\gamma\,\nabla\,u) &= 0 \quad \text{in } D, \\ u &= f \quad \text{on } \partial D, \end{aligned} \tag{1.3}$$

from observations of the *Neumann data*

$$\gamma\frac{\partial u_{\gamma,f}}{\partial\nu}\bigg|_{\partial D},$$

for solutions $u = u_{\gamma,f}$ corresponding to each f in some class $\mathcal{F}$ of functions. This inverse problem arises for example in medical imaging, as the mathematical formulation of electrical impedance tomography (EIT) – for example see [55]. In the context of EIT, the coefficient γ is the *conductivity*. Since different medical tissues exhibit vastly different conductivities, finding γ is enough to build a good 3-dimensional image of the patient's internals. Notice that the application of a voltage f and the measurements of the current flux $\gamma\frac{\partial u_{\gamma,f}}{\partial\nu}$ are done only at the boundary ∂D (the domain D demarcates the patient's body), so that this is a non-invasive imaging technique. Moreover, a single set of electrodes suffice to apply the voltages and measure the currents, making the process very cheap compared to other imaging techniques. The most realistic representation of EIT would take the function class $\mathcal{F}$ to consist of a finite collection of indicator functions $\{\mathbb{1}_{I_j}\}_{j\leq n}$ for sets I_j corresponding to the locations of electrodes, and measure the current flux only in these regions also.

Fig. 1.1 A 10-day old baby undergoing an EIT scan. The white rectangles are electrodes.

Figure by Inéz Frerichs, in S. Heinrich, H. Schiffmann, A. Frerichs, A. Klockgether-Radke, I. Frerichs, 'Body and head position effects on regional lung ventilation in infants: an electrical impedance tomography study.' Intensive Care Med., 32:1392-1398, 2006.

The Calderón problem, as well as providing the mathematical formulation of EIT, also describes electrical resistance tomography as used in geophysics, for example for investigating the state of fault lines as in [88]. Indeed, the noiseless formulation outlined above was originally described by Alberto Calderón in the context of oil prospecting [13]. The numerous applications of the problem have seen it widely researched within the inverse problems literature: see the many references in Section 3.0.

Define the *Dirichlet-to-Neumann map* Λ_γ as the operator taking a voltage profile f to the associated current profile $\gamma\frac{\partial u_{\gamma,f}}{\partial \nu}$. (In what follows, one can also consider applying a current profile and measuring the corresponding voltages, as is sometimes more practical in medical contexts; in the mathematical formulation this corresponds to substituting the Neumann-to-Dirichlet map for the Dirichlet-to-Neumann map and does not fundamentally change the problem.) The map Λ_γ takes one weak derivative, as proved in Chapter 3, so if we take $\mathcal{F}$ to be an L^2–Sobolev space $H^r(\partial D)$ for some $r \in \mathbb{R}$ (see Section 1.6 for the definitions of weak derivatives and Sobolev spaces), we arrive at the following statistical version of the Calderón problem:

Recover γ given a noisy observation of the map $\Lambda_\gamma : H^r(\partial D) \to H^{r-1}(\partial D)$.

The noise we consider is Gaussian white noise indexed by an appropriate space of operators. This strikes a balance between realism and tractability, as explained in Section 3.1 and Appendix 3.D, where it is argued that this Gaussian white noise model is statistically close to a model more realistically representing the electrode measurements. The notion of distance providing the sense in which two models can be statistically close is the Le Cam discrepancy, described in Section 1.3.1.

The results in Chapter 3 show that the posterior mean is a consistent estimator (as the noise level tends to zero) of the conductivity γ in this Gaussian white noise model.

1.3 Statistical decision theory: a brief overiew

The main question addressed by this thesis can be phrased thus: Under the frequentist assumption of a fixed data-generating parameter θ_0, how well do Bayesian posteriors estimate θ_0 in the above two models? To answer it we must settle on an appropriate notion of what it means to estimate a parameter "well". The notion used here is given by the *minimax* paradigm described in Section 1.3.2. First, we pin down the precise meaning of a *model*, and give a notion of equivalence for different models.

1.3.1 Le Cam equivalence

The concepts throughout this section are drawn from Le Cam's 1986 monograph [51]. The expository paper of Mariucci [57] gives a gentler introduction to the area.

Definitions. Statistical model/experiment A *statistical model* or (more formally) *experiment*, consists of a family of data generating processes P_θ (probability measures on some measurable space $(\mathcal{X}, \mathcal{F})$) indexed by the parameter $\theta \in \Theta$ for some set Θ. Formally, an experiment is therefore the triple $(\mathcal{X}, \mathcal{F}, \{P_\theta\}_{\theta\in\Theta})$.

Markov Kernel For measurable spaces $(\mathcal{X}_1, \mathcal{F}_1), (\mathcal{X}_2, \mathcal{F}_2)$, a *Markov kernel with source* $(\mathcal{X}_1, \mathcal{F}_1)$ *and target* $(\mathcal{X}_2, \mathcal{F}_2)$ is a map $T : \mathcal{X}_1 \times \mathcal{F}_2 \to [0, 1]$ such that $T(x, \cdot)$ is a probability measure for each $x \in \mathcal{X}_1$, and $T(\cdot, A)$ is measurable for each $A \in \mathcal{F}_2$.

Le Cam discrepancy The *Le Cam discrepancy* between experiments $\mathcal{E}_1$ and $\mathcal{E}_2$, where $\mathcal{E}_i = (\mathcal{X}_i, \mathcal{F}_i, \{P_{i,\theta}\}_{\theta\in\Theta})$ for $i = 1, 2$, for a common parameter space Θ, is

$$\delta(\mathcal{E}_1, \mathcal{E}_2) = \inf_T \sup_{\theta\in\Theta} \|TP_{1,\theta} - P_{2,\theta}\|_{\mathrm{TV}},$$

where the infimum is over all Markov kernels with source $(\mathcal{X}_1, \mathcal{F}_1)$ and target $(\mathcal{X}_2, \mathcal{F}_2)$. The measure $TP_{1,\theta}$ is defined as

$$TP_{1,\theta}(A) = \int_{\mathcal{X}_1} T(x, A)\, \mathrm{d}P_{1,\theta}(x),$$

and $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm on signed measures,

$$\|\nu\|_{\mathrm{TV}} = \sup_A |\nu(A)|.$$

The Le Cam discrepancy satisfies the triangle inequality, but is not symmetric.

Le Cam distance The *Le Cam distance* between experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ on a common parameter space Θ is

$$\Delta(\mathcal{E}_1, \mathcal{E}_2) = \max(\delta(\mathcal{E}_1, \mathcal{E}_2), \delta(\mathcal{E}_2, \mathcal{E}_1)).$$

We say two experiments are *(statistically) equivalent* if the Le Cam distance between them is zero, and that they are *asymptotically equivalent* if the measures $P_{i,\theta}$, $i = 1, 2$ also depend on a parameter n and $\Delta(\mathcal{E}_1^n, \mathcal{E}_2^n) \to 0$ as $n \to \infty$.

If we identify experiments whose Le Cam distance is zero, Δ is a proper metric.

Kullback–Leibler divergence In the above definitions the total variation distance is used as the basic tool to define distances between probability measures. Another useful notion of the discrepancy between two probability measures is the *Kullback–Leibler divergence* K, defined for distributions P, Q as

$$K(P,Q) = E_{X\sim P} \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(X)\Big),$$

where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon–Nikodym derivative (= likelihood ratio) between the distributions. If P is not absolutely continuous with respect to Q, so that $\frac{\mathrm{d}P}{\mathrm{d}Q}$ fails to exist, by convention $K(P,Q)$ is defined to equal ∞.

Remark. Given any 'action set' A, any bounded loss function $L : \Theta \times A \to [0,1]$, and any decision rule $\rho_2 : \mathcal{X}_2 \to A$, there exists a decision rule $\rho_1 : \mathcal{X}_1 \to A$ (allowed to depend possibly also on some external randomness) such that, denoting the risk functions by $R_j(\rho_j, \theta) = E_{X\sim P_{j,\theta}} L(\theta, \rho_j(X))$, $j = 1, 2$, we have

$$R_1(\rho_1, \theta) \le R_2(\rho_2, \theta) + \delta(\mathcal{E}_1, \mathcal{E}_2), \quad \forall \theta \in \Theta.$$

This captures the intuitive definition that the Le Cam discrepancy is the worst-case error we incur when reconstructing a decision rule in $\mathcal{E}_2$ using data from $\mathcal{E}_1$. See [57] Theorem 2.7.

The next lemma gathers some key tools used to control Le Cam discrepancies.

Lemma 1.1. *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be experiments with a common parameter set Θ: write $\mathcal{E}_j = (\mathcal{X}_j, \mathcal{F}_j, \{P_{j,\theta}\}_{\theta\in\Theta})$.*

a. Suppose further that the experiments are defined on a common probability space, i.e. that $\mathcal{X}_1 = \mathcal{X}_2$ and $\mathcal{F}_1 = \mathcal{F}_2$. Then

$$\Delta(\mathcal{E}_1, \mathcal{E}_2) \le \sup_{\theta\in\Theta} \|P_{1,\theta} - P_{2,\theta}\|_{\mathrm{TV}} \le \sup_{\theta\in\Theta} \sqrt{K(P_{1,\theta}, P_{2,\theta})/2}. \tag{1.4}$$

b. Let $F : \mathcal{X}_1 \to \mathcal{X}_2$ be any (deterministic) measurable map. Then

$$\delta(\mathcal{E}_1, \mathcal{E}_2) \le \sup_{\theta\in\Theta} \|P_{1,\theta} \circ F^{-1} - P_{2,\theta}\|_{\mathrm{TV}}. \tag{1.5}$$

c. Let $F : \mathcal{X}_1 \to \mathcal{X}_2$ be a measurable map. Suppose that $P_{1,\theta} \circ F^{-1} = P_{2,\theta}$ for each $\theta \in \Theta$ and suppose that $F(X)$ is a sufficient statistic for $X \sim P_{1,\theta}$. Then $\Delta(\mathcal{E}_1, \mathcal{E}_2) = 0$.

Proof. Given a (deterministic) measurable function $F : \mathcal{X}_1 \to \mathcal{X}_2$, denote by T_F the Markov kernel $T_F(x, A) = \mathbb{1}\{F(x) \in A\}$.

a. The first inequality is immediate from the definition, since the Markov kernel T_{Id} corresponding to the identity map $\mathrm{Id} : \mathcal{X}_1 \to \mathcal{X}_2 = \mathcal{X}_1$ satisfies $T_{\mathrm{Id}}P = P$ for all probability measures P on $(\mathcal{X}_1, \mathcal{F}_1)$. The second inequality is Pinsker's inequality (e.g. Proposition 6.1.7a in [38]).

b. Observe that $T_F P_{1,\theta}(A) = P_{1,\theta}(F(X) \in A) = P_{1,\theta} \circ F^{-1}(A)$. The result follows.

c. See [57], Property 3.12. □

Remark. Informally, Lemma 1.1b says that given any data X from an experiment $\mathcal{E}_1$ and any function F, the experiment $\mathcal{E}_2$ with data $Y = F(X)$ satisfies $\delta(\mathcal{E}_1, \mathcal{E}_2) = 0$. If F is bijective, the two experiments are equivalent. This gives a sense in which the discrepancy from $\mathcal{E}_1$ to $\mathcal{E}_2$ measures whether data from $\mathcal{E}_2$ contains all the information that would be available in a sample of data from $\mathcal{E}_1$.

Example. Denote by $\mathcal{B}, \mathcal{B}^n$ the Borel σ-algebras on $\mathbb{R}, \mathbb{R}^n$ respectively. Define

$$\begin{aligned}
\mathcal{E}_1 &= (\mathbb{R}^n, \mathcal{B}^n, \{P_{1,\theta}\}_{\theta\in\mathbb{R}}), && \text{with data } X = (X_1, \dots, X_n) \overset{iid}{\sim} N(\theta, 1) \text{ under } P_{1,\theta},\\
\mathcal{E}_2 &= (\mathbb{R}, \mathcal{B}, \{P_{2,\theta}\}_{\theta\in\mathbb{R}}), && \text{with data point } Y \sim N(\theta, 1/n) \text{ under } P_{2,\theta}.
\end{aligned}$$

Then $\Delta(\mathcal{E}_1, \mathcal{E}_2) = 0$ by Lemma 1.1c with $F((X_1, \dots, X_n)) = \frac{1}{n}\sum X_i$. An alternative direct argument uses Lemma 1.1b to show $\delta(\mathcal{E}_1, \mathcal{E}_2) = 0$, and in the other direction $\delta(\mathcal{E}_2, \mathcal{E}_1) = 0$ is witnessed by the Markov kernel $T : \mathbb{R} \times \mathcal{B}^n$ defined by $T(y, A) = \Pr(y + \varepsilon \in A)$, where $\varepsilon = (\varepsilon_i)_{i\le n}$ with $\varepsilon_i \overset{iid}{\sim} N(0, (n-1)/n)$ independently of Y, and addition $y + \varepsilon$ is defined pointwise. Note that $T(\cdot, A)$ is continuous for each $A \in \mathcal{B}^n$, hence is measurable, and observe that TP_θ is the law of $Y + \varepsilon =^d X$.

See Appendix 3.D for more examples of (asymptotic) equivalence results.

1.3.2 Minimax decision theory

Minimax decision theory gives an optimality criterion for statistical estimators based on their worst-case performance. Given the statistical model $(\mathcal{X}, \mathcal{F}, \{P_\theta\}_{\theta\in\Theta})$ we suppose throughout that Θ is a metric space, equipped with a metric d and with the induced Borel σ–algebra.

Definition (Estimators, tests)**.** An *estimator* of the parameter θ is any measurable map $\hat{\theta} : \mathcal{X} \to \Theta$.

More generally, for any function of the parameter (for example $G(\theta)$ in an inverse problem with forward map G) we define an estimator as a measurable map from $\mathcal{X}$ to the codomain of the function.

A *test* is a $\{0,1\}$–valued measurable function of the data, interpreted as a rule for choosing between competing hypotheses H_0 and H_1. Note that, by composing with the indicator function $\mathbb{1}_A$ of a measurable set $A \subset \Theta$, we can construct tests from any given estimator $\hat{\theta}$.

Remark. Typically, the measures $\{P_\theta\}_{\theta\in\Theta}$ also include a parameter governing the sample size or noise level, and we seek estimators which perform well in the large dataset/small noise setting. For simplicity of exposition, throughout this introductory section we will always take this parameter to be $n \to \infty$, hence will consider data $X \sim P_\theta^n$ (with corresponding expectation and variance operators E_θ^n, Var_θ^n); in the EIT setting we instead have a noise level $\varepsilon \to 0$ and in the diffusion setting the time horizon $n\Delta$ largely takes on the role of sample size. Estimators are allowed to depend on this parameter n; this dependence is often left implicit.

Definition. An estimator $\hat{\theta}$ is *minimax optimal* for estimating a parameter θ in the metric d if

$$\sup_{\theta\in\Theta} E_\theta^n(d(\hat{\theta},\theta)) = \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} E_\theta^n(d(\tilde{\theta},\theta)),$$

where the infimum is over all estimators $\tilde{\theta}$. The quantity on the right in the above is called the *minimax rate*. An estimator $\hat{\theta}$ such that

$$\sup_{\theta\in\Theta} E_\theta^n(d(\hat{\theta},\theta)) \leq C \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} E_\theta^n(d(\tilde{\theta},\theta)) \quad \forall n,$$

for some constant C will be described as *rate optimal*, or typically still just minimax optimal.

In this thesis, an estimator $\hat{\theta}$ will also be described as minimax optimal, and ε_n described as the minimax rate, if there exist constants c, C such that

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} P_\theta^n(d(\hat{\theta},\theta) > C\varepsilon_n) = 0 \tag{1.6}$$

$$\liminf_{n\to\infty} \inf_{\tilde{\theta}} \sup_{\theta\in\Theta} P_\theta^n(d(\tilde{\theta},\theta) > c\varepsilon_n) > 0. \tag{1.7}$$

This latter usage is not universally standard, but note the following partial equivalence results: if (1.7) holds, then ε_n lower bounds the usual minimax rate (up to a constant factor) by Markov's inequality, while if $\hat{\theta}$ is minimax in the usual sense with rate ε_n, then

(1.6) holds for rate $\varepsilon_n' = C_n \varepsilon_n$ for a sequence C_n tending to infinity arbitrarily slowly, again by Markov's inequality.

Why assess estimators according to their worst-case performance? Given we do not know the true parameter, it is comforting to know an estimator will perform well independent of the truth, provided the base assumption $\theta \in \Theta$ holds. The following examples also illustrate pitfalls of some other criteria for judging estimators.

Examples. 1. Consider the constant estimator $\hat{\theta} = \hat{\theta}(X) = 0$. This has "perfect" performance if θ happens to equal zero, and in all other cases it is useless. Thus, judging an estimator by its best case performance yields meaningless guarantees.

2. The *Cramér–Rao* lower bound says that the variance of any unbiased estimator of θ (i.e. of any $\hat{\theta}$ such that $E_\theta \hat{\theta} = \theta$ for all $\theta \in \Theta$) is at least the inverse of the *Fisher information* (the variance of the θ–derivative of the log likelihood). In the model $X_i \overset{iid}{\sim} N(\theta, 1)$, this says that $\mathrm{Var}_\theta^n(\hat{\theta}_n) \geq n^{-1}$ for any unbiased estimator $\hat{\theta}_n$ built from n observations $X_1, \ldots, X_n$. Consider *Hodges' estimator* for this model: define

$$\hat{\theta}_n = \begin{cases} \bar{X}_n := \frac{1}{n} \sum X_i & \text{if } |\bar{X}_n| \geq n^{-1/4} \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that $\hat{\theta}_n$ is asymptotically unbiased (i.e. $E_\theta^n \hat{\theta}_n \to \theta \ \forall \theta \in \Theta$) and the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is $N(0,1)$ if $\theta \neq 0$ or $N(0,0)$ if $\theta = 0$.[2] So Hodges' estimator is 'superefficient': at each fixed $\theta \in \Theta$ its variance asymptotically matches or outperforms the Fisher information lower bound suggested by the finite sample Cramér–Rao theorem.

However, few would argue that Hodges' estimator is better than the sample mean $\bar{X}_n$ in view of the fact that the asymptotic risk (rescaled by $\sqrt{n}$) of $\hat{\theta}_n$ at a sequence $\theta_n = n^{-1/4}$ is infinite (for $\bar{X}_n$ the asymptotic risk rescaled by $\sqrt{n}$ remains finite). The uniformity in performance demanded by the minimax criterion rules out the type of pointwise but non-uniform convergence displayed by the risk function of Hodges' estimator.

[2]Proof: We have $\sqrt{n}(\bar{X}_n - \theta) =^d N(0,1)$, and $\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(\hat{\theta}_n - \bar{X}_n) + \sqrt{n}(\bar{X}_n - \theta)$. Note $P_\theta(\hat{\theta}_n \neq \bar{X}_n) = P_\theta(|\bar{X}_n| \leq n^{-1/4})$. For $\theta \neq 0$, choose n large enough that $|\theta| > 2n^{-1/4}$ and apply the triangle inequality to see

$$P_\theta(\hat{\theta}_n \neq \bar{X}_n) \leq P_\theta(n^{1/2}|\bar{X}_n - \theta| > n^{1/4}) \to 0,$$

so that $\sqrt{n}(\hat{\theta}_n - \bar{X}_n) \to^p 0$. The desired result follows by Slutsky's lemma (a special case of the continuous mapping theorem; e.g. see [84, Lemma 2.8]). When $\theta = 0$, similar calculations show $P_\theta(\hat{\theta}_n = 0) \to 1$, so the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is a Dirac mass at 0.

Lower bounds on minimax rates are typically proved using corresponding information-theoretic lower bounds in hypothesis testing, such as the following result for a data model $X^{(n)} \sim P_\theta^n$, assumed to be dominated so that P_θ^n has a density $p_\theta^{(n)}$ with respect to some reference measure. Recalling $K(P, Q)$ denotes the Kullback–Leibler divergence between distributions P and Q, we, in a slight abuse of notation, also write $K(p, q)$ for the Kullback–Leibler divergence between distributions with densities p and q.

Theorem 1.2. *There exists a pair of positive constants (c, μ) (we may take $c = 1/7$, $\mu = 1/250$) such that if for all $n > N$, some $N \in \mathbb{N}$, there are parameters $\theta_0, \theta_1 \in \Theta$ (both allowed to depend on n) which satisfy*

(1) $d(\theta_0, \theta_1) \geq \varepsilon_n$,

(2) $K(p_{\theta_1}^{(n)}, p_{\theta_0}^{(n)}) \leq \mu$,

then the minimax rate is lower bounded by ε_n in the precise sense that

$$\inf_{\tilde\theta} \sup_{\theta \in \Theta} P(d(\tilde\theta, \theta) > \varepsilon_n) \geq c, \quad \textit{and} \tag{1.8}$$

$$\inf_{\tilde\theta} \sup_{\theta} E d(\tilde\theta, \theta) \geq c\varepsilon_n. \tag{1.9}$$

Proof. The (standard) proof, as follows, can be found for example in Tsybakov [80, Chapter 2] or Giné & Nickl [38, Theorem 6.3.2].

Under condition (1), noting that $\mathbb{1}\{d(\tilde\theta, \theta_1) < d(\tilde\theta, \theta_0)\}$ yields a test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, we see

$$\inf_{\tilde\theta} \sup_{\theta \in \Theta} P_\theta^n\Big(d(\tilde\theta, \theta) \geq \tfrac{1}{2}\varepsilon_n\Big) \geq \inf_{\psi} \max(P_{\theta_0}^n(\psi \neq 0), P_{\theta_1}^n(\psi \neq 1)),$$

where the latter infimum is over all tests ψ. Introducing the event $A = \Big\{\frac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} \geq 1/2\Big\}$, we see

$$P_{\theta_0}^n(\psi \neq 0) \geq E_{\theta_1}^n\Big[\tfrac{p_{\theta_0}^{(n)}}{p_{\theta_1}^{(n)}} \mathbb{1}_A \psi\Big] \geq \tfrac{1}{2}[P_{\theta_1}^n(\psi = 1) - P_{\theta_1}^n(A^c)]$$

Thus, writing $p_1 = P_{\theta_1}^n(\psi = 1)$, we see

$$\begin{aligned}\max(P_{\theta_0}^n(\psi \neq 0), P_{\theta_1}^n(\psi \neq 1)) &\geq \max(\tfrac{1}{2}(p_1 - P_{\theta_1}^n(A^c)), 1 - p_1) \\ &\geq \inf_{p \in [0,1]} \max(\tfrac{1}{2}(p - P_{\theta_1}^n(A^c)), 1 - p).\end{aligned}$$

The infimum is attained when $\frac{1}{2}(p - P^n_{\theta_1}(A^c)) = 1 - p$ and takes the value $\frac{1}{3}P^n_{\theta_1}(A)$ so that

$$\inf_{\tilde{\theta}} \sup_{\theta \in \Theta} P^n_\theta\Big(d(\tilde{\theta}, \theta) \geq \tfrac{1}{2}\varepsilon_n\Big) \geq \tfrac{1}{6}P^n_{\theta_1}(A). \tag{1.10}$$

Next observe

$$\begin{aligned} P^n_{\theta_1}(A) &= P^n_{\theta_1}\Big[\tfrac{p^{(n)}_{\theta_1}}{p^{(n)}_{\theta_0}} \leq 2\Big] = 1 - P^n_{\theta_1}\Big[\log\Big(\tfrac{p^{(n)}_{\theta_1}}{p^{(n)}_{\theta_0}}\Big) > \log 2\Big] \geq 1 - P^n_{\theta_1}\Big[|\log(\tfrac{p^{(n)}_{\theta_1}}{p^{(n)}_{\theta_0}})| > \log 2\Big] \\ &\geq 1 - (\log 2)^{-1} E^n_{\theta_1}\Big|\log\Big(\tfrac{p^{(n)}_{\theta_1}}{p^{(n)}_{\theta_0}}\Big)\Big|, \end{aligned}$$

where we have used Markov's inequality to attain the final expression. By the second Pinsker inequality (Proposition 6.1.7b in [38]), using condition (2) we can continue the chain of inequalities to see

$$P^n_{\theta_1}(A) \geq 1 - (\log 2)^{-1}\Big[K(p^{(n)}_{\theta_1}, p^{(n)}_{\theta_0}) + \sqrt{2K(p^{(n)}_{\theta_1}, p^{(n)}_{\theta_0})}\Big] \geq 1 - (\log 2)^{-1}(\mu + \sqrt{2\mu}).$$

For any $c < 1/6$, we may choose $\mu = \mu(c)$ small enough that (1.10) is lower bounded by

$$\tfrac{1}{6}\Big(1 - \frac{\mu + \sqrt{2\mu}}{\log 2}\Big) > c,$$

and in particular a numerical calculation shows that $\mu < 1/250$ suffices for $c = 1/7$. This yields (1.8), and (1.9) follows by Markov's inequality. □

1.4 General theory for Bayesian inverse problems

In the context of Bayesian techniques, an appropriate notion of the quality of estimation, adapted from the notion of the minimax rate, is the 'contraction rate' of a posterior. In this section we review general techniques for proving contraction rates in Bayesian inverse problems: first, we outline the main contraction rate theorem from the seminal work of Ghosal–Ghosh–van der Vaart [32] for a non-inverse problem, then we give adaptations to this result to allow it to apply to inverse problems.

1.4.1 Direct observations

Recall we assume (Θ, d) is a metric space, equipped with the Borel σ–algebra, and we consider a dominated data model. The *likelihood* is the density, viewed as a function of θ: $L_n(\theta) = L_n(\theta; X^{(n)}) = p^{(n)}_\theta(X^{(n)})$. The object of consideration is the posterior $\Pi(\cdot \mid X^{(n)})$

corresponding to some prior Π via Bayes' rule: under mild assumptions to ensure the expression on the right-hand side is well-defined (e.g. see [34] §1.3) we have

$$\Pi(B \mid X) = \frac{\int_B p_\theta^{(n)}(X^{(n)})\,\mathrm{d}\Pi(\theta)}{\int_\Theta p_\theta^{(n)}(X^{(n)})\,\mathrm{d}\Pi(\theta)}. \tag{1.11}$$

Define Kullback–Leibler type balls

$$B_{KL}^n(\varepsilon) = \left\{\theta \in \Theta \text{ s.t. } K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \le n\varepsilon^2, \mathrm{Var}_{\theta_0}^n\left(\log \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}}\right) \le n\varepsilon^2\right\}. \tag{1.12}$$

A core tool we will use for deducing contraction rates in the two models considered in this thesis is the following abstract result, slightly adapted from Theorem 2.1 of [32]. See also Chapter 8 in the monograph [34] of Ghosal & van der Vaart for a number of results of this flavour.

Theorem 1.3. *Suppose the positive sequence (ε_n) satisfies $\varepsilon_n \to 0$, $n\varepsilon_n^2 \to \infty$. Let Π be a prior (or sequence of priors with suppressed index n) on the metric space (Θ, d) (equipped with the Borel σ–algebra). Suppose there exists $\zeta > 0$ and (measurable) sets $\Theta_n \subset \Theta$ such that*

(i) $\Pi(\Theta_n^c) \le e^{-(2\zeta+8)n\varepsilon_n^2}$,

(ii) There exists an estimator $\hat\theta$ such that for some constant C

$$\sup_{\theta\in\Theta_n} P_\theta^n(d(\hat\theta,\theta) > \tfrac{1}{2}C\varepsilon_n) \le e^{-(2\zeta+8)n\varepsilon_n^2},$$

(iii) $\Pi(B_{KL}^n(\varepsilon_n)) \ge e^{-\zeta n\varepsilon_n^2}$.

Then as $n \to \infty$,

$$P_{\theta_0}^n\left(\Pi(d(\theta,\theta_0) > C\varepsilon_n \mid X^{(n)}) \ge 2e^{-(\zeta+4)n\varepsilon_n^2}\right) \to 0, \tag{1.13}$$

and ε_n is called a posterior contraction rate *for the prior Π. If (iii) holds uniformly for $\theta_0 \in \tilde\Theta$, some $\tilde\Theta \subset \Theta$, then (1.13) also holds with $\sup_{\theta_0\in\tilde\Theta}$ in front.*

Remark (Optimality). Contraction at rate ε_n guarantees the existence of an estimator converging to the true parameter at rate ε_n as follows. If we allow *randomised* estimators (i.e. $\tilde\theta(X^{(n)})$ is a random variable even once $X^{(n)} = x^{(n)}$ is specified) then (1.13) immediately implies that $\tilde\theta_n$ corresponding to a single draw from the posterior distribution will

achieve $P^n_{\theta_0}(d(\tilde{\theta}_n, \theta_0) > C\varepsilon_n) \to 0$. Restricting to "proper" (nonrandomised) estimators, we can argue as in Theorem 8.7 of [34].[3]

It follows that no rate ε_n faster than the minimax rate can be achieved uniformly in Theorem 1.3. Typically series expansion or Gaussian priors can achieve a rate ε_n equalling the minimax rate up to a log factor (as in Chapters 2 and 3), and so Theorem 1.3 is near optimal, both in the sense that the posterior cannot (uniformly) contract any faster, and in the sense that an estimator achieving the frequentist minimax rate exists.

Remark (Necessity of the conditions). Even in the simplest parametric models a condition along the lines of (iii) is needed: if the prior puts no mass on a neighbourhood of the true θ_0, the posterior will put no mass there too. Any positive mass around the true θ_0 is enough to achieve posterior *consistency*, wherein the mass of arbitrary but fixed open neighbourhoods $U \ni \theta_0$ tends to 1, but to achieve a contraction *rate* a lower bound away from zero is needed to ensure the influence of the prior does not overshadow that of the data.

Conditions (i) and (ii) should be considered as a pair. Together, they govern the complexity of the model. In a 'parametric' case (i.e. when Θ is a subset of a finite dimensional vector space) we can typically take $\Theta_n = \Theta$ for all n and use the *maximum likelihood estimator*, for which there is general theory guaranteeing good asymptotic performance. In nonparametric (=infinite dimensional) models, frequentist methods typically require a trade-off between bias and variance; conditions (i) and (ii) give a Bayesian version of this, with (i) giving the bias of the prior towards simple sets Θ_n, while (ii) verifies that the parameter can be well estimated within these sets.

In the literature, including the original celebrated result of [32], it is common in place of (ii) to assume an entropy condition, which perhaps makes it even clearer that conditions (i) and (ii) govern the complexity of the model. The above formulation using concentration of an estimator was introduced in Giné & Nickl [37], and is better suited to the problems addressed here because in Chapter 2 it allows us to access bodies of work on martingale, Markov, and path-continuity properties of a diffusion which are not as well suited to proving an entropy condition, while in Chapter 3, the estimator formulation makes it easier to accommodate necessary boundedness restrictions on the conductivity function.

[3] Consider the estimator $\hat{\theta}$ given by taking the centre of a (nearly) smallest posterior ball of mass at least 1/2 (choose arbitrarily in case of non-uniqueness; the 'nearly' conceals that we need to do this in a measurable way, but it is not important to the result). For C as in Theorem 1.3, if $d(\hat{\theta}, \theta_0) > 2C\varepsilon_n$ then the posterior balls of radius $C\varepsilon_n$ around $\hat{\theta}$ and θ_0 are disjoint; the latter has posterior mass tending to 1 on a sequence of events of $P^n_{\theta_0}$–probability tending to 1, hence the former has posterior mass less than 1/2 for large enough n on this sequence of events, yielding a contradiction. So $P^n_{\theta_0}(d(\hat{\theta}, \theta_0) > 2C\varepsilon_n) \to 0$.

Before the proof, let's unpack the result. The posterior $\Pi(\cdot \mid X^{(n)})$ is a measure-valued $P_{\theta_0}^n$–random variable: that is, for each value $x^{(n)}$ that $X^{(n)}$ can take under $P_{\theta_0}^n$, we obtain a measure $\Pi_{x^{(n)}} = \Pi(\cdot \mid X^{(n)} = x^{(n)})$. The mass the posterior gives to the set $\{\theta : d(\theta, \theta_0) > C\varepsilon_n\}$ is therefore a $[0,1]$–valued $P_{\theta_0}^n$–random variable. Theorem 1.3 says that this random variable tends to zero in $P_{\theta_0}^n$–probability, and even gives a rate $(2e^{-(\zeta+4)n\varepsilon_n^2})$ of convergence to zero.

We further illustrate with a simple (parametric) example.

Example. **Prior:** Under Π, $\theta \sim U(0,1)$ on the parameter set $\Theta = (0,1)$.

Model: Under P_θ^n, observe $X^{(n)} = (X_1, \ldots, X_n)$ with $X_i \overset{iid}{\sim} N(\theta, 1)$ for $1 \le i \le n$. Thus the likelihood is $L_n(\theta) = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum(X_i-\theta)^2}$.

Posterior: The posterior has density (w.r.t. Lebesgue measure) given by

$$\pi(\theta \mid X^{(n)}) = c^{-1} L_n(\theta) \mathbb{1}_{\theta \in [0,1]}, \quad c = c(X^{(n)}) = \int_0^1 L_n(\theta)\, \mathrm{d}\theta$$

Figure 1.2 shows the posterior density for two somewhat representative draws from P_{θ_0} for $\theta_0 = 0.7$, together with the probability each posterior gives to the set $(0.6, 0.8)$. With enough draws from the true distribution, we hope the posterior will be a curve very tightly centred around θ_0, hence the posterior mass of this set will be close to 1. This can be proved using Theorem 1.3: balls around θ_0 of radius a constant multiple of[4] $\varepsilon_n = n^{-1/2}\log(n)^{1/2}$ have posterior mass tending to 1. (The balls $B_{KL}(\varepsilon)$ can be shown to contain $\{|\theta - \theta_0| \le \varepsilon\}$ so that the small ball condition (iii) can be verified explicitly in this model, while for $\Theta_n = \Theta$, (ii) can be shown to hold for the estimator $\hat{\theta} = 0 \vee \bar{X}_n \wedge 1$ using the standard normal tail inequality[5] $\Pr(Z > u) \le e^{-u^2/2}$.)

A main step in the proof of Theorem 1.3 is to demonstrate an 'evidence lower bound'. Rewriting (1.11) as

$$\Pi(B \mid X) = \frac{\int_B (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\, \mathrm{d}\Pi(\theta)}{\int_\Theta (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\, \mathrm{d}\Pi(\theta)},$$

the following lemma says that for ε_n bounded roughly between zero and $n^{-1/2}$, the denominator is not too small, on an event of probability tending to 1.

[4]This rate is not sharp: this posterior in fact contracts at rate $C_n n^{-1/2}$ for any sequence C_n tending to ∞ arbitrarily slowly (e.g. see [34], Chapter 8). This illustrates a downside of insufficiently refined methods of analysing Bayesian posteriors: often superfluous log factors are required, either relative to the rate truly achieved by the posterior (these can often be removed by more careful analysis) or by the posterior rate relative to the minimax rate (sometimes more careful choices of priors can achieve the minimax rate even if the simplest choices do not).

[5]Proved for $u > 1$ by introducing a factor $t > 1$ in the integral $\int_u^\infty e^{-t^2/2}\, \mathrm{d}t$ so that it can be explicitly evaluated, and for $0 < u < 1$ by noting $\Pr(Z > u) \le 1/2 \le e^{-u^2/2}$.

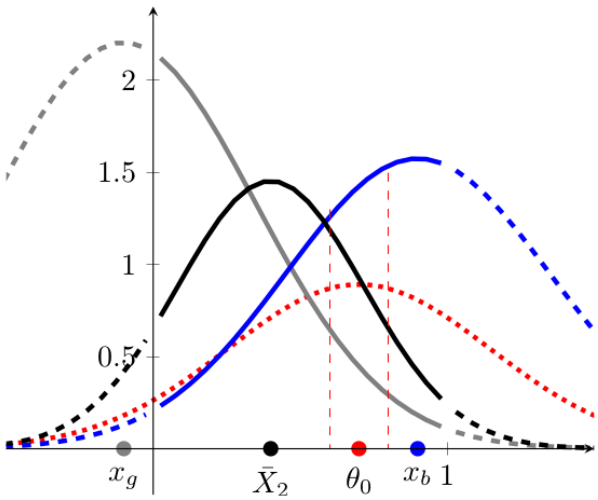

Fig. 1.2 Plotted in red dotted lines is the true density p_{θ_0} of the data points, centred at $\theta_0 = 0.7$. Also in red are dashed vertical lines demarcating the set $(0.6, 0.8)$. In grey and blue are the posteriors corresponding to data points x_g and x_b respectively (the dashed lines show the squared exponential functions which are truncated to give the posteriors, marked with solid lines), while the black plot, centred at $\bar{X}_2 = (x_g + x_b)/2$, is the two-point posterior for these draws.
The posterior probability of the set $(0.6, 0.8)$ is 0.091 if we observe x_g, 0.283 if we observe x_b, and 0.185 if we observe both points.

Lemma 1.4 (Evidence lower bound, ELBO)**.** *Suppose the positive sequence ε_n satisfies $\varepsilon_n \to 0$ and write B^n_{KL} for $B^n_{KL}(\varepsilon_n)$. Define the event*

$$A_n = \left\{ \int_\Theta (p_\theta^{(n)}/p_{\theta_0}^{(n)})\, \mathrm{d}\Pi(\theta) \geq \Pi(B^n_{KL}) e^{-2n\varepsilon_n^2} \right\}.$$

Then $P^n_{\theta_0}(A_n^c) \leq (n\varepsilon_n^2)^{-1}$.

Remark. The result remains essentially true, though with $P_{\theta_0}(A_n^c)$ tending to zero at a different rate, if we define A_n instead by $A_n = \{\int_\Theta (p_\theta^{(n)}/p_{\theta_0}^{(n)})\, \mathrm{d}\Pi(\theta) \geq \Pi(B^n_{KL}) e^{-Bn\varepsilon_n^2}\}$ for any $B > 1$. That is to say, the exact value 2 in the exponent is not important for the proof. This propagates through to Theorem 1.3, where the exponent $-(2\zeta + 8)n\varepsilon_n^2$ is not sharp.

Proof. Write $\Pi' = \Pi/\Pi(B^n_{KL})$ for the renormalised restriction of Π to B^n_{KL}. Then by Jensen's inequality we have

$$\int_\Theta (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\, \mathrm{d}\Pi(\theta) \geq \Pi(B^n_{KL}) \exp\left(\int_{B^n_{KL}} \log(p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)}))\, \mathrm{d}\Pi'(\theta) \right).$$

Write $Z = \int_{B_{KL}^n} \log(p_\theta^{(n)}/p_{\theta_0}^{(n)})\,\mathrm{d}\Pi'(\theta) = -\int_{B_{KL}^n} \log(p_{\theta_0}^{(n)}/p_\theta^{(n)})\,\mathrm{d}\Pi'(\theta)$. Applying Fubini's Theorem and using the definition of B_{KL}^n, we see that

$$E_{\theta_0}^n Z \geq -\sup_{\theta\in B_{KL}^n} E_{\theta_0}^n \log(p_{\theta_0}^{(n)}/p_\theta^{(n)}) \geq -n\varepsilon_n^2.$$

Further, applying Jensen's inequality and twice applying Fubini's Theorem, we see

$$\begin{aligned}
\mathrm{Var}_{\theta_0}^n Z &= E_{\theta_0}^n\Big(\int_{B_{KL}^n} \log(p_\theta^{(n)}/p_{\theta_0}^{(n)})\,\mathrm{d}\Pi'(\theta) - E_{\theta_0}^n Z\Big)^2 \\
&= E_{\theta_0}^n\Big(\int_{B_{KL}^n} (\log(p_\theta^{(n)}/p_{\theta_0}^{(n)}) - E_{\theta_0}^n \log(p_\theta^{(n)}/p_{\theta_0}^{(n)}))\,\mathrm{d}\Pi'(\theta)\Big)^2 \\
&\leq E_{\theta_0}^n \int_{B_{KL}^n} \Big(\log(p_\theta^{(n)}/p_{\theta_0}^{(n)}) - E_{\theta_0}^n \log(p_\theta^{(n)}/p_{\theta_0}^{(n)})\Big)^2 \mathrm{d}\Pi'(\theta) \\
&= \int_{B_{KL}^n} \mathrm{Var}_{\theta_0}\Big(\log(p_{\theta_0}^{(n)}/p_\theta^{(n)})\Big)\,\mathrm{d}\Pi'(\theta) \leq n\varepsilon_n^2,
\end{aligned}$$

the inequality in the final line following from the definition of B_{KL}^n. Note that Fubini's theorem indeed applies both times we have used it, once because the integrand is non-negative and once because we can show the expression considered is integrable, using that $|x| \leq 1 + x^2$ for $x \in \mathbb{R}$.

Together, these bounds on the mean and variance of Z tell us that

$$P_{\theta_0}\Big(\exp(Z) < \exp(-2n\varepsilon_n^2)\Big) \leq P_{\theta_0}\Big(|Z - EZ| > n\varepsilon_n^2\Big) \leq (n\varepsilon_n^2)^{-1},$$

as required, where we have applied Chebyshev's inequality to obtain the final inequality. □

We are ready to prove the contraction result Theorem 1.3.

Proof of Theorem 1.3. First we note that the estimator concentration condition (ii) implies the existence of exponentially powerful tests of $\theta = \theta_0$ vs suitably separated alternatives. Let $\psi_n(X^{(n)}) = \mathbb{1}\{d(\hat{\theta}, \theta_0) > \frac{1}{2}C\varepsilon_n\}$. Then $E_{\theta_0}^n \psi_n = P_{\theta_0}^n(d(\hat{\theta}, \theta_0) > \frac{1}{2}C\varepsilon_n) \leq e^{-(2\zeta+8)n\varepsilon_n^2}$, and by the triangle inequality

$$\sup_{\substack{\theta\in\Theta_n,\\ d(\theta,\theta_0)>C\varepsilon_n}} E_\theta^n[1-\psi_n] \leq \sup_{\substack{\theta\in\Theta_n,\\ d(\theta,\theta_0)>C\varepsilon_n}} P_\theta^n(d(\hat{\theta},\theta) > \tfrac{1}{2}C\varepsilon_n) \leq e^{-(2\zeta+8)n\varepsilon_n^2}. \tag{1.14}$$

Now consider the following decomposition: writing $S = \{\theta \in \Theta_n : d(\theta, \theta_0) > C\varepsilon_n\}$, we have, for A_n as in Lemma 1.4,

$$\Pi(d(\theta,\theta_0) > C\varepsilon_n \mid X^{(n)}) \leq \mathbb{1}_{A_n^c} + \psi_n + \Pi(\Theta_n^c \mid X^{(n)})\mathbb{1}_{A_n} + \Pi(S \mid X^{(n)})\mathbb{1}_{A_n}[1-\psi_n],$$

so that the probability under $P_{\theta_0}^n$ of the event $\Big\{\Pi(d(\theta,\theta_0) > C\varepsilon_n \mid X^{(n)}) \geq 2e^{-(\zeta+4)n\varepsilon_n^2}\Big\}$ is upper bounded by

$$P_{\theta_0}^n(A_n^c) + E_{\theta_0}^n \psi_n + \\ P_{\theta_0}^n\Big(\Pi(\Theta_n^c \mid X^{(n)})\mathbb{1}_{A_n} \geq e^{-(\zeta+4)n\varepsilon_n^2}\Big) + P_{\theta_0}^n\Big(\Pi(S \mid X^{(n)})\mathbb{1}_{A_n}[1-\psi_n] \geq e^{-(\zeta+4)n\varepsilon_n^2}\Big).$$

In view of Lemma 1.4 and the construction of the tests ψ_n, the first two terms in the above tend to zero.

On the event A_n, for any set B the denominator in the expression

$$\Pi(B \mid X^{(n)}) = \frac{\int_B (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\,\mathrm{d}\Pi(\theta)}{\int_\Theta (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\,\mathrm{d}\Pi(\theta)}$$

can be bounded by $\Pi(B_{KL}^n)^{-1}e^{2n\varepsilon_n^2} \leq e^{(\zeta+2)n\varepsilon_n^2}$. Thus, also using Fubini's theorem and noting that $E_{\theta_0}^n[(p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})] = E_\theta^n[1] = 1$, we see

$$E_{\theta_0}^n[\Pi(\Theta_n^c \mid X^{(n)})\mathbb{1}_{A_n}] \leq e^{(\zeta+2)n\varepsilon_n^2} E_{\theta_0}^n\Big[\int_{\Theta_n^c} (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\,\mathrm{d}\Pi(\theta)\Big] = e^{(\zeta+2)n\varepsilon_n^2}\Pi(\Theta_n^c),$$

so that by Markov's inequality and condition (i), we deduce

$$P_{\theta_0}^n\Big(\Pi(\Theta_n^c \mid X^{(n)})\mathbb{1}_{A_n} \geq e^{-(\zeta+4)n\varepsilon_n^2}\Big) \leq e^{(2\zeta+6)n\varepsilon_n^2}\Pi(\Theta_n^c) \to 0.$$

It remains to bound $P_{\theta_0}^n(\Pi(S|X^{(n)})\mathbb{1}_{A_n}[1-\psi_n] \geq e^{-(\zeta+4)n\varepsilon_n^2})$. Appealing to Fubini's theorem and (1.14) we see that

$$\begin{aligned} E_{\theta_0}^n[\Pi(S \mid Y)\mathbb{1}_{A_n}(1-\psi)] &= E_{\theta_0}^n\Big[\mathbb{1}_{A_n}(1-\psi_n)\frac{\int_S (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\,\mathrm{d}\Pi(\theta)}{\int_\Theta (p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\,\mathrm{d}\Pi(\theta)}\Big] \\ &\leq e^{(\zeta+2)n\varepsilon_n^2} E_{\theta_0}^n\Big[\int_S \tfrac{p_\theta^{(n)}}{p_{\theta_0}^{(n)}}(X^{(n)})(1-\psi_n)(X^{(n)})\,\mathrm{d}\Pi(\theta)\Big] \\ &\leq e^{(\zeta+2)n\varepsilon_n^2}\int_S E_\theta^n[(1-\psi_n)(X^{(n)})]\,\mathrm{d}\Pi(\theta) \\ &\leq e^{-(\zeta+6)n\varepsilon_n^2}, \end{aligned}$$

hence by Markov's inequality

$$P_{\theta_0}^n\Big(\Pi(S \mid X^{(n)})\mathbb{1}_{A_n}(1-\psi) \geq e^{-(\zeta+4)n\varepsilon_n^2}\Big) \leq e^{-2n\varepsilon_n^2} \to 0.$$

Uniformity over $\tilde{\Theta}$ follows from the fact that, given a uniform version of condition (iii), all the conditions and all the rates attained are uniform in θ_0. □

Remarks. 1. As mentioned, in the diffusion setting of Chapter 2 we will bypass the inverse nature of the problem by expressing as regression, so proving the conditions of Theorem 1.3 hold will suffice in that chapter, rather than requiring Theorem 1.5.

2. In order to apply the theorem, we typically prove a small ball result of the form

$$\bar{d}(\theta, \theta_0) \leq f(\varepsilon_n) \implies \theta \in B^n_{KL}$$

for some metric $\bar{d}$ (often but not necessarily taken to be the same metric d in which contraction is proved) and for some increasing function f (usually linear as in Lemma 2.14, but see also the proof of Lemma 3.11 where a quadratic function is used), since it is generally straightforward to lower bound the prior mass given to balls in a metric $\bar{d}$, and harder to directly verify the mass on abstract Kullback–Leibler type balls.

3. While the above contraction rate theory is fairly satisfactory for estimation, it is deficient from the point of view of *uncertainty quantification*. In particular, 'credible sets', sets with posterior probability $1 - \alpha$ for some $0 < \alpha < 1$, need not have the advertised coverage rate. One could attempt to prove a nonparametric *Bernstein–von Mises* type theorem, which would give the asymptotic shape of the posterior and allow the construction of credible sets which (asymptotically) are also confidence sets at the advertised level. See for example the pioneering works of Castillo & Nickl [15, 16] or, for results in nonlinear inverse problems similar to those considered here, see Nickl [61], and Nickl & Söhl [63]. Note that establishing a posterior contraction rate as in Theorem 1.3 constitutes a key first step towards a proof of a Bernstein–von Mises result, since it allows one to localise the posterior around the true parameter.

1.4.2 Inverse problems

Recall now that we are interested in Bayesian estimation in an inverse problem in which the measure P^n_θ depends on θ only through $G(\theta)$. Placing a prior on θ induces a prior on $G(\theta)$, and Theorem 1.3 allows us to deduce contraction around $G(\theta_0)$ under some conditions on this prior. A 'stability' result (continuity of the inverse map) will then allow us to deduce contraction around the true θ_0, yielding the following theorem. Recall $B^n_{KL}(\varepsilon)$ was defined in (1.12).

Theorem 1.5. *Let d be a metric on Θ and let $\tilde{d}$ be a metric on $G(\Theta)$. Suppose for some positive sequences ε_n, ξ_n satisfying $\varepsilon_n \to 0, \xi_n \to 0$, $n\varepsilon_n^2 \to \infty$, we have, for some sets $\Theta_n \subset \Theta$ and some constants $C, \zeta > 0$,*

(a) $\Pi(\Theta_n^c) \leq e^{-(2\zeta+8)n\varepsilon_n^2}$,

(b) There exists an estimator $\hat{G}$ of $G(\theta)$ such that

$$\sup_{\theta \in \Theta_n} P_\theta^n(\tilde{d}(\hat{G}, G(\theta)) > \tfrac{1}{2}C\varepsilon_n) \leq e^{-(2\zeta+8)n\varepsilon_n^2},$$

(c) $\Pi(B_{KL}^n(\varepsilon_n)) \geq e^{-\zeta n\varepsilon_n^2}$,

(d) For $\theta \in \Theta_n$, $\tilde{d}(G(\theta), G(\theta_0)) \leq C\varepsilon_n \implies d(\theta, \theta_0) \leq \xi_n$.

Then

$$P_{\theta_0}^n\Big(\Pi(d(\theta, \theta_0) > \xi_n \mid X^{(n)}) \geq 3e^{-(\zeta+4)n\varepsilon_n^2}\Big) \to 0, \quad \text{as } n \to \infty. \tag{1.15}$$

Further assume that the metric d arises from a norm $\|\cdot\|$, that the conditions (c) and (d) are true for all θ_0 in some (norm-) bounded subset $\tilde{\Theta}$ of Θ, that $e^{-n\varepsilon_n^2}\xi_n^{-1} \to 0$ as $n \to \infty$, and that the prior has finite second moment (satisfying a uniform bound if we have a sequence of priors),

$$\sup_n E^\Pi[\|\theta\|^2] < \infty.$$

Then the posterior mean $E^\Pi[\theta \mid X^{(n)}]$ estimates θ_0 at rate ξ_n, uniformly in $\tilde{\Theta}$; precisely

$$\sup_{\theta_0 \in \tilde{\Theta}} P_{\theta_0}^n\Big(\|E^\Pi[\theta \mid X^{(n)}] - \theta_0\| > 2\xi_n\Big) \to 0. \tag{1.16}$$

Remarks.

1. Since Θ is infinite dimensional, the sense in which the posterior mean is to be interpreted must be clarified. Here we consider the *Bochner* mean: choosing a sequence of finitely valued random variables u_n (whose expectations are naturally defined as weighted averages) such that $E^\Pi[\|u_n - \theta\| \mid X] \to 0$, we define $E^\Pi[\theta \mid X] = \lim_n E^\Pi[u_n \mid X]$ (see e.g. [38, §2.6.1] for details).

2. As with Remark 2 after the proof of Theorem 1.3, it is desirable to prove a result of the form

$$d(\theta, \theta_0) \leq f(\varepsilon_n) \implies \theta \in B_{KL}^n.$$

Since B_{KL}^n depends on θ only through $G(\theta)$, a continuity result for the forward map, controlling $\tilde{d}(G(\theta), G(\theta_0))$ in terms of $d(\theta, \theta_0)$ can help us achieve this. See for example Lemma 3.11 and its dependence on Lemma 3.6.

3. It is sometimes inconvenient to specify priors for which (c) holds for all $\theta_0 \in \Theta$, with boundary issues often causing problems, hence the allowance for a smaller set $\tilde{\Theta}$ on which we target uniform convergence. An alternative (equivalent) perspective, taken in Chapter 3, is to specify the prior on a superset of Θ.

4. Typical stability results, controlling $d(\theta, \theta_0)$ in terms of $\tilde{d}(G(\theta), G(\theta_0))$, are polynomial (i.e. of the form $d(\theta, \theta_0) \leq K\tilde{d}(G(\theta), G(\theta_0))^\alpha$ for some constants K, α) or even 'logarithmic', as in Lemma 3.7 where prove an estimate of the form $d(\theta, \theta_0) \leq K\{\log[1/\tilde{d}(G(\theta), G(\theta_0))]\}^{-\alpha}$. Given such a stability result, by possibly first increasing ξ_n by a constant factor, we can "decouple" conditions (b) and (d).

The proof of the contraction rate is straightforward given Theorem 1.3. The proof of the consistency of the posterior mean is more involved, and is given here following the structure of Monard–Nickl–Paternain [59].

Proof. First, using conditions (a) to (c), Theorem 1.3 applies on the metric space $(G(\Theta), \tilde{d})$ to yield

$$P^n_{\theta_0}\Big(\Pi(\tilde{d}(G(\theta), G(\theta_0)) > C\varepsilon_n \mid X^{(n)}) \geq 2e^{-(\zeta+4)n\varepsilon_n^2}\Big) \to 0.$$

Then, by condition (d),

$$\Pi(d(\theta, \theta_0) > \xi_n \mid X^{(n)}) \leq \Pi(\Theta_n^c \mid X^{(n)}) + \Pi(\tilde{d}(G(\theta), G(\theta_0)) > C\varepsilon_n \mid X^{(n)}),$$

and, as in the proof of Theorem 1.3, for A_n as in Lemma 1.4,

$$P^n_{\theta_0}(\Pi(\Theta_n^c \mid X^{(n)}) > e^{-(\zeta+4)n\varepsilon_n^2}) \leq P^n_{\theta_0}(A_n^c) + e^{(2\zeta+6)n\varepsilon_n^2}\Pi(\Theta_n^c) \to 0.$$

This yields the contraction rate (1.15).

To prove consistency of the posterior mean $E[\theta \mid X^{(n)}]$, introduce the event

$$\mathcal{A} = A_n \cap \{\Pi(\|\theta - \theta_0\| > \xi_n \mid X^{(n)}) \leq 3e^{-(\zeta+4)n\varepsilon_n^2}\},$$

and decompose

$$P^n_{\theta_0}(\|E^\Pi[\theta \mid X^{(n)}] - \theta_0\| > 2\xi_n) \leq P^n_{\theta_0}(\mathcal{A}^c) + P^n_{\theta_0}(\|E^\Pi[\theta \mid X^{(n)}] - \theta_0\|\mathbb{1}_{\mathcal{A}} > 2\xi_n).$$

In view of (1.15) and Lemma 1.4, the first term on the right vanishes as $n \to \infty$.

For the second term, we have the following chain of inequalities, appealing to several standard inequalities as listed (note that any norm is convex as a function of one of its

arguments):

$$
\begin{aligned}
&P_{\theta_0}^n\Big(\mathbb{1}_{\mathcal{A}}\|E^{\Pi}[\theta \mid X^{(n)}]-\theta_0\|>2\xi_n\Big)\\
\leq &P_{\theta_0}^n\Big(\mathbb{1}_{\mathcal{A}}E^{\Pi}[\|\theta-\theta_0\| \mid X^{(n)}]>2\xi_n\Big) && \text{(Jensen)}\\
\leq &P_{\theta_0}^n\Big(\mathbb{1}_{\mathcal{A}}E^{\Pi}[\|\theta-\theta_0\|\mathbb{1}_{\|\theta-\theta_0\|>\xi_n} \mid X^{(n)}]>\xi_n\Big)\\
\leq &P_{\theta_0}^n\Big(\mathbb{1}_{\mathcal{A}}E^{\Pi}[\|\theta-\theta_0\|^2 \mid X^{(n)}]^{1/2}\Pi[\|\theta-\theta_0\|>\xi_n \mid X^{(n)}]^{1/2}>\xi_n\Big) && \text{(Cauchy–Schwarz)}\\
\leq &\xi_n^{-1}E_{\theta_0}^n\Big[\mathbb{1}_{\mathcal{A}}E^{\Pi}[\|\theta-\theta_0\|^2 \mid X^{(n)}]^{1/2}\Pi(\|\theta-\theta_0\|>\xi_n \mid X^{(n)})^{1/2}\Big] && \text{(Markov)}\\
\leq &\xi_n^{-1}E_{\theta_0}^n\Big[\mathbb{1}_{\mathcal{A}}E^{\Pi}[\|\theta-\theta_0\|^2 \mid X^{(n)}]\Big]^{1/2}E_{\theta_0}^n\Big[\mathbb{1}_{\mathcal{A}}\Pi(\|\theta-\theta_0\|>\xi_n \mid X^{(n)})\Big]^{1/2} && \text{(Cauchy–Schwarz).}
\end{aligned}
$$

From the definition of $\mathcal{A}$ it is immediate that $E_{\theta_0}^n\Big[\mathbb{1}_{\mathcal{A}}\Pi(\|\theta-\theta_0\|>\xi_n \mid X^{(n)})\Big] \leq 3e^{-(\zeta+4)n\varepsilon_n^2}$, and by Lemma 1.4 and Fubini's theorem observe that

$$
\begin{aligned}
E_{\theta_0}^n\Big[\mathbb{1}_{\mathcal{A}}E^{\Pi}[\|\theta-\theta_0\|^2 \mid X^{(n)}]\Big] &\leq E_{\theta_0}^n\Big[e^{(\zeta+2)n\varepsilon_n^2}\int_{\Theta}\|\theta-\theta_0\|^2(p_\theta^{(n)}/p_{\theta_0}^{(n)})(X^{(n)})\,\mathrm{d}\Pi(\theta)\Big]\\
&\leq 2e^{(\zeta+2)n\varepsilon_n^2}(\|\theta_0\|^2+E^{\Pi}[\|\theta\|^2]).
\end{aligned}
$$

Overall, we deduce

$$
P_{\theta_0}^n\Big(\|E^{\Pi}[\theta \mid X^{(n)}]-\theta_0\|\mathbb{1}_{\mathcal{A}}>2\xi_n\Big) \leq \xi_n^{-1}\sqrt{6}e^{-n\varepsilon_n^2}(\|\theta_0\|^2+E^{\Pi}[\|\theta\|^2])^{1/2} \to 0
$$

and the result follows, noting that as with Theorem 1.3, uniformity in θ_0 is immediate from uniformity of the conditions assumed and of the rates attained in the proof. □

1.5 Computation for Bayesian inverse problems

This thesis concerns theory for Bayesian methods, and the practicality of these methods for inverse problems setting was justified by observing that they bypass the need to invert the forward operator G. Of course, this justification requires that natural Bayesian estimators are themselves tractable, and the purpose of this section is to briefly support that claim in a general setting. The later chapters give references regarding computation in the specific models considered therein.

In linear inverse problems, optimisation based estimators such as the posterior mode (the 'maximum a posteriori probability, MAP, estimator', which for many common priors coincides with the *Tikhonov regulariser* for an appropriate penalisation term) can for typical priors be expressed as an explicit function of the data X, and as a result are computationally feasible. In nonlinear inverse problems, such estimators (whose

theoretical performance can be guaranteed using techniques related to those discussed in this thesis – see e.g. Nickl–van de Geer–Wang [64]) remain computationally feasible if the objective function is convex. However, in the EIT setting considered in Chapter 3, the objective function is not convex and so the MAP estimator is not tractable. Fortunately, the posterior mean, whose theoretical performance was considered in Theorem 1.5, is generically computable via *Monte Carlo methods* (i.e. using samples from the posterior), in particular *Markov chain Monte Carlo* (MCMC): for a function ψ, given samples $\theta^i,\ i = 1, \dots, K$ drawn i.i.d. from the posterior or drawn from a Markov chain whose invariant distribution is the posterior, the central limit theorem (see [58], Theorem 17.0.1 in the Markov Chain case) tells us that under mild assumptions

$$E_{\Pi(\cdot|X)}[\psi(\theta)] \approx \frac{1}{K}\sum_{i=1}^{K}\psi(\theta^i), \tag{1.17}$$

with (stochastic) approximation error of order $K^{-1/2}$.

Assume that the posterior $\Pi(\cdot \mid X^{(n)})$ has a density $\pi(\cdot|X^{(n)})$ with respect to some reference measure. Typically, it is easy to compute $\pi(\theta|X^{(n)})$, up to a normalising constant, from the expression $\pi(\theta|X^{(n)}) \propto \pi(\theta)p_\theta^{(n)}(X^{(n)})$ (integrating over Θ to calculate the normalising constant is often difficult). We can thus access the basic *accept-reject algorithm*, and the *Metropolis–Hastings algorithm* to follow, as methods for sampling from the posterior. Both are written here for a general target (unnormalised) density ν, since the algorithms apply more broadly than just to the case $\nu = \pi(\cdot \mid X^{(n)})$ used in Bayesian statistics.

Algorithm 1 Accept-reject

input a density q, easy to sample from and satisfying $\nu(\theta) \le Mq(\theta)$ for all $\theta \in \Theta$, for some known constant M.
repeat
  sample $\phi \sim q$.
  accept ϕ with probability $\nu(\phi)/(Mq(\phi))$.
until K samples have been accepted
output the accepted samples

Informally, the fact that the accepted samples have distribution with density proportional to ν can be justified as follows. A rephrasing of the algorithm samples $X \sim q$ and $U \sim [0,1]$ independently and defines $Y = Mq(X)U$. Then (X, Y) is uniform on the subgraph of Mq, which is a superset of the subgraph of ν. If (X, Y) lies in the subgraph

of ν, accept the pair; this leads to points uniformly sampled on the subgraph of ν, whose X coordinates are therefore sampled from the normalised density proportional to ν.

The only difficulty in applying the accept-reject algorithm is choosing the proposal distribution q. When Θ is a bounded subset of $\mathbb{R}^N$ for some $N \in \mathbb{N}$, we may take q to be the uniform distribution on Θ as a default choice, but the acceptance probability $M^{-1} \int_\Theta \nu$ becomes vanishingly small as N increases or Θ becomes unbounded. When Θ is infinite dimensional, no default choice of q can work generically.

The Metropolis–Hastings algorithm adds some flexibility which can help us bypass the dimensionality issues suffered by the accept-reject algorithm.

Algorithm 2 Metropolis–Hastings

input an initial value θ^0, and a family of proposal densities $(q(\theta, \cdot))_{\theta \in \Theta}$ which are easy to sample from.
for $0 \leq j \leq K$, independently sample $\phi^j \sim q(\theta^j, \cdot)$, and $U \sim U([0,1])$.
  if $U \leq [\nu(\phi^j)q(\theta^j, \phi^j)]/[\nu(\theta^j)q(\phi^j, \theta^j)]$ **then** set $\theta^{j+1} \leftarrow \phi^j$
  else set $\theta^{j+1} \leftarrow \theta^j$
output $\theta^1, \ldots, \theta^K$

It can be shown that the Markov chain $(\theta^j)_{j \geq 0}$ has invariant density proportional to ν, hence (under extra assumptions ensuring the central limit theorem holds) the expression (1.17) is justified for these samples, with error $CK^{-1/2}$ for a constant C. As with the accept-reject algorithm, it is crucial to choose q carefully: the constant C grows as the acceptance probability (the probability we set θ^{j+1} as ϕ^j rather than as θ^j) decreases. Unlike the accept-reject algorithm, though, there is a default choice for the proposal distribution q in the Metropolis–Hastings which works for many of the infinite-dimensional models of interest: the *preconditioned Crank–Nicholson (pCN) proposal*, in which we set $\phi^j = \sqrt{1-\beta^2}\theta^j + \beta\xi_j$ for some i.i.d. random variables ξ_j and some constant $\beta \in [0,1]$. See Cotter et al. [22] for an argument that this algorithm works well when the target distribution is the posterior corresponding to a Gaussian prior and the variables ξ_j are drawn from the prior, and note that calculating the acceptance probability requires evaluating the forward map G at ϕ^j and θ^j, but does not require inverting G.

1.6 Elliptic PDEs: a brief introduction

The models considered in this thesis arise from PDEs, so this section introduces some of the elementary techniques used in PDE theory, with a particular focus on techniques

allowing the derivation of the types of stability results described in Section 1.4.2. The theory in this section is described for complex-valued functions,[6] as used in Chapter 3, but passes virtually unchanged to the real-valued functions used in Chapter 2.

We consider in particular *boundary value problems*, which are PDEs of the form

$$L[u] = g \text{ on } D, \;\; u = f \text{ on } \partial D,$$

for some partial differential operator L and some smooth bounded domain D (i.e. a connected bounded open subset of $\mathbb{R}^d$, with smooth boundary denoted ∂D). Throughout this thesis all domains will be smooth and bounded, so we will omit explicit mention and simply say 'domain'.

The operator L will be taken to be a second order operator, written in divergence form.

Definition. A *second order partial differential operator in divergence form* is a map L, taking smooth enough functions $u : D \to \mathbb{C}$ as inputs, of the form

$$L[u] = -\sum_{j,k} \partial_k(a_{jk}\partial_j u) + \sum_j b_j \partial_j u + cu, \tag{1.18}$$

for functions a_{jk}, b_j, c, where (without loss of generality, by changing b_j if necessary) we assume that $a_{kj} = a_{jk}^*$ for all j, k, where * denotes the complex conjugate, and where ∂_j denotes the partial derivative in the jth direction.

The PDE on which we will apply the techniques from this section is the Dirichlet problem (1.3) ($\nabla \cdot (\gamma \nabla u) = 0$ on D, $u = f$ on ∂D); we will assume in Chapter 3 that $\gamma \geq m$ for some constnat $m > 0$ which ensures this PDE is 'uniformly elliptic' and so we focus on this class of PDEs.

Definition (Uniform ellipticity)**.** The operator L of (1.18) is *uniformly elliptic* if there exists $m > 0$, called the *ellipticity constant*, such that

$$\sum_{jk} a_{jk}(x)\xi_j\xi_k^* \geq m\|\xi\|^2 \quad \text{for all } \xi \in \mathbb{C}^d, x \in D.$$

($\mathbb{C}^d$ is equipped with the usual Euclidean norm.)

For PDEs whose partial differential operator is uniformly elliptic, a 'weak' solution theory is more appropriate than a classical solution theory, and so we introduce the

[6] While the *codomain* of the functions considered is complex, their *domains* will be real. Thus real derivatives, rather than more rigid complex derivatives, are used.

notion of weak derivatives and the Sobolev spaces $H^r(D)$ consisting of "r–times weakly differentiable functions".

1.6.1 Weak derivatives and Sobolev spaces

The maximal function space considered is the space of *distributions* on D, i.e. the (linear) dual space of the space $C_c^\infty(D)$ of complex-valued smooth functions compactly supported in D. We define *weak* or *distributional* (partial) derivatives ∂_j on the space of distributions by 'duality', i.e. $\partial_j u$ is the (unique) distribution v defined by

$$\int_D v\phi = -\int_D u\frac{\partial\phi}{\partial x_j} \quad \forall\phi \in C_c^\infty(D), \tag{1.19}$$

where the *classical* or *strong* derivative appears on the right-hand side ($\frac{\partial\phi}{\partial x_j}(x) = \lim_{h\to 0}\frac{\phi(x+he_j)-\phi(x)}{h}$ for e_j a unit vector in the x_j direction). Integrals initially being defined for functions only, for a distribution v we interpret $\int_D v\phi$ as notation for the duality pairing, i.e. $\int v\phi$ gives result of applying the linear map $v : C_c^\infty(D) \to \mathbb{C}$ to ϕ. We naturally associate to a measurable function f the distribution $\tilde{f} : C_c^\infty(D) \to \mathbb{C}$, $\tilde{f}(\phi) = \int_D f\phi$, consistently with the above notation. Any distribution v can be associated with at most one function f up to a Lebesgue null set (i.e. if $\tilde{f} = v = \tilde{g}$ then $f = g$ Lebesgue almost everywhere[7]), and whenever such a function associated with the distribution $\partial_j\tilde{u}$ exists, it will also be called the weak derivative of the function u. Integration by parts shows that if the classical derivative $\frac{\partial u}{\partial x_j}$ exists then it is also the weak derivative. Henceforth all derivatives, whether denoted ∂_j or $\frac{\partial}{\partial x_j}$, will be defined in a weak sense unless otherwise specified, and we will not distinguish between a function and its associated distribution.

Integration by parts remains true for weak derivatives via the *divergence theorem* (see e.g. [17] eq. (38)).

Theorem 1.6 (The divergence theorem)**.** *For a domain D with outward unit normal ν on the boundary ∂D,*

$$\int_D \nabla\cdot F = \int_{\partial D} F\cdot\nu,$$

where $\nabla\cdot$ denotes the divergence operator on vector fields F, $\nabla\cdot F = \sum_{j=1}^d \partial_j F_j$.

[7] Proved as follows: let $\phi \in C_c^\infty(\mathbb{R}^d)$ be a 'bump function', with $\phi = 0$ if $|x| > 1$, $\phi(x) \geq 0$ for all x, and $\int_{\mathbb{R}^d}\phi = 1$. Then $\phi_\varepsilon = \varepsilon^{-d}\phi(\varepsilon^{-1}(\cdot))$ is called a *mollifier*. It can be shown that $\int_D f(x)\phi_\varepsilon(a-x)\,\mathrm{d}x \to f(a)$ as $\varepsilon \to 0$ for Lebesgue-almost all $a \in D$, provided that f is locally integrable (which means that the integral is well-defined for ε sufficiently small) – see [27, Appendix C.4]. But $\int_D f(x)\phi_\varepsilon(a-x)\,\mathrm{d}x = \int_D g(x)\phi_\varepsilon(a-x)\,\mathrm{d}x$ for all a and all ε sufficiently small by assumption, hence taking limits we see $f(a) = g(a)$ almost everywhere.

Remark. We can for example apply to $F = u\,\nabla\, v - v\,\nabla\, u$ to see that

$$\int_D u\,\Delta\, v = \int_D v\,\Delta\, u + \int_{\partial D} u\frac{\partial v}{\partial \nu} - v\frac{\partial u}{\partial \nu},$$

hence the description as 'integration by parts' (∇ denotes the usual gradient operator and Δ the Laplacian). More generally, writing

$$\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}} \equiv \partial_1^{\alpha_1}\cdots\partial_d^{\alpha_d}, \tag{1.20}$$

for a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)$, $\alpha_j \in \mathbb{N}\cup\{0\}$ for $j \le d$, of order $|\alpha| = \sum_{j\le d}\alpha_j$, and assuming u, v are compactly supported in D to avoid having to define the appropriate boundary operators (cf. [53, Chapter II, Theorem 2.1, p114]), we have

$$\int_D u\partial^\alpha v = (-1)^{|\alpha|}\int_D v\partial^\alpha u.$$

The Sobolev spaces $H^r(D)$, $r \in \mathbb{R}$, are constructed to capture the notion of the number of weak derivatives a function has: the derivative ∂_j maps $H^{r+1}(D) \to H^r(D)$ continuously for L^2–Sobolev spaces $H^r(D)$, $r \ge 0$ (indeed, it is immediate from the definition below of $H^r(D)$ that $\|\partial_j u\|_{H^r(U)} \le \|u\|_{H^{r+1}(U)}$ for $r \ge 0$; for $r < 0$ see Theorem 12.1 in [53] Chapter 1, p71). The space $H^0(D) = L^2(D)$ is the Lebesgue space of functions

$$L^2(D) = \{f : D \to \mathbb{C} \text{ measurable s.t. } \|f\|^2_{L^2(D)} := \int_D |f(x)|^2 < \infty\},$$

where functions which are equal almost everywhere are understood to be identified, and for $r \in \mathbb{R}$ we define as follows.

Definitions $(H^r(D), H^r_0(D), H^r_{\text{loc}}(D))$**.** Definitions are drawn from Lions & Magenes [53]. For $r \in \mathbb{N}\cup\{0\}$ we define

$$H^r(D) = \{f \in L^2(D) \text{ s.t. } \partial^\alpha f \in L^2(D) \text{ for all multi-indices } \alpha \text{ satisfying } |\alpha| \le r\}.$$

$H^r(D)$ is a Hilbert space, equipped with the inner product

$$\langle f, g\rangle_{H^r(D)} = \sum_{|\alpha|\le r}\int_D \partial^\alpha f \cdot \partial^\alpha g^*,$$

and associated norm $\|\cdot\|_{H^r(D)}$ (see [53, Chapter I, §1.1, p1]). We use the analogous notation to denote the inner products and norms for each Hilbert space considered here.

For $r \in \mathbb{R}$, $r \geq 0$ we define $H^r(D)$ via interpolation (see [53, Chapter I, §9.1, p40] for details).

$H_0^r(D)$ is defined as the $\|\cdot\|_{H^r(D)}$–closure of $C_c^\infty(D) \subset H^r(D)$ (see [53, Chapter I, §11.1, p55]), and the spaces $H^r(D)$, $r < 0$ are defined as the (topological) dual spaces, equipped with the dual norms,

$$H^r(D) = (H_0^r(D))^* = \{\text{distributions } f \text{ s.t. } \|f\|_{H^r(D)} = \sup_{\substack{g \in H_0^{|r|}(D), \\ \|g\|_{H^{|r|}(D)}=1}} \int fg^* < \infty\}, \quad r < 0.$$

(See [53, Chapter I, §12.1, p70].)

$H^r_{\text{loc}}(D)$ is defined as the set of distributions f such that $f\phi \in H^r(D)$ for all $\phi \in C_c^\infty(D)$ (see [53, Chapter II, §3.2, p125]) or, equivalently, as the set of distributions f such that $f|_U \in H^r(U)$ for all domains $U \Subset D$, where the symbol $\Subset$ is read 'compactly contained' and $U \Subset D$ means the closure $\bar{U}$ is a subset of the interior $\operatorname{int} D = D$.

A fundamental result of PDE theory is Poincaré's inequality, which says that the H^1 norm is equivalent to the H^1 seminorm on the subset $H_0^1(D) \subset H^1(D)$. There is also a version for functions whose average on D is zero, but only the given version will be used in this thesis.

Theorem 1.7 (Poincaré's inequality)**.** *There exists a constant $C = C(D)$ such that for all $u \in H_0^1(D)$,*

$$\|u\|_{H^1(D)} \leq C\|\nabla u\|_{L^2(D)}.$$

Proof. See Corollary 6.31 in [3]. Roughly, the idea of the proof is that the fundamental theorem of calculus shows that a classically differentiable function cannot take values much larger than those of its derivative if the boundary values are zero, and this extends to $H_0^1(D)$ functions by density of $C_c^\infty(D)$ (density can be proved using convolution with mollifiers, similarly to the footnote in Section 1.6.1). □

1.6.2 Boundary values and trace theorems

Recall we are considering boundary value problems

$$L[u] = g \text{ on } D, \; u = f \text{ on } \partial D,$$

and we seek weak solutions in some Sobolev space $H^r(D)$. Since elements u of $H^r(D)$ are in fact *equivalence classes* of functions differing on Lebesgue null sets, we must clarify the precise sense in which "$u = f$ on ∂D" is to be understood. The definition of $H_0^r(D)$

given in the previous section allows us to address the case $f = 0$, and more generally, we make sense of boundary values through the *trace operator*. For the following, recall that a function h is *harmonic* on the domain D if $\Delta h = 0$ on D.

Theorem 1.8 (Trace theorem)**.** *There exists a linear operator* tr *acting on the Sobolev scale of functions* $\bigcup_{s\in\mathbb{R}} H^s(D)$ *which acts on* $u \in C(\bar{D})$ *as restriction to the boundary* ∂D *(*$\operatorname{tr} u = u|_{\partial D}$*) and for which the following is true. Given* $s \in \mathbb{R}$ *suppose* w *lies in* $H^{s+1/2}(D)$*; in the case* $s \leq 0$ *further assume* w *is harmonic. Then* $\operatorname{tr} w$ *is in* $H^s(\partial D)$ *and satisfies*

$$\|\operatorname{tr} w\|_{H^s(\partial D)} \leq C\|w\|_{H^{s+1/2}(D)} \tag{1.21}$$

for some constant $C = C(s, D)$*. Further,*

- *The map* tr *has a continuous right inverse for* $s > 0$*; that is, there exists* $C' = C'(s, D)$ *such that for any* $f \in H^s(\partial D)$*, there exists* $F \in H^{s+1/2}(D)$ *satisfying* $\operatorname{tr} F = f$ *and* $\|F\|_{H^{s+1/2}(D)} \leq C'\|f\|_{H^s(\partial D)}$*.*
- *We may also define (outward) normal partial derivatives on the boundary* $\frac{\partial^j}{\partial \nu^j}$*,* $j > 0$*, similarly. These satisfy, for* $s > 0$ *and some constant* $C'' = C''(s, D, j)$*,*

  $$\|\tfrac{\partial^j w}{\partial \nu^j}\|_{H^s(\partial D)} \leq C''\|w\|_{H^{s+j+1/2}(D)}.$$

  For $s < 0$ *and* w *harmonic, the estimate holds for* $j = 1$*.*

The space $H^s(\partial D)$ in which the boundary functions live can be defined via the differential geometry notion of a *smooth atlas*. That is, we choose open sets U covering ∂D, then on each set we choose smoothly varying local coordinate functions $\phi_U : U \to \mathbb{R}^{d-1}$. Such a smooth atlas immediately allows the definition of classical derivatives on the manifold ∂D, since these only depend on local behaviour of a function.[8] It is not immediately clear that this idea can be extended to weak derivatives, which are defined in a non-local way, but indeed it can. We omit the details here, since in any case when the boundary Sobolev spaces are needed (in Chapter 3) we use an alternative definition in terms of the eigenfunctions of the Laplace–Beltrami operator on ∂D. See [53], Chapter I, Section 7.3 (p34) for a careful definition built on the ideas outlined in this paragraph.

Proof of Theorem 1.8. One proof method, in the case $s = 1$, is to show that the restriction operator satisfies the estimate $\|u|_{\partial D}\|_{L^2(\partial D)} \leq C\|u\|_{H^1(D)}$ for all $u \in C^1(\bar{D}) \subset H^1(D)$, for some constant C independent of u (see [27], Section 5.5, Theorem 1). Since $C^1(\bar{D})$ is

[8] Given $f : \partial D \to \mathbb{C}$, we say f is differentiable if $f \circ \phi_U^{-1} : \phi_U(U) \to \mathbb{C}$ is (classically) differentiable as a map from $\mathbb{R}^{d-1}$ to $\mathbb{C}$ for all 'charts' (U, ϕ_U) in the atlas.

dense in $H^1(D)$, this allows us to extend the restriction operator to $H^1(D)$ by taking limits, yielding the continuous trace map $\mathrm{tr} : H^1(D) \to L^2(\partial D)$.

The sharper and more general result (1.21) can be found in [53]. See Chapter I Theorem 9.2 (p41) for the case $s > 0$; for w harmonic and $s \leq -3/2$, see Chapter II Theorem 6.5 (and Remark 6.4, pages 175-177), and for w harmonic and $-3/2 < s < 1/2$ see Chapter II Theorem 7.3 (p187). In these latter cases, in the notation of [53], take the operator A to be the Laplacian Δ, and consider the normal system given by the singleton $B_0 = \mathrm{tr}$ (or the singleton $B_0 = \frac{\partial}{\partial \nu}$ for the case of normal derivatives; note condition (1.11) in Chapter II means we cannot attain higher orders of normal derivatives in the same way). □

Remark. Given the trace theorem, we have the following characterisation of $H_0^r(D)$ for $r > 0$, equivalent to the definition given in the previous section (see [53], Chapter I, Theorem 11.5 on p62):

$$H_0^r(D) = \{f \in H^r(D) : \tfrac{\partial^j f}{\partial \nu^j}\Big|_{\partial D} = 0, 0 \leq j < r - 1/2\},$$

with the normal boundary derivatives defined in a trace sense and $\partial^0 f/\partial \nu^0 := \mathrm{tr}\, f$.

1.6.3 Weak solutions via the Lax–Milgram theorem

Given a partial differential operator L defined by $L[u] = -\sum_{j,k} \partial_k(a_{jk}\partial_j u) + \sum_j b_j \partial_j u + cu$, as in (1.18), we associate to L a sesquilinear operator $B = B_L$ (i.e. B is linear in the first argument and conjugate linear in the second argument, satisfying $B(u, \lambda v + \mu w) = \lambda^* B(u, v) + \mu^* B(u, w)$) defined by

$$B(u, v) := \int_D \Big(cuv^* + \sum_j v^* b_j \partial_j u + \sum_{j,k} a_{jk} \partial_j u \partial_k v^*\Big). \tag{1.22}$$

Then a *weak solution* of the elliptic partial differential equation

$$L[u] = g \text{ on } D, \;\; u = f \text{ on } \partial D \tag{1.23}$$

is understood to mean a function $u \in H^1(D)$ whose trace is f and such that

$$B(u, v) = \int_D g v^* \quad \forall v \in H_0^1(D). \tag{1.24}$$

Integration by parts (for classical derivatives) shows that any classical solution (i.e. any $C^2(D)$ function solving (1.23) for the derivatives of L taken in a classical sense) is also a

weak solution. The weak formulation is more robust, more appropriately characterises the regularity of solutions, and is amenable to Hilbert space theory, in particular the *Lax–Milgram theorem.*

Theorem 1.9 (Lax–Milgram)**.** *Let $(H, \|\cdot\|_H)$ be a Hilbert space (with complex scalars). Let $B : H \times H \to \mathbb{C}$, $A : H \to \mathbb{C}$ be sesquilinear and conjugate linear respectively. Suppose*

- *B is bounded: for some $C > 0$ and all $u, v \in H$, $|B(u, v)| \leq C\|u\|_H \|v\|_H$.*
- *B is coercive: for some $c > 0$ and all $u \in H$, $B(u, u) \geq c\|u\|_H^2$.*
- *A is bounded: for some $K > 0$ and all $v \in H$, $|A(v)| \leq K\|v\|_H$. The smallest such K is the* operator norm *of A, denoted $\|A\|_{H\to\mathbb{C}}$.*

Then the equation

$$B(u, v) = A(v) \quad \forall v \in H$$

has a unique solution $u \in H$. Moreover, this solution satisfies

$$\|u\|_H \leq K/c.$$

Proof. Given existence, the norm bound follows from the calculation $K\|u\|_H \geq |A(u)| = |B(u, u)| \geq c\|u\|_H^2$.

For existence, when B is a 'hermitian form', i.e. $B(u, v) = B(v, u)^*$ (as will be the case when this result is used in Theorem 1.10), the proof follows from a single application of the Riesz representation theorem, since B defines an inner product on H in this case. See [27] §6.2, Theorem 1 for a proof in the case where B is not hermitian (note the proof there is given for H having real scalars and B bilinear, but the same argument works in the complex scalar case considered here). □

We are ready to prove the existence of weak solutions to the PDE (1.23). Note a version of the following theorem remains true under weaker assumptions on the coefficients $\{c, b_j\}$, but the proof is slightly more involved and is not needed in this thesis: see [27, §6.2] for details.

Theorem 1.10. *Let L be as in (1.18) for bounded coefficient functions a_{jk}, b_j, c satisfying $c(x) \geq 0$, $b_j(x) = 0$, for all $x \in D$ and $1 \leq j \leq d$, and suppose L is uniformly elliptic with ellipticity constant m. Let B be the associated sesquilinear operator, and let A be a bounded conjugate linear operator $H_0^1(D) \to \mathbb{C}$. There is a unique $u \in H_0^1$ solving*

$$B(u, v) = A(v) \quad \forall v \in H_0^1(D),$$

and this solution satisfies

$$\|u\|_{H_0^1(D)} \leq \tfrac{1}{mk}\|A\|_{H_0^1(D)\to\mathbb{C}},$$

where $k = k(D)$ is the constant of the Poincaré inequality (Theorem 1.7). In particular, for $f = 0$ and $g \in H^{-1}(D)$, the elliptic PDE (1.23) has a unique weak solution u satisfying

$$\|u\|_{H_0^1(D)} \leq \tfrac{1}{mk}\|g\|_{H^{-1}(D)}.$$

Proof. We apply Theorem 1.9 on the Hilbert space $H_0^1(D)$. For the more general result it suffices to show the sesquilinear B defined in (1.22) is bounded and coercive. Boundedness is an immediate consequence of the Cauchy–Schwarz inequality: we have

$$|B(u,v)| \leq d^2 \max(\|c\|_\infty, \max_{j,k}\|a_{jk}\|_\infty)\|u\|_{H^1(D)}\|v\|_{H^1(D)},$$

where $\|\cdot\|_\infty$ denotes the usual supremum norm and we recall d is the dimension of D. For coercivity, observe by nonnegativity of c and ellipticity,

$$B(u,u) = \int_D (c|u|^2 + \sum_{jk} a_{jk}\partial_j u \partial_k u^*) \geq m\|\nabla u\|^2_{L^2(D)},$$

so that Theorem 1.7 implies that B is coercive with coercivity constant mk where k is the Poincaré constant for the domain D.

Finally, in the particular case $A(v) = \int_D gv^*$, recalling $H^{-1}(D)$ is defined as the dual of $H_0^1(D)$, we have

$$|A(v)| \leq \|g\|_{H^{-1}(D)}\|v\|_{H_0^1(D)}$$

hence $\|A\|_{H_0^1(D)\to\mathbb{C}} \leq \|g\|_{H^{-1}(D)}$. □

Remarks. 1. The above arguments can also be used to show that the 'Neumann problem' ($L[w] = 0$ on D, $\partial w/\partial\nu = 0$ on ∂D) has a solution $w \in H^1(D)$ which is unique up to an additive constant: existence is by the Lax–Milgram theorem, and uniqueness is because we find $\|\nabla w\|^2_{L^2(D)} = 0$ for $w \in H^1(D)$ satisfying $B(v,w) = 0$ for all $v \in H^1(D)$.

2. To prove existence of weak solutions to (1.23) with more general boundary data f, it suffices to find a function F whose trace is f and which is smooth enough that Theorem 1.10 implies the existence of a weak solution $w \in H_0^1$ to $L[w] = g - L[F]$ in D. Then $u = F + w$ will solve (1.23) with boundary data f. For $f \in H^s(\partial D)$, some $s \geq 1/2$, such an F exists since the invertibility of the trace operator in Theorem 1.8 yields $F \in H^1(D)$ with $\operatorname{tr} F = f$ and, recalling that the derivatives ∂_j

map $H^r(D)$ to $H^{r-1}(D)$ for $r \in \mathbb{R}$, we see $L[F] \in H^{-1}(D)$. For rougher f, such an F can be found using theory for harmonic functions, for example as in the proof of Lemma 3.17.

1.6.4 Regularity estimates

For ordinary differential equations (i.e. $d \equiv \dim D = 1$ in the above), if u is a classical solution of the equation $L[u] = g$ then u has two more classical derivatives than g. For PDEs, this is no longer true in general. However, for elliptic PDEs, the weak formulation of this statement is true under mild conditions: Section 6.3 in [27] gives some such conditions for the case of boundary data $f = 0$. Here, we instead include the following result, which addresses only the case $c = 0 = b_j \ \forall j$, but considers non-trivial boundary data, hence is ideally suited to the Dirichlet problem (1.3) studied in Chapter 3.

Theorem 1.11. *Suppose $L = -\sum_{j,k} \partial_k(a_{jk}\partial_j u)$ is uniformly elliptic on a smooth domain D, with smooth coefficient functions a_{jk}. For $s \in \mathbb{R}$, assume $f \in H^{s+3/2}(\partial D)$ and $g \in H^s(D)$. In the case $s < 0$, further assume g is compactly supported in D. Then there exists $u \in H^{s+2}(D)$ solving (1.23), and this solution satisfies*

$$\|u\|_{H^{s+2}(D)} \leq C(\|g\|_{H^s(D)} + \|f\|_{H^{s+3/2}(\partial D)}). \tag{1.25}$$

The constant C depends on the coefficients of L, on the domain D and, in the case $s < 0$, on the support of g.

Proof. Observe, in the notation of [53, Remark 7.2, Chapter II, p188], $N = N^* = \{0\}$. This follows from Theorem 1.10: since the zero function is a unique solution in H^1_0 to $L[w] = 0$ in D, $w = 0$ on ∂D, this is also the unique solution in the smaller space $C^\infty_c(\bar{D})$. (In the notation of [53, Chapter II, Theorem 2.1, p114], take $T_j = \mathrm{tr}$.) The remark then states that a solution exists and for $s \geq 0$ gives the continuity estimate (1.25). In the case $s < 0$, the continuity estimate of [53] has a different norm in place of the $H^s(D)$ norm on the right-hand side, but for g of compact support the two norms coincide up to a constant depending only on the support of g. □

Remarks. 1. If, instead of Dirichlet boundary data $\mathrm{tr}\, u = f$, we are given Neumann boundary data $\partial u/\partial \nu = f$, there exists a solution u satisfying

$$\|u\|_{H^{s+2}(D)/\mathbb{C}} \leq C(\|g\|_{H^s(D)} + \|f\|_{H^{s+1/2}(\partial D)}),$$

where $\|u\|_{H^{s+2}(D)/\mathbb{C}} \equiv \inf_{z\in\mathbb{C}}\|u-z\|_{H^{s+2}(D)}$ is the usual quotient norm. The proof only differs in that to show N, N^* consist of the constant functions we use Remark 1 after Theorem 1.10 .

2. In Chapter 3, we use Theorem 1.11 for equivalence classes of functions defined up to a constant. Fixing a representative f of the equivalence class and applying to $(f-z)$ for each $z \in \mathbb{C}$, we see that the result holds with quotient norms on u and f.

3. For this theorem we assume a_{jk} is smooth for all j, k. In Chapter 3, we will not assume the conductivity γ is smooth, since this is highly undesirable in an imaging context (indeed, when using EIT we are generally looking for jump discontinuities in the conductivity, corresponding to the boundaries of different materials). We in fact only apply Theorem 1.11 to $L = \Delta$, hence there is no conflict.

Whilst the above theory only gives weak regularity (i.e. regularity in Sobolev spaces), the *Sobolev embedding theorem* (a name collectively given to a number of inequalities) allows the deduction of classical regularity as a result. Here is one version of the theorem.

Theorem 1.12 (Sobolev embedding)**.** *Suppose $s > d/2$. Then any $u \in H^s(D)$ has a uniformly continuous representative. Moreover, the induced embedding $H^s(D) \hookrightarrow C_u(D)$, where $C_u(D)$ denotes the space of uniformly continuous functions on D, is continuous; that is, for some constant $C = C(s, D)$ and all $u \in H^s(D)$,*

$$\|u\|_\infty \leq C\|u\|_{H^s(D)}.$$

Proof. See for example Theorem 7.3.4c in [3]. Note that the Besov space $B^s_{2,2}(D)$ is equal to $H^s(D)$. □

Remarks. 1. The result can be applied to yield higher order derivatives. For $s > 1+d/2$, if $u \in H^{s+1}(D)$, then $\partial_j u \in H^s(D) \hookrightarrow C_u(D)$. By considering mollifiers as in the footnote in Section 1.6.1, we can choose $u_k \in C^\infty(D)$ such that $u_k \to u$ almost everywhere and $\partial_j u_k \to \partial_j u$ in sup-norm. Then u_k converges to some $v \in C^1(D)$ by standard arguments; by uniqueness of limits we see $v = u$ Lebesgue almost everywhere so that u has a $C^1(D)$ representative. This argument bootstraps up, so that if $u \in H^s(D)$ for all $s > 0$, in fact $u \in C^\infty(D)$.

2. The Sobolev embedding is also true on sufficiently smooth manifolds; in particular, for $H^s(\partial D)$.

1.7 Background reading

This section gathers some background material in the areas studied in this thesis. The focus is on survey works.

Frequentist analysis of Bayesian procedures Ghosal & van der Vaart [34], Giné & Nickl [38, chapter 8]. For diffusions in particular: van Zanten [87]

Stochastic diffusions Durrett [26], Rogers & Williams [72, 73], Bass [7], Bhattacharya & Waymire [9]

Bayesian computation Cotter, Roberts, Stuart & White [22], Stuart [77]

PDE inverse problems Katchalov–Kury–Lassas [47], Isakov [44]

The Calderón problem/EIT Uhlmann [81], Salo [75]

PDEs and functional analysis Evans [27], Gilbarg & Trudinger [36], Aubin [5]

Chapter 2

Contraction rates for scalar diffusions with high-frequency data

Notation

Most of the notation to be used in this chapter is informally gathered here.

X: A solution to $\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t$.

$\dot{X}$: The periodised diffusion $\dot{X} = X \mod 1$.

b, σ: Drift function, diffusion coefficient.

$\mu = \mu_b$; π_b: Invariant distribution/density of $\dot{X}$.

$P_b^{(x)}$: Law of X on $C([0,\infty])$ (on $C([0,\Delta])$ in Section 2.4) for initial condition $X_0 = x$.

E_b; P_b; Var_b: Expectation/probablity/variance according to the law of X started from μ_b.

E_μ; Var_μ, and similar: Expectation/variance according to the subscripted measure.

$\mathbb{W}_\sigma^{(x)}$: Notation for $P_b^{(x)}$ when $b = 0$.

$p_b(t,x,y), \dot{p}_b(t,x,y)$: Transition densities of $X, \dot{X}$ (with respect to Lebesgue measure).

$\tilde{p}_b$: Density (with respect to $\mathbb{W}_\sigma^{(x)}$) of $P_b^{(x)}$ on $C([0,\Delta])$.

$I_b(x) = \int_0^x (2b/\sigma^2(y))\,\mathrm{d}y$.

$X^{(n)} = (X_0, \ldots, X_{n\Delta})$; $x^{(n)} = (x_0, \ldots, x_{n\Delta})$; $p_b^{(n)}(x^{(n)}) = \pi_b(x_0)\prod_{i=1}^n p_b(\Delta, x_{(i-1)\Delta}, x_{i\Delta})$.

b_0: The true parameter generating the data.

μ_0, π_0, p_0 etc.: Shorthand for $\mu_{b_0}, \pi_{b_0}, p_{b_0}$ etc.

$\sigma_L > 0$; $\sigma_U < \infty$: A lower and upper bound for σ.

L_0: A constant such that $n\Delta^2 \log(1/\Delta) \leq L_0$ for all n.

$C^1_{\text{per}}([0,1])$: The space of continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(x+1) = f(x)$ for $x \in \mathbb{R}$.

$\Theta = \Theta(K_0)$: The maximal paramater space: $\Theta = \{f \in C^1_{\text{per}}([0,1]) \text{ s.t. } \|f\|_{C^1_{\text{per}}} \leq K_0\}$.

$\Theta_s(A_0) = \{f \in \Theta : \|f\|_{B^s_{2,\infty}} \leq A_0\}$, for $B^s_{2,\infty}$ a (real scalar, 1–periodic) Besov space.

$\mathcal{I} = \{K_0, \sigma_L, \sigma_U\}$.

S_m: Wavelet approximation space of resolution m, generated by periodised Meyer-type wavelets: $S_m = \text{span}\{\psi_{lk} : -1 \leq l < m, 0 \leq k < 2^l\}$, where $\psi_{-1,0}$ is used as notation for the constant function 1.

$D_m = \dim(S_m) = 2^m$; $\pi_m =$(L^2–)orthogonal projection onto S_m.

$w_m(\delta) = \delta^{1/2}(\log(\delta^{-1})^{1/2} + \log(m)^{1/2})$ if $m \geq 1$, $w_m := w_1$ if $m < 1$.

$\mathbb{1}_A$: Indicator of the set (or event) A.

$K(p,q)$: Kullback–Leibler divergence between densities p, q: $K(p,q) = E_p[\log(p/q)]$.

$\text{KL}(b_0, b) = E_{b_0} \log(p_0/p_b)$.

$B^{(n)}_{KL}(\varepsilon) = \left\{ b \in \Theta \text{ s.t. } K(p^{(n)}_0, p^{(n)}_b) \leq (n\Delta+1)\varepsilon^2, \text{Var}_{b_0}\left(\log\left(p^{(n)}_0/p^{(n)}_b\right)\right) \leq (n\Delta+1)\varepsilon^2 \right\}$.

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

Π: The prior distribution.

$\Pi(\cdot \mid X^{(n)})$: The posterior distribution given data $X^{(n)}$.

$\langle \cdot, \cdot \rangle_2$: the $L^2([0,1])$ inner product, $\langle f, g \rangle_2 = \int_0^1 f(x)g(x)\,\mathrm{d}x$ (this chapter uses real-valued functions, so no need for the complex conjugate of g).

$\|\cdot\|_2$: The $L^2([0,1])$–norm, $\|f\|_2^2 = \int_0^1 f(x)^2\,\mathrm{d}x$.

$\|\cdot\|_\mu$: The $L^2(\mu)$–norm, $\|f\|_\mu^2 = \int_0^1 f(x)^2\mu(\mathrm{d}x) = \int_0^1 f(x)^2\pi_b(x)\,\mathrm{d}x$.

$\|\cdot\|_\infty$: The L^∞– (supremum) norm.

$\|\|_{C^1_{\text{per}}}$: The C^1_{per}–norm, $\|f\|_{C^1_{\text{per}}} = \|f\|_\infty + \|f'\|_\infty$.

$\|\cdot\|_n$: The empirical L^2–norm $\|f\|_n^2 = \sum_{k=1}^n f(X_{k\Delta})^2$.

2.0 Introduction

Let's reprise the stochastic diffusion model described in Section 1.1, and expand on its key features. Consider a scalar diffusion process $(X_t)_{t\geq 0}$ starting at some X_0 and evolving according to the stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \tag{2.1}$$

where W_t is a standard Brownian motion. It is of considerable interest to estimate the parameters b and σ, which are arbitrary functions (until we place further assumptions on their form), so that the model is naturally *nonparametric*. The problems of estimating σ and b can essentially be decoupled in the setting to be considered here (see Section 2.1), so in this chapter we consider estimation of the drift function b when the diffusion coefficient σ is assumed to be given.

It is realistic to assume that we do not observe the full trajectory $(X_t)_{t\leq T}$ but rather the process sampled at discrete time intervals $(X_{k\Delta})_{k\leq n}$. The estimation problem for b and σ has been studied extensively and minimax rates have been attained in two sampling frameworks: *low-frequency*, where Δ is fixed and asymptotics are taken as $n \to \infty$ (see Gobet–Hoffmann–Reiss [39]), and *high-frequency*, where asymptotics are taken as $n \to \infty$ and $\Delta = \Delta_n \to 0$, typically assuming also that $n\Delta^2 \to 0$ and $n\Delta \to \infty$ (see Hoffmann [43], Comte et al. [21]). See also e.g. [23], [40], [68], [82] for more papers addressing nonparametric estimation for diffusions.

For typical frequentist methods, one must know from which sampling regime the data is drawn. In particular, the low-frequency estimator of [39] is consistent in the high-frequency setting but numerical simulations suggest it does not attain the minimax rate (see the discussion in Chorowski [18]), while the high-frequency estimators of [43] and [21] are not even consistent with low-frequency data. Often real data arrives all at once so that the appropriate asymptotic regime is not clear, hence it is desirable to use estimators which perform well independent of the regime. The only previous result known to the author in this direction in the nonparametric setting considered here is found in [18], where Chorowski is able to estimate the diffusion coefficient σ but not the drift, and obtains the minimax rate when σ has 1 derivative but not for smoother diffusion coefficients. See also Chorowski & Trabs [19], wherein an estimator adapting to the sampling regime is given but only for the sampling period bounded below, and Coca [20] for an estimator adapting to the sampling regime in a Lévy process setting.

For this chapter we consider estimation of the parameters in a diffusion model from a nonparametric Bayesian perspective. An attraction of Bayesian methods for the discretely

sampled diffusion model is that the statistician need only specify a prior, and the prior need not reference the sampling regime, so Bayesian methodology provides a natural candidate for a unified approach to the high- and low-frequency settings. Note also that, building on ideas outlined in Section 1.5, Bayesian methods for diffusion estimation can be implemented in practice (e.g. see Papaspiliopoulos et al. [65]). Under the frequentist assumption of a fixed true parameter, the results of this chapter imply that Bayesian methods can adapt both to the sampling regime and also to unknown smoothness of the drift function (see the remarks after Propositions 2.4 and 2.2 respectively for details).

It has previously been shown that in the low-frequency setting we have a *posterior contraction rate*, guaranteeing that posteriors corresponding to reasonable priors concentrate their mass on neighbourhoods of the true parameter shrinking at the fastest possible rate (up to log factors) – see Nickl & Söhl [62]. To complete a proof that such posteriors contract at a rate adapting to the sampling regime, it remains to prove a corresponding contraction rate in the high-frequency setting. This forms the key contribution of the current chapter: we prove that a large class of "reasonable" priors will exhibit posterior contraction at the optimal rate (up to log factors) in L^2–distance. This in turn guarantees that point estimators based on the posterior will achieve the frequentist minimax optimal rate (see the remark after Theorem 2.1) in both high- and low-frequency regimes.

The broad structure of the proof, inspired by that in [62], is as described Section 1.4.1. The main ingredients are:

- A concentration inequality for a (frequentist) estimator, from which we construct tests of the true b_0 against a set of suitable (sufficiently separated) alternatives. See Section 2.3.
- A small ball result, to relate the L^2–distance to the information-theoretic Kullback–Leibler "distance". See Section 2.4.

Although the main proof structure reflects that used in [62] for the low-frequency case, the details are very different. Estimators for the low-frequency setting are typically based on the mixing properties of $(X_{k\Delta})$ viewed as a Markov chain and the spectral structure of its transition matrix (see Gobet–Hoffmann–Reiss [39]) and fail to take full advantage of the local information one sees when $\Delta \to 0$. Here we instead use an estimator introduced in Comte et al. [21], which uses the assumption $\Delta \to 0$ to view estimation of b as a regression problem. To prove this estimator concentrates depends on a key insight of this chapter: the Markov chain concentration results used in the low-frequency setting (which give *worse* bounds as $\Delta \to 0$) must be supplemented by Hölder type continuity results,

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

Similarly, the small ball result in the low-frequency setting is proved using Markov chain mixing. Here, we instead adapt the approach of van der Meulen & van Zanten [83]. They demonstrate that the Kullback–Leibler divergence in the discrete setting can be controlled by the corresponding divergence in the continuous data model; a key result of this chapter is that in the high-frequency setting this control extends to give a bound on the variance of the log likelihood ratio.

2.1 Framework and assumptions

For a scalar diffusion X satisfying (2.1), assume the following. All functions in this chapter are taken to be real-valued.

Assumption 1. $\sigma \in C^2_{\text{per}}([0,1])$ is given. Continuity guarantees the existence of an upper bound $\sigma_U < \infty$; further assume the existence of a lower bound $\sigma_L > 0$ so that $\sigma_L \leq \sigma(x) \leq \sigma_U$ for all $x \in [0,1]$. Here $C^2_{\text{per}}([0,1])$ denotes 1–periodic $C^2(\mathbb{R})$ functions.

Assumption 2. b is periodic, and continuously differentiable with given norm bound. Precisely, assume $b \in \Theta$, where, for some arbitrary but known constant K_0,

$$\Theta = \Theta(K_0) = \{f \in C^1_{\text{per}}([0,1]) \text{ s.t. } \|f\|_{C^1_{\text{per}}} = \|f\|_\infty + \|f'\|_\infty \leq K_0\}.$$

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

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

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

with $B^s_{p,q}$ denoting a periodic Besov space and $\|\cdot\|_{B^s_{p,q}}$ denoting the associated norm: see Section 2.1.1 for a definition of the periodic Besov spaces we use (readers unfamiliar with Besov spaces may substitute the L^2–Sobolev space – defined as in Section 1.6.1 but with real scalars – $H^s((0,1)) = B^s_{2,2}((0,1)) \subseteq B^s_{2,\infty}((0,1))$ for $B^s_{2,\infty}$ and only mildly weaken the results). We generally assume the regularity index s is unknown. Our results will therefore aim to be *adaptive*, at least in the smoothness index (to be fully adaptive we would need to adapt to K_0 also).

Under Assumptions 1 and 2, there is a unique strong solution to (2.1) (see, for example, Bass [7] Theorem 24.3). Moreover, this solution is also weakly unique (= unique in law) and satisfies the Markov property (see [7] Proposition 25.2 and Theorem 39.2). Denote by $P_b^{(x)}$ the law (on the cylindrical σ–algebra of $C([0,\infty])$) of the unique solution to (2.1) started from $X_0 = x$.

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

Beyond guaranteeing existence and uniqueness of a solution, the assumptions also guarantee the existence of transition densities for the discretely sampled process (see Gihman & Skorohod [35], Chapter 3, §13, Theorem 2 for an explicit formula for the transition densities). Morever, there also exists an invariant distribution μ_b, with density π_b, for the periodised process $\dot{X} = X \mod 1$. Defining $I_b(x) = \int_0^x \frac{2b}{\sigma^2}(y)\,\mathrm{d}y$ for $x \in [0,1]$, the density is

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

(see Bhattacharya et al. [8], equations 2.15 to 2.17; note we have chosen a different normalisation constant so the expressions appear slightly different).

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

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

Assumption 3. $X_0 \sim \mu_b$.

Write P_b for the law of the full process X under Assumptions 1 to 3, and write E_b for expectation according to this law. Note μ_b is not invariant for P_b, but nevertheless $E_b(f(X_t)) = E_b(f(X_0))$ for any 1–periodic function f (e.g. see the proof of Theorem 2.6). Since we will be estimating the 1–periodic function b, the assumption that $X_0 \in [0,1)$ is unimportant.

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

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

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

Remarks (Comments on assumptions). *Periodicity assumption.* We assume b and σ are periodic so that we need only estimate b on $[0, 1]$. In the frequentist setting, it is common to instead assume that b satisfies some growth condition ensuring recurrence, then estimate the restriction of b to $[0, 1]$, as in Comte et al. [21]. In the Bayesian framework, it is not sufficient to assume a growth condition; rather, the exact growth rate (not just a bound for it) must be known in order to model the drift sufficiently well with the prior. The periodic setting considered here is therefore more naturally suited to Bayesian methods.

Assuming that $\sigma \in C^2_{\text{per}}$ is given. If we observe continuous data $(X_t)_{t \leq T}$ then σ is known exactly (at least at any point visited by the process) via the expression for the quadratic variation $\langle X \rangle_t = \int_0^t \sigma^2(X_s)\,\mathrm{d}s$. With high-frequency data we cannot perfectly reconstruct the diffusion coefficient from the data, but we can estimate it at a much faster rate than the drift. When b and σ are both assumed unknown, if b is s-smooth and σ is s'-smooth, the minimax errors for b and σ respectively scale as $(n\Delta)^{-s/(1+2s)}$ and $n^{-s'/(1+2s')}$, as can be shown by slightly adapting Theorems 5 and 6 from Hoffmann [43] so that they apply in the periodic setting we use here; since we assume that $n\Delta^2 \to 0$, it follows that $n\Delta \leq n^{1/2}$ for large n, hence we can estimate σ at a faster rate than b regardless of their relative smoothnesses.

Further, note that the problems of estimating b and of estimating σ in the high-frequency setting are essentially independent. For example, the smoothness of σ does not affect the rate for estimating b, and vice-versa – see [43]. We are therefore not substantially simplifying the problem of estimating b through the assumption that σ is given.

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

Assuming a known bound on $\|b\|_{C^1_{\text{per}}}$. The assumption that b has one derivative is a typical minimal assumption to ensure that the diffusion equation (2.1) has a strong solution and that this solution has an invariant density. The assumption of a *known* bound for the C^1_{per}–norm of the function is undesirable, but needed for the proofs in this chapter, in particular to ensure the existence of a uniform lower bound π_L on the

invariant densities (this lower bound is essential for the Markov chain mixing results as its reciprocal controls the mixing time in Theorem 2.6). Using rescaled Gaussian process priors as in Chapter 3, it should be possible to remove this assumption.

Assuming $X_0 \sim \mu_b$. It can be shown (see the proof of Theorem 2.6) that the law of $\dot{X}_t$ converges to μ_b at exponential rate from any starting distribution, so assuming $X_0 \sim \mu_b$ is not restrictive (as mentioned, our fixing $X_0 \in [0,1)$ is arbitrary but unimportant).

Assuming $n\Delta^2 \log(1/\Delta) \leq L_0$. It is typical in the high-frequency setting to assume $n\Delta^2 \to 0$ (indeed the minimax rates in [43] are only proved under this assumption) but for technical reasons we need the above in Section 2.3.

2.1.1 Spaces of approximation

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

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

We denote $\psi_{-1,0} \equiv 1$ for convenience. Denote by $\langle \cdot, \cdot \rangle_2$ the $L^2([0,1])$ inner product and by $\|\cdot\|_2$ the L^2–norm, i.e. $\langle f, g \rangle_2 = \int_0^1 f(x)g(x)\,\mathrm{d}x$ and $\|f\|_2 = \langle f, f \rangle_2^{1/2}$ for $f, g \in L^2([0,1])$. One definition of the (periodic) Besov norm $\|f\|_{B^s_{2,\infty}}$ is, for $f_{lk} := \langle f, \psi_{lk} \rangle_2$,

$$\|f\|_{B^s_{2,\infty}} = |f_{-1,0}| + \sup_{l \geq 0} 2^{ls} \left(\sum_{k=0}^{2^l - 1} f_{lk}^2 \right)^{1/2}, \tag{2.2}$$

with $B^s_{2,\infty}$ defined as those 1–periodic $f \in L^2(\mathbb{R})$ for which this norm is finite. See Giné & Nickl [38] Sections 4.2.3 and 4.3.4 for a construction of periodised Meyer-type wavelets and a proof that this wavelet norm characterisation agrees with other possible definitions of the desired Besov space.

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

$$\|\pi_m b - b\|_2 \leq K \|b\|_{B^s_{2,\infty}} 2^{-ms}, \tag{2.3}$$

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

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

$$\sup_{b\in\Theta} \|\pi_m b - b\|_\infty \to 0 \quad \text{as} \quad m \to \infty. \tag{2.4}$$

This follows from Proposition 4.3.24 in [38] since K_0 uniformly bounds $\|b\|_{C^1_{\text{per}}}$ for $b \in \Theta$.

Alternative approximation spaces. The key property we need for our approximation spaces is that (2.3) and (2.4) hold. The latter is only used for some proofs, and priors built using other function spaces for which an appropriate adaptation of (2.3) holds will achieve the same posterior contraction rates. A version holds for many other function spaces, including for S_m the set of trigonometric polynomials of degree at most m, or, provided $s \le s_{\max}$ for some given $s_{\max} \in \mathbb{R}$, for S_m generated by periodised Daubechies wavelets, if we replace 2^m by $D_m = \dim(S_m)$;

2.2 Main contraction theorem

Let Π be a (prior) probability distribution on some σ–algebra $\mathcal{S}$ of subsets of Θ. Given $b \sim \Pi$ assume that $(X_t : t \ge 0)$ follows the law P_b as described in Section 2.1. Write $p_b(\Delta, x, y)$ for the transition densities

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

and recall p_0 is used as shorthand for p_{b_0}. Assume that the mapping $(b, \Delta, x, y) \mapsto p_b(\Delta, x, y)$ is jointly measurable with respect to the σ–algebras $\mathcal{S}$ and $\mathcal{B}_\mathbb{R}$, where $\mathcal{B}_\mathbb{R}$ is the Borel σ–algebra on $\mathbb{R}$ (see e.g. [54] for detailed discussions of measurability in the diffusion setting). Then it can be shown by standard arguments that the Bayesian posterior distribution given the data is

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

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

A main result of this chapter is the following. Theorem 2.1A is designed to apply to adaptive sieve priors, while Theorem 2.1B is designed for use when the smoothness of the parameter b is known. See Section 2.2.1 for explicit examples of these results in use and see Section 2.5 for their proofs.

Theorem 2.1. *Consider data $X^{(n)} = (X_{k\Delta})_{0\le k\le n}$ sampled from a solution X to (2.1) under Assumptions 1 to 4. Let the true parameter be b_0. Assume the appropriate sets below are measurable with respect to the σ–algebra $\mathcal{S}$.*

A. Let Π be a sieve prior on Θ, i.e. let $\Pi = \sum_{m=1}^{\infty} h(m)\Pi_m$, where $\Pi_m(S_m \cap \Theta) = 1$, for S_m a periodic Meyer-type wavelet space of resolution m as described in Section 2.1.1, and h some probability mass function on $\mathbb{N}$. Suppose we have, for all $\varepsilon > 0$ and $m \in \mathbb{N}$, and for some constants $\alpha, \beta_1, \beta_2, B_1, B_2 > 0$,

(i) $B_1 e^{-\beta_1 D_m} \le h(m) \le B_2 e^{-\beta_2 D_m}$,

(ii) $\Pi_m(\{b \in S_m : \|b - \pi_m b_0\|_2 \le \varepsilon\}) \ge (\varepsilon\alpha)^{D_m}$,

where π_m is the L^2–orthogonal projection onto S_m and $D_m = \dim(S_m) = 2^m$. Then for some constant $M = M(A_0, s, \mathcal{I}, L_0, \beta_1, \beta_2, B_1, B_2, \alpha)$ we have, for any $b_0 \in \Theta_s(A_0)$,

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

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

B. Suppose now $b_0 \in \Theta_s(A_0)$ where $s \ge 1$ and $A_0 > 0$ are both known. Let $j_n \in \mathbb{N}$ be such that $D_{j_n} \sim (n\Delta)^{1/(1+2s)}$, i.e. for some positive constants L_1, L_2 and all $n \in \mathbb{N}$ let $L_1(n\Delta)^{1/(1+2s)} \le D_{j_n} \le L_2(n\Delta)^{1/(1+2s)}$. Let $(\Pi^{(n)})_{n\in\mathbb{N}}$ be a sequence of priors satisfying, for some constant $\alpha > 0$ and for $\varepsilon_n = (n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}$,

(I) $\Pi^{(n)}(\Theta_s(A_0) \cap \Theta) = 1$ for all n,

(II) $\Pi^{(n)}(\{b \in \Theta : \|\pi_{j_n} b - \pi_{j_n} b_0\|_2 \le \varepsilon_n\}) \ge (\varepsilon_n \alpha)^{D_{j_n}}$.

Then we achieve the same rate of contraction; i.e. for some $M = M(A_0, s, \mathcal{I}, L_0, \alpha)$,

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

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

Remark. Optimality. The minimax lower bounds of Hoffmann [43] do not strictly apply because we have assumed σ is given. Nevertheless, the minimax rate in this model should be $(n\Delta)^{-s/(1+2s)}$. This follows by adapting arguments for the continuous data case from Kutoyants [50] Section 4.5 to apply to the periodic model and observing that with high-frequency data we cannot outperform continuous data. Thus (recalling the remark after Theorem 1.3), the rates attained in Theorem 2.1 are optimal, up to the log factors, and there exists an estimator of b attaining these rates.

2.2.1 Explicit examples of priors

Our results guarantee that the following priors will exhibit posterior contraction. Throughout this section we continue to adopt Assumptions 1 to 4, and for technical convenience, we add an extra assumption on b_0. Precisely, recalling that $\{\psi_{lk}\}$ form a family of Meyer-type wavelets as in Section 2.1.1 and $\psi_{-1,0}$ denotes the constant function 1, we assume the following.

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

$$b_0 = \sum_{\substack{l\geq -1\\ 0\leq k<2^l}} \tau_l \beta_{lk}\psi_{lk}, \quad \text{with } |\beta_{lk}| \leq B \text{ for all } l \geq -1 \text{ and all } 0 \leq k < 2^l. \tag{2.5}$$

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

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

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

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

$$b(x) = \sum_{\substack{-1\leq l<m\\ 0\leq k<2^l}} \tau_l u_{lk}\psi_{lk}(x), \qquad x \in [0,1], \tag{2.6}$$

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

Example (Basic sieve prior). Let $\tau_{-1} = \tau_0 = 1$ and $\tau_l = 2^{-3l/2}l^{-2}$ for $l \geq 1$. Let h be a probability distribution on $\mathbb{N}$ as described in Theorem 2.1A, for example, $h(m) = \gamma e^{-2^m}$, where γ is a normalising constant. Let $\Pi = \sum_{m=1}^{\infty} h(m)\Pi_m$ where Π_m is as above.

Proposition 2.2. *The preceding prior meets the conditions of Theorem 2.1A for any b_0 satisfying Assumption 5 with the same τ_l used to define the prior, and for an appropriate constant K_0. Thus, if also $b_0 \in \Theta_s(A_0)$ for some constant A_0, then for some M,*

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

The proof can be found in Section 2.5.1.

Remark. Adaptive estimation. If we assume $b_0 \in \Theta_{s_{\min}}(A_0)$, for some $s_{\min} > 3/2$, Assumption 5 automatically holds with τ_l as in Section 2.2.1 for some constant $B = B(s_{\min}, A_0)$, as can be seen from the wavelet characterisation (2.2). Thus, in contrast to the low-frequency results of [62], the above prior adapts to unknown s in the range $s_{\min} \leq s < \infty$.

When $s > 1$ is known, we fix the rate of decay of wavelet coefficients to ensure a draw from the prior lies in $\Theta_s(A_0)$ by hand, rather than relying on the hyperparameter to choose the right resolution of wavelet space. We demonstrate with the following example. The proofs of Propositions 2.3 and 2.4, also given in Section 2.5.1, mimic that of Proposition 2.2 but rely on Theorem 2.1B in place of Theorem 2.1A.

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

Proposition 2.3. *Assume $\bar{L}_n/(n\Delta)^{1/(1+2s)}$ is bounded away from zero. Then for any $s > 1$, the preceding sequence of priors meets the conditions of Theorem 2.1B for any b_0 satisfying Assumption 5 with the same τ_l used to define the prior, and for an appropriate constant K_0. Thus, for some constant M,*

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

Remark. Assumption 5 with $\tau_l = 2^{-l(s+1/2)}$ in fact forces $b_0 \in B^s_{\infty,\infty} \subsetneq B^s_{2,\infty}$ with fixed norm bound. Restricting to this smaller set does not change the minimax rate, as can be seen from the fact that the functions by which Hoffmann perturbs in the lower bound proofs in [43] lie in the smaller class addressed here. In principle, one could remove this assumption by taking $\tau_l = 2^{-ls}$ and taking the prior $\Pi^{(n)}$ to be the law of $b \sim \Pi_{\bar{L}_n}$ conditional on $b \in \Theta_s(A_0)$.

Example (Prior on the invariant density). In some applications it may be more natural to place a prior on the invariant density and only implicitly model the drift function. With minor adjustments, Theorem 2.1B can still be applied to such priors. We outline the necessary adjustments.

(i) b is not identifiable from π_b and σ^2. We therefore introduce the identifiability constraint $I_b(1) = 0$; fixing $I_b(1)$ as any constant, we can reduce to the case $I_b(1) = 0$ by a translation, hence we make this choice for simplicity (this assumption is standard in the periodic setting, for example see van Waaij & van Zanten [86]).

With this restriction, we have $\pi_b(x) = \frac{e^{I_b(x)}}{G_b \sigma^2(x)}$ for a normalising constant G_b, so that $b = ((\sigma^2)' + \sigma^2(\log \pi_b)')/2$.

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

$$H_0 = \sum_{\substack{l \geq -1 \\ 0 \leq k < 2^l}} \tau_l h_{lk} \psi_{lk}, \quad \text{with } |h_{lk}| \leq B \text{ for all } l \geq -1 \text{ and all } 0 \leq k < 2^l, \tag{2.7}$$

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

(iii) Induce a prior on $b = ((\sigma^2)' + \sigma^2 H')/2$ by putting the prior $\Pi^{(n)} = \Pi_{\bar{L}_n}$ on H, where $\bar{L}_n$ is as in Proposition 2.3.

(iv) To ensure $b \in \Theta_s(A_0)$ we place further restrictions on σ; for example, we could assume σ^2 is smooth. More tightly, it is sufficient to assume (in addition to Assumption 1) that $\sigma^2 \in \Theta_{s+1}(A_1)$ and $\|\sigma^2\|_{C^s_{\text{per}}} \leq A_1$, where C^s_{per} is the Hölder norm, for some $A_1 > 0$. These conditions on σ can be bypassed with a more careful statement of Theorem 2.1B and a more careful treatment of the bias.

Proposition 2.4. *Make changes (i) to (iv) as listed. Then, the obtained sequence of priors meets the conditions of Theorem 2.1B for an appropriate constant K_0, hence $\Pi^{(n)}\Big(\{b \in \Theta : \|b - b_0\|_2 \leq M(n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}\} \mid X^{(n)}\Big) \to 1$ in P_{b_0}–probability for some constant M.*

Remarks. Minimax rates. The assumption (2.7) restricts b_0 beyond simply lying in $\Theta_s(A_0)$. As with Nickl & Söhl [62] Remark 5, this further restriction does not change the minimax rates, except for a log factor induced by the weights l^{-2}.

Adaptation to sampling regime. The prior of Proposition 2.4 is the same as the prior on b in [62]. However, since here we assume σ is given while in [62] it is an unknown parameter, the results of [62] do not immediately yield contraction of this prior at a near-minimax rate in the low-frequency setting. In particular, when σ is known the minimax rate for estimating b with low-frequency data is $n^{-s/(2s+3)}$ (for example see Söhl & Trabs [76]), rather than the slower rate $n^{-s/(2s+5)}$ attained in Gobet–Hoffmann–Reiss [39] when σ is unknown (this improvement is possible because one bypasses the delicate interweaving of the problems of estimating b and σ with low-frequency data). Nevertheless, the prior of Proposition 2.4 will indeed exhibit near-minimax contraction also in the low-frequency setting. An outline of the proof is as follows. The small ball

results of [62] still apply, with minor changes to the periodic model used here in place of their reflected diffusion, so it is enough to exhibit tests of the true parameter against suitably separated alternatives. The identification $b = ((\sigma^2)' + \sigma^2(\log \pi_b)')/2$ means one can work with the invariant density rather than directly with the drift. Finally one shows the estimator from [76] exhibits sufficiently good concentration properties (alternatively, one could use general results for Markov chains from Ghosal & van der Vaart [33]).

It remains an interesting open problem to simultaneously estimate b and σ with a method which adapts to the sampling regime. Extending the proofs of this chapter to the case where σ is unknown would show that the Bayesian method fulfils this goal. The key difficulty in making this extension arises in the small ball section (Section 2.4), because Girsanov's Theorem does not apply to diffusions with differing diffusion coefficients.

Intermediate sampling regime. Strictly speaking, we only demonstrate robustness to the sampling regime in the extreme cases where $\Delta > 0$ is fixed or where $n\Delta^2 \to 0$. The author is not aware of any papers addressing the intermediate regime (where Δ tends to 0 at a slower rate than $n^{-1/2}$) for a nonparametric model: the minimax rates do not even appear in the literature. Since the Bayesian method adapts to the extreme regimes, one expects that it attains the correct rates in this intermediate regime (up to log factors). However, the proof would require substantial extra work, primarily in exhibiting an estimator with good concentration properties in this regime. Kessler's work on the intermediate regime in the parametric case [48] would be a natural starting point for exploring this regime in the nonparametric setting.

2.3 Concentration of a drift estimator

In this section we introduce an estimator and prove it exhibits adequate concentration to satisfy condition (ii) of the general contraction rate result Theorem 1.3. The estimator is adapted from one of Comte et al. [21], constructed by considering drift estimation as a regression-type problem. Specifically, defining

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

we can write

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

Note $R_{k\Delta}$ is a discretization error which vanishes as $\Delta \to 0$ and $Z_{k\Delta}$ takes on the role of noise. Define the *empirical norm* and the related *empirical loss function*

$$\|u\|_n = \Big(\frac{1}{n}\sum_{k=1}^{n} u(X_{k\Delta})^2\Big)^{1/2}, \quad \gamma_n(u) = \frac{1}{n}\sum_{k=1}^{n}[\Delta^{-1}(X_{(k+1)\Delta}-X_{k\Delta})-u(X_{k\Delta})]^2, \quad u:[0,1]\to\mathbb{R},$$

leaving out the $k=0$ term in each case for notational convenience. Recalling that S_m is a periodic Meyer-type wavelet space of resolution m as described in Section 2.1.1 and K_0 is an upper bound for the C^1_{per}–norm of any $b \in \Theta$, for l_n to be chosen we define $\tilde{b}_n$ as a solution to the minimisation problem

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

choosing arbitrarily among minimisers in the (generic) case that there is no unique minimiser. The main result of this section is the following (recall that π_m is the L^2–orthogonal projection map onto S_m and $D_m = \dim(S_m) = 2^m$).

Theorem 2.5. *Consider data $X^{(n)} = (X_{k\Delta})_{0\le k\le n}$ sampled from a solution X to (2.1) under Assumptions 1 to 4. Let $\varepsilon_n \to 0$ be a sequence of positive numbers and let $l_n \to \infty$ be a sequence of positive integers such that $n\Delta\varepsilon_n^2/\log(n\Delta) \to \infty$ and, for some constant L and all n, $D_{l_n} \le Ln\Delta\varepsilon_n^2$. For these l_n, let $\tilde{b}_n$ be defined as above and let $\Theta_n \subseteq \{b \in \Theta \text{ s.t. } \|\pi_{l_n}b - b\|_2 \le \varepsilon_n\}$ contain b_0.*

Then for any $D > 0$ there is a $C = C(\mathcal{I}, L_0, D, L) > 0$ such that for all n sufficiently large

$$\sup_{b\in\Theta_n} P_b\Big(\|\tilde{b}_n - b\|_2 > C\varepsilon_n\Big) \le e^{-Dn\Delta\varepsilon_n^2}.$$

Remark. Previous proofs of Bayesian contraction rates using the concentration of estimators approach (see [37],[62],[70]) have used duality arguments, i.e. the fact that $\|f\|_2 = \sup_{v:\|v\|_2=1}\langle f, v\rangle_2$, to demonstrate that the linear estimators considered satisfy a concentration inequality of the desired form. A key insight in this chapter is that for the model we consider we can achieve the required concentration using the above *minimum contrast* estimator (see Birgé & Massart [10]), for which we need techniques which differ substantially from duality arguments.

2.3.1 General concentration results

We will use three forms of concentration results as building blocks for Theorem 2.5. The first comes from viewing the data $(X_{j\Delta})_{0\le j\le n}$ as a Markov chain and applying Markov

chain concentration results; such results are similar to those used in Nickl & Söhl [62] for the low-frequency case, but here we need to track the dependence of constants on Δ. The second form are useful only in the high-frequency case because they use a quantitative form of Hölder continuity for diffusion processes. An inequality of the third form, based on martingale properties, is introduced only where needed (in Lemma 2.12).

Markov chain concentration results applied to diffusions

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

Theorem 2.6. *Let X solve the scalar diffusion equation (2.1), and grant Assumptions 1 to 3. There exists a constant $\kappa = \kappa(\mathcal{I})$ such that, for all n sufficiently large, all bounded 1–periodic functions $f : \mathbb{R} \to \mathbb{R}$, and all $s \geq 0$,*

$$P_b\Big(\Big|\sum_{k=1}^n \big(f(X_{k\Delta}) - E_\mu[f]\big)\Big| \geq s\Big) \leq 2\exp\Big(-\frac{1}{\kappa}\Delta \min\Big(\frac{s^2}{n\|f\|_\mu^2}, \frac{s}{\|f\|_\infty}\Big)\Big), \tag{2.8}$$

or equivalently

$$P_b\Big(\Big|\sum_{j=1}^n f(X_{j\Delta}) - E_\mu[f]\Big| \geq \max(\sqrt{\kappa v^2 x}, \kappa u x)\Big) \leq 2e^{-x}, \tag{2.9}$$

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

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

$$P_b\Big(\sup_{f\in\mathcal{F}}\Big|\sum_{j=1}^n \big(f(X_{j\Delta}) - E_\mu[f]\big)\Big| \geq \tilde{\kappa}\max\Big\{\sqrt{V^2(d+x)}, U(d+x)\Big\}\Big) \leq 4e^{-x}. \tag{2.10}$$

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

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

Theorem 2.7 (Paulin [66], Proposition 3.4 and Theorem 3.4)**.** *Let $M_1, \dots, M_n$ be a time-homogeneous Markov chain taking values in S with transition kernel $P(x, \mathrm{d}y)$ and invariant measure μ. Suppose M is uniformly ergodic, i.e. $\sup_{x\in S}\|P^n(x,\cdot) - \mu\|_{\mathrm{TV}} \leq K\rho^n$ for some constants $K < \infty$, $\rho < 1$, where $P^n(x,\cdot)$ is the $n-$step transition kernel and $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm for signed measures. Write $t_{mix} = \min\{n \geq 0 : \sup_{x\in S}\|P^n(x,\cdot) - \mu\|_{\mathrm{TV}} < 1/4\}$. Suppose $M_1 \sim \mu$ and $f : S \to \mathbb{R}$ is bounded. Let*

$V_f = \mathrm{Var}[f(M_1)]$, *let* $C = \|f - E[f(M_1)]\|_\infty$. *Then for* $s \geq 0$,

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

Proof of Theorem 2.6. Since f is assumed periodic we see that $f(X_{k\Delta}) = f(\dot{X}_{k\Delta})$, where we recall $\dot{X} = X \mod 1$. Denote by $\dot{p}_b(t,y,z)$ the transition densities of $\dot{X}$, i.e. $\dot{p}_b(t,y,z) = \sum_{j\in\mathbb{Z}} p_b(t,y,z+j)$ (see the proof of Proposition 9 in Nickl & Söhl [62] for an argument that the sum converges). Theorem 2.6 in Bhattacharya et al. [8] tells us that if $\dot{X}_0$ has a density η_0 on $[0,1]$, then $\dot{X}_t$ has a density η_t satisfying

$$\|\eta_t - \pi_b\|_{\mathrm{TV}} \leq \tfrac{1}{2}\|\eta_0/\pi_b - 1\|_{\mathrm{TV}} \exp\Big(-\tfrac{1}{2M_b}t\Big),$$

where $M_b := \sup_{z\in[0,1]}\Big\{(\sigma^2(z)\pi_b(z))^{-1}\int_0^z \pi_b(x)\,\mathrm{d}x \int_z^1 \pi_b(y)\,\mathrm{d}y\Big\}$ (writing, in an abuse of notation, $\|p-q\|_{\mathrm{TV}}$ for the total variation distance between measures with densities p,q). We can regularise to extend the result so that it also applies when the initial distribution of $\dot{X}$ is a point mass: if $\dot{X}_0 = y$ then $\dot{X}_1$ has density $\dot{p}_b(1,y,\cdot)$, hence the result applies to show

$$\|\dot{p}_b(t,y,\cdot) - \pi_b\|_{\mathrm{TV}} \leq \tfrac{1}{2}\|\dot{p}_b(1,y,\cdot)/\pi_b - 1\|_{\mathrm{TV}} \exp\Big(-\tfrac{1}{2M_b}(t-1)\Big).$$

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

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

$$\|\dot{p}_b(t,y,\cdot) - \pi_b\|_{\mathrm{TV}} \leq K\exp\Big(-\tfrac{1}{2M}t\Big)$$

for some constant $K = K(\mathcal{I})$, uniformly across $y \in [0,1]$. It follows that, for each fixed Δ, the discrete time Markov chain $(\dot{X}_{k\Delta})_{k\geq 0}$ is uniformly ergodic with mixing time $t_{\mathrm{mix}} \leq 1 + 2M\log(4K)\Delta^{-1} \leq K'\Delta^{-1}$ for some constant K'. Theorem 2.7 applies to tell us

$$P_b\Big(|\sum_{i=1}^n f(X_{k\Delta}) - E_\mu[f]| \geq s\Big) \leq 2\exp\Big(-\frac{s^2}{2K'\Delta^{-1}(8(n+2K'\Delta^{-1})V_f + 20sC)}\Big).$$

Since $n\Delta \to \infty$ by assumption, we see $8(n+2K'\Delta^{-1}) \leq K''n$ for some constant K''. Using the bound $2/(a+b) \geq \min(1/a, 1/b)$ for $a,b>0$ and upper bounding the centred moments V_f and C by the uncentred moments $\|f\|_\mu^2$ and $\|f\|_\infty$, we deduce (2.8).

The result (2.9) is obtained by a change of variables. For the supremum result (2.10), we use a standard chaining argument, e.g. as in Baraud [6] Theorem 2.1, where we use

(2.9) in place of Baraud's Assumption 2.1, noting that Baraud only uses Assumption 2.1 to prove an expression mirroring (2.9), and the rest of the proof follows through exactly. Precisely, following the proof, we can take $\tilde{\kappa} = 36\kappa$. □

Remark. The proof simplifies if we consider only those b satisfying $I_b(1) = 0$. In this case, the invariant density (upon changing normalising constant to some G_b) reduces to the more familiar form $\pi_b(x) = (G_b\sigma^2(x))^{-1}e^{I_b(x)}$. The diffusion is moreover reversible under this condition, so we can use Theorem 3.3 from [66] instead of Theorem 3.4 to attain the same results but with better constants.

Hölder continuity properties of diffusions

Define

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

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

Lemma 2.8. *Let X solve the scalar diffusion equation (2.1), and grant Assumptions 1 and 2. Then there exist positive constants λ, C and τ, all depending on $\mathcal{I}$ only, such that for any $u > C\max(\log(m), 1)^{1/2}$ and any initial value x,*

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

Remarks.

i. We will need to control all increments $X_{(j+1)\Delta} - X_{j\Delta}$ simultaneously, hence we include the parameter m, which we will take to be the time horizon $n\Delta$ when applying this result. Simply controlling over $[0, 1]$ and using a union bound does not give sharp enough results.

ii. The lemma applies for any distribution of X_0, not just point masses, by an application of the tower law for conditional expectation.

The modulus of continuity w_m matches that of Brownian motion, and indeed the proof, given in Appendix 2.B, is to reduce to the corresponding result for Brownian motion. First, by applying the scale function one transforms X into a local martingale, reducing Lemma 2.8 to the following result, also useful in its own right.

Lemma 2.9. *Let Y be a local martingale with quadratic variation satisfying $|\langle Y\rangle_t - \langle Y\rangle_s| \leq A|t - s|$ for a constant $A \geq 1$. Then there exist positive constants $\lambda = \lambda(A)$ and*

$C = C(A)$ such that for any $u > C\max(\log(m), 1)^{1/2}$,

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

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

Lemma 2.9 follows from the corresponding result for Brownian motion by a time change (i.e. the (Dambis–)Dubins-Schwarz Theorem). It is well known that Brownian motion has modulus of continuity $\delta^{1/2}(\log\delta^{-1})^{1/2}$ in the sense that there almost surely exists a constant $C > 0$ such that $|B_t - B_s| \leq C|t-s|^{1/2}(\log(|t-s|^{-1}))^{1/2}$, for all $t, s \in [0,1]$ sufficiently close, but Lemmas 2.8 and 2.9 depend on the following quantitative version of this statement, proved using Gaussian process techniques. The proofs of Lemmas 2.9 and 2.10 are given in Appendix 2.B.

Lemma 2.10. *Let B be a standard Brownian motion on $[0,m]$. There are positive (universal) constants λ and C such that for $u > C\max(\log(m), 1)^{1/2}$,*

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

2.3.2 Proof of the estimator concentration result Theorem 2.5

It is enough to show that, uniformly across $b \in \Theta_n$, for any $D > 0$ there is a $C > 0$ such $P_b\big(\|\tilde{b}_n - b\|_2 > C\varepsilon_n\big) \leq 16e^{-Dn\Delta\varepsilon_n^2}$, because by initially considering a $D' > D$ and finding the corresponding C', we can eliminate the factor of 16 in front of the exponential.

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

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

$$I_n = \Big\{t \in S_{l_n} \text{ s.t. } \|t\|_\mu = 1, \|t\|_\infty \leq C_1\varepsilon_n^{-1}\Big\}, \tag{2.11}$$

$$\mathcal{A}_n = \{\tilde{t}_n \in I_n\} \cup \{\tilde{t}_n = 0\}, \tag{2.12}$$

$$\Omega_n = \Big\{\Big|\|t\|_n^2 - 1\Big| \leq \tfrac{1}{2},\ \forall t \in I_n\Big\}, \tag{2.13}$$

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

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

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

1. Show the theorem holds (deterministically) on $\mathcal{A}_n^c$, for a large enough constant C.
2. Show that $P_b(\Omega_n^c) \leq 4e^{-Dn\Delta\varepsilon_n^2}$ for a suitable choice of C_1.
3. Show that, for any D, we can choose a C such that $P_b\Big(\|\tilde{b}_n - b\|_2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C\varepsilon_n\Big) \leq 12e^{-Dn\Delta\varepsilon_n^2}$.

Step 1: Intuitively we reason thus. The event $\mathcal{A}_n^c$ can only occur if the $L^2(\mu)$–norm of $\tilde{b}_n - \pi_{l_n} b$ is small compared to the L^∞–norm. Since we have assumed a uniform supremum bound on functions $b \in \Theta$, in fact $\mathcal{A}_n$ holds unless the $L^2(\mu)$–norm is small in absolute terms. But if $\|\tilde{b}_n - \pi_{l_n} b\|_\mu$ is small, then so is $\|\tilde{b}_n - b\|_2$. We formalise this reasoning now.

For a constant C_2 to be chosen, define

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

On $\mathcal{A}_n'$ we have $\|\tilde{t}_n\|_\infty \leq (\|\tilde{b}_n\|_\infty + \|\pi_{l_n} b\|_\infty) C_2^{-1}\varepsilon_n^{-1}$. Note $\|\tilde{b}_n\|_\infty \leq K_0 + 1$ by definition. Since, for n large enough, $\|\pi_{l_n} b - b\|_\infty \leq 1$ uniformly across $b \in \Theta_n \subseteq \Theta$ by (2.4) so that $\|\pi_{l_n} b\|_\infty \leq \|b\|_\infty + 1 \leq K_0 + 1$, we deduce that on $\mathcal{A}_n'$, $\|\tilde{t}_n\|_\infty \leq (2K_0 + 2) C_2^{-1}\varepsilon_n^{-1}$. Since also $\|\tilde{t}_n\|_\mu = 1$ (or $\tilde{t}_n = 0$) by construction, we deduce $\mathcal{A}_n' \subseteq \mathcal{A}_n$ if $C_2 \geq C_1^{-1}(2K_0 + 2)$.

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

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

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

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

For $t \in I_n$ we have $\Big|\|t\|_n^2 - 1\Big| = n^{-1}\Big|\sum_{k=1}^n t^2(X_{k\Delta}) - E_\mu[t^2]\Big|$. Thus Theorem 2.6 can be applied to $\Omega_n^c = \{\sup_{t \in I_n} n^{-1}|\sum_{k=1}^n t^2(X_{k\Delta}) - E_\mu[t^2]| > 1/2\}$. Each $t \in I_n$ has $\|t^2\|_\infty \leq C_1^2\varepsilon_n^{-2}$ and $\|t^2\|_\mu^2 = E_\mu[t^4] \leq \|t^2\|_\infty \|t\|_\mu^2 \leq C_1^2\varepsilon_n^{-2}$. Since the indexing set I_n lies

in a vector space of dimension D_{l_n}, we apply the theorem with $x = Dn\Delta\varepsilon_n^2$ to see

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

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

Step 3: Since $b \in \Theta_n$ and π_{l_n} is L^2–orthogonal projection, we have $\|\tilde b_n - b\|_2^2 \leq \|\tilde b_n - \pi_{l_n} b\|_2^2 + \varepsilon_n^2$. Recall that $\|\cdot\|_2 \leq \pi_L^{-1/2}\|\cdot\|_\mu$ and note that on $\mathcal{A}_n \cap \Omega_n$, we further have $\frac{1}{2}\|\tilde b_n - \pi_{l_n} b\|_\mu^2 \leq \|\tilde b_n - \pi_{l_n} b\|_n^2$.

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

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

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

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

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

Bounding $P_b(\|\pi_{l_n} b - b\|_n > C\varepsilon_n)$: We show that for any $D > 0$ there is a constant C such that $P_b(\|\pi_{l_n} b - b\|_n > C\varepsilon_n) \leq 2e^{-Dn\Delta\varepsilon_n^2}$, for all n sufficiently large. Since $E_b\|g\|_n^2 = \|g\|_\mu^2$ for any 1–periodic deterministic function g and $\|\pi_{l_n} b - b\|_\mu^2 \leq \pi_U\|\pi_{l_n} b - b\|_2^2 \leq \pi_U\varepsilon_n^2$ for $b \in \Theta_n$, it is enough to show that

$$P_b\Big(\Big|\|\pi_{l_n} b - b\|_n^2 - E_b\|\pi_{l_n} b - b\|_n^2\Big| > C\varepsilon_n^2\Big) \leq 2e^{-Dn\Delta\varepsilon_n^2} \tag{2.14}$$

for some different C. As in Step 2, we apply Theorem 2.6, but now working with the single function $(\pi_{l_n} b - b)^2$. For large enough n we have the bounds $\|\pi_{l_n} b - b\|_\infty \leq 1$ (derived from (2.4)), and $\|(\pi_{l_n} b - b)^2\|_\mu \leq \|\pi_{l_n} b - b\|_\infty\|\pi_{l_n} b - b\|_\mu \leq \pi_U^{1/2}\varepsilon_n$ (because $b \in \Theta_n$) and so applying the theorem with $x = Dn\Delta\varepsilon_n^2$ gives

$$P_b\Bigg(\Bigg|\sum_{k=1}^n\Big[(\pi_{l_n} b - b)^2(X_{k\Delta}) - \|\pi_{l_n} b - b\|_\mu^2\Big]\Bigg| \geq \max\{a,b\}\Bigg) \leq 2e^{-Dn\Delta\varepsilon_n^2},$$

for $a = \sqrt{\kappa n\Delta^{-1}\pi_U\varepsilon_n^2 Dn\Delta\varepsilon_n^2} = n\varepsilon_n^2\sqrt{\kappa\pi_U D}$ and $b = \kappa\Delta^{-1}Dn\Delta\varepsilon_n^2 = n\varepsilon_n^2\kappa D$, for some constant $\kappa = \kappa(\mathcal{I})$. We see that a/n and b/n are both upper bounded by a constant multiple of ε_n^2, hence, by choosing C large enough, (2.14) holds.

Bounding $P_b\Big(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n\cap\Omega_n} > C\varepsilon_n^2\Big)$: We show that $P_b\Big(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n\cap\Omega_n} > C\varepsilon_n^2\Big) \leq 10e^{-Dn\Delta\varepsilon_n^2}$ for some constant C.

Recall an application of (2.4) showed us that $\|\pi_{l_n}b\|_\infty \leq K_0 + 1$ for sufficiently large n, hence we see that $\pi_{l_n}b$ lies in $\tilde{S}_{l_n}$, so by definition $\gamma_n(\tilde{b}_n) \leq \gamma_n(\pi_{l_n}b)$. We now use this to show that

$$\frac{1}{4}\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n\cap\Omega_n} \leq \frac{7}{4}\|\pi_{l_n}b - b\|_n^2 + 8\nu_n(\tilde{t}_n)^2\mathbb{1}_{\mathcal{A}_n} + \frac{8}{n}\sum_{k=1}^n R_{k\Delta}^2, \tag{2.15}$$

where $\nu_n(t) = \frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})Z_{k\Delta}$ and we recall that $\tilde{t}_n = (\tilde{b}_n - \pi_{l_n}b)\|\tilde{b}_n - \pi_{l_n}b\|_\mu^{-1}$. The argument, copied from [21] Sections 3.2 and 6.1, is as follows. Using $\Delta^{-1}(X_{(k+1)\Delta} - X_{k\Delta}) = b(X_{k\Delta}) + Z_{k\Delta} + R_{k\Delta}$ and $\gamma_n(\tilde{b}_n) - \gamma_n(b) \leq \gamma_n(\pi_{l_n}b) - \gamma_n(b)$, one shows that

$$\|\tilde{b}_n - b\|_n^2 \leq \|\pi_{l_n}b - b\|_n^2 + 2\nu(\tilde{b}_n - \pi_{l_n}b) + \tfrac{2}{n}\sum_{k=1}^n R_{k\Delta}(\tilde{b}_n - \pi_{l_n}b)(X_{k\Delta}). \tag{2.16}$$

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

$$\frac{2}{n}\sum_{k=1}^n R_{k\Delta}(\tilde{b}_n - \pi_{l_n}b)(X_{k\Delta}) \leq \frac{8}{n}\sum_{k=1}^n R_{k\Delta}^2 + \frac{1}{8}\|\tilde{b}_n - \pi_{l_n}b\|_n^2,$$
$$2\nu(\tilde{b}_n - \pi_{l_n}b) = 2\|\tilde{b}_n - \pi_{l_n}b\|_\mu \nu(\tilde{t}_n) \leq 8\nu_n(\tilde{t}_n)^2 + \tfrac{1}{8}\|\tilde{b}_n - \pi_{l_n}b\|_\mu^2.$$

Next recall that on $\mathcal{A}_n \cap \Omega_n$, we have $\|\tilde{b}_n - \pi_{l_n}b\|_\mu^2 \leq 2\|\tilde{b}_n - \pi_{l_n}b\|_n^2$, and further recall $\|\tilde{b}_n - \pi_{l_n}b\|_n^2 \leq 2\|\tilde{b}_n - b\|_n^2 + 2\|\pi_{l_n}b - b\|_n^2$. Putting all these bounds into (2.16) yields (2.15), where on the right hand side we have only included indicator functions where they will help us in future steps. Next, by a union bound, we deduce

$$\begin{aligned}
&P_b(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n\cap\Omega_n} > C\varepsilon_n^2) \\
&\qquad \leq P_b(\|\pi_{l_n}b - b\|_n^2 > C'\varepsilon_n^2) + P_b(\nu_n(\tilde{t}_n)^2\mathbb{1}_{\mathcal{A}_n} > C'\varepsilon_n^2) + P_b\Big(\frac{1}{n}\sum_{k=1}^n R_{k\Delta}^2 > C'\varepsilon_n^2\Big),
\end{aligned}$$

for some constant C' (we can take $C' = C/96$). We have already shown that $P_b(\|\pi_{l_n}b - b\|_n > C\varepsilon_n) \leq 2e^{-Dn\Delta\varepsilon_n^2}$ for a large enough constant C, thus, since $\tilde{t}_n \in I_n$ on the event $\mathcal{A}_n$, the following two lemmas conclude the proof.

Lemma 2.11. *Under the conditions of Theorem 2.5, for each $D > 0$ there exists a constant $C = C(\mathcal{I}, L_0, D) > 0$ for which, for n sufficiently large, $P_b\Big(\frac{1}{n}\sum_{k=1}^n R_{k\Delta}^2 > C\varepsilon_n^2\Big) \leq 2e^{-Dn\Delta\varepsilon_n^2}$.*

Lemma 2.12. *Under the conditions of Theorem 2.5 and with I_n defined as in (2.11), for each $D > 0$ there exists a constant $C = C(\mathcal{I}, L, D) > 0$ for which, for n sufficiently large, $P_b(\sup_{t\in I_n}(\nu_n(t)) > C\varepsilon_n) \leq 6e^{-Dn\Delta\varepsilon_n^2}$.*

Proof of Lemma 2.11. Recall $R_{k\Delta} = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(b(X_s) - b(X_{k\Delta}))\,\mathrm{d}s$, and recall any $b \in \Theta$ is Lipschitz, with Lipschitz constant at most K_0, so $|R_{k\Delta}| \leq K_0 \max_{s\leq\Delta}|X_{k\Delta+s} - X_{k\Delta}|$. It is therefore enough to bound $\sup\{|X_t - X_s| : s, t \in [0, n\Delta],\ |t - s| \leq \Delta\}$.

We apply the Hölder continuity result (Lemma 2.8) with $u = D^{1/2}\lambda^{-1/2}(n\Delta\varepsilon_n^2)^{1/2}$ for $\lambda = \lambda(\mathcal{I})$ the constant of the lemma, noting that the assumption $n\Delta\varepsilon_n^2/\log(n\Delta) \to \infty$ ensures that u is large enough compared to $m = n\Delta$ that the conditions for the lemma are met, at least when n is large. We see that

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

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

Now observe that $\log(n\Delta)^{1/2} \leq (\log(\Delta^{-1})^{1/2})$ for large enough n because $n\Delta^2 \to 0$ (so $n\Delta \leq \Delta^{-1}$ eventually). Further, from the assumption $n\Delta^2\log(\Delta^{-1}) \leq L_0$ we are able to deduce that $\Delta^{1/2}\log(\Delta^{-1})^{1/2}(n\Delta\varepsilon_n^2)^{1/2} \leq L_0^{1/2}\varepsilon_n$. It follows that on $\mathcal{D}$, we have $R_{k\Delta} \leq C\varepsilon_n$ for a suitably chosen constant C (independent of k and n), which implies the desired concentration. □

Proof of Lemma 2.12. Recall the definitions

$$Z_{k\Delta} = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)\,\mathrm{d}W_s, \qquad \nu_n(t) = \frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})Z_{k\Delta}.$$

The martingale-derived concentration result Lemma 2 in Comte et al. [21] (the model assumptions in [21] are slightly different to those made here, but the proof of the lemma equally applies in our setting) tells us $P_b(\nu_n(t) \geq \xi, \|t\|_n^2 \leq u^2) \leq \exp\Big(-\frac{n\Delta\xi^2}{2\sigma_U^2u^2}\Big)$, for any t, u, and for any drift function $b \in \Theta$, so that

$$P_b(\nu_n(t) \geq \xi) \leq \exp\left(-\frac{n\Delta\xi^2}{2\sigma_U^2u^2}\right) + P_b(\|t\|_n^2 > u^2). \qquad (\star)$$

We can apply Theorem 2.6 to see that, for some constant $\kappa = \kappa(\mathcal{I})$,

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

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

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

$$\begin{aligned} P_b(\nu_n(t) \ge \xi) &\le \exp\Big(-\frac{n\Delta}{4\sigma_U^2}\min(\xi^2\|t\|_\mu^{-2}, \xi\|t\|_\infty^{-1})\Big) + 2\exp\big(-\tfrac{1}{\kappa}n\Delta\min(\xi^2\|t\|_\mu^{-2}, \xi\|t\|_\infty^{-1})\big) \\ &\le 3\exp\big(-\tfrac{1}{\kappa'}n\Delta\min(\xi^2\|t\|_\mu^{-2}, \xi\|t\|_\infty^{-1})\big), \end{aligned}$$

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

By changing variables we attain the bound $P_b(\nu_n(t) \ge \max(\sqrt{v^2x}, ux)) \le 3\exp(-x)$, where $v^2 = \kappa'(n\Delta)^{-1}\|t\|_\mu^2$ and $u = \kappa'(n\Delta)^{-1}\|t\|_\infty$. Then, as in Theorem 2.6, a standard chaining argument allows us to deduce that

$$P_b\Big(\sup_{t\in I_n}\nu_n(t) \ge \tilde{\kappa}\Big(\sqrt{V^2(D_{l_n}+x)} + U(D_{l_n}+x)\Big)\Big) \le 6e^{-x},$$

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

2.4 Small ball probabilities

Now we show that the Kullback–Leibler divergence between the laws corresponding to different parameters b_0, b can be controlled in terms of the L^2–distance between the parameters, so that the 'small ball condition' (condition (iii) of Theorem 1.3) can be made more explicit. Recall from Chapter 1 that $K(p,q)$ denotes the Kullback–Leibler divergence between probability distributions with densities p and q, i.e. $K(p,q) =$

$E_p \log(\frac{p}{q}) = \int \log(\frac{p(x)}{q(x)})p(x)\,\mathrm{d}x$. Also write

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

Recalling that $p_b^{(n)}(x^{(n)}) = \pi_b(x_0)\prod_{i=1}^n p_b(\Delta, x_{(i-1)\Delta}, x_{i\Delta})$ is the density on $\mathbb{R}^{n+1}$ of $X^{(n)}$ under P_b, two Kullback–Leibler type neighbourhoods akin to the one defined in (1.12) and appropriate to the high-frequency setting considered here are defined for $\varepsilon > 0$ as

$$B_{KL}^{(n)}(\varepsilon) = \Big\{b \in \Theta : K(p_0^{(n)}, p_b^{(n)}) \le (n\Delta + 1)\varepsilon^2,\ \ \mathrm{Var}_{b_0}\Big(\log\frac{p_0^{(n)}}{p_b^{(n)}}\Big) \le (n\Delta + 1)\varepsilon^2\Big\},$$
$$B_\varepsilon = \Big\{b \in \Theta : K(\pi_0, \pi_b) \le \varepsilon^2, \mathrm{Var}_{b_0}(\log\tfrac{\pi_0}{\pi_b}) \le \varepsilon^2, \mathrm{KL}(b_0, b) \le \Delta\varepsilon^2, \mathrm{Var}_{b_0}(\log\tfrac{p_0}{p_b}) \le \Delta\varepsilon^2\Big\}.$$

Note that $\mathrm{KL}(b_0, b)$ and B_ε implicitly depend on n via Δ.

The main result of this section is the following.

Theorem 2.13. *Consider data $X^{(n)} = (X_{k\Delta})_{0\le k\le n}$ sampled from a solution X to (2.1) under Assumptions 1 to 4. Let $\varepsilon_n \to 0$ be a sequence of positive numbers such that $n\Delta\varepsilon_n^2 \to \infty$. Then there is a constant $A = A(\mathcal{I})$ such that, for all n sufficiently large, $\{b \in \Theta \text{ s.t. } \|b - b_0\|_2 \le A\varepsilon_n\} \subseteq B_{KL}^{(n)}(\varepsilon_n)$.*

Proof. Apply Lemma 2.21 from Appendix 2.A, which says that

$$\mathrm{Var}_{b_0}\log\Big(\frac{p_0^{(n)}(X^{(n)})}{p_b^{(n)}(X^{(n)})}\Big) \le 3\,\mathrm{Var}_{b_0}\Big(\log\frac{\pi_0(X_0)}{\pi_b(X_0)}\Big) + 3n\,\mathrm{Var}_{b_0}\Big(\log\frac{p_0(X_0, X_\Delta)}{p_b(X_0, X_\Delta)}\Big);$$

noting also that $K(p_0^{(n)}, p_b^{(n)}) = K(\pi_0, \pi_b) + n\,\mathrm{KL}(b_0, b)$ by linearity, we observe that $B_{\varepsilon_n/\sqrt{3}} \subseteq B_{KL}^{(n)}(\varepsilon_n)$. It is therefore enough to show that for some $A = A(\mathcal{I})$ we have $\{b \in \Theta \text{ s.t. } \|b - b_0\|_2 \le A\varepsilon_n\} \subseteq B_{\varepsilon_n/\sqrt{3}}$. This follows immediately by applying Lemma 2.14 below to $\xi_n = \varepsilon_n/\sqrt{3}$. □

Lemma 2.14. *Under the conditions of Theorem 2.13, there is an $A = A(\mathcal{I})$ such that, for all n sufficiently large, $\{b \in \Theta \text{ s.t. } \|b - b_0\|_2 \le A\varepsilon_n\} \subseteq B_{\varepsilon_n}$.*

The key idea in proving Lemma 2.14 is to use the Kullback–Leibler divergence between the laws $P_{b_0}^{(x)}, P_b^{(x)}$ of the continuous-time paths to control the Kullback–Leibler divergence between p_b and p_0. This will help us because we can calculate the Kullback–Leibler divergence between the full paths using Girsanov's Theorem, which gives us an explicit formula for the likelihood ratios.

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

Theorem 2.15 (Girsanov's Theorem)**.** *Suppose b_0 and b lie in Θ, and σ satisfies Assumption 1. Then the laws $P_{b_0,T}^{(x)}$ and $P_{b,T}^{(x)}$ are mutually absolutely continuous with, for $X \sim P_{b,T}^{(x)}$, the almost sure identification*

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

Proof. See Liptser & Shiryaev [54], Theorem 7.19, noting that the assumptions are met because b, b_0 and σ are all Lipschitz and bounded, and σ is bounded away from 0. □

We write

$$\tilde{p}_0^{(x)} = \frac{\mathrm{d}P_{b_0}^{(x)}}{\mathrm{d}\mathbb{W}_\sigma^{(x)}}, \qquad \tilde{p}_b^{(x)} = \frac{\mathrm{d}P_b^{(x)}}{\mathrm{d}\mathbb{W}_\sigma^{(x)}} \tag{2.17}$$

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

Proof of Lemma 2.14. We break the proof into a series of lemmas. We will upper bound the variances in the definition of B_{ε_n} by the corresponding uncentred second moments. For some constant $A = A(\mathcal{I})$ we show the following.

1. $A^2\,\mathrm{KL}(b_0, b) \le \Delta\|b - b_0\|_2^2$, which shows that $\mathrm{KL}(b_0, b) \le \Delta\varepsilon_n^2$ whenever $\|b - b_0\|_2 \le A\varepsilon_n$. This is the content of Lemma 2.16.

2. If $\|b - b_0\|_2 \le A\varepsilon_n$ then we have $E_{b_0}[\log(p_0/p_b)^2] \le \Delta\varepsilon_n^2$. This is the content of Lemma 2.17. Note that the other steps do not need any assumptions on ε_n, but this step uses that $n\Delta\varepsilon_n^2 \to \infty$.

3. $A^2 \max\{K(\pi_0, \pi_b), E_{b_0}[\log(\pi_0/\pi_b)^2]\} \le \|b_0 - b\|_2^2$. From this it follows that $K(\pi_0, \pi_b) \le \varepsilon_n^2$ and $E_{b_0}[\log(\pi_0/\pi_b)^2] \le \varepsilon_n^2$ whenever $\|b - b_0\|_2 \le A\varepsilon_n$. This is the content of Lemma 2.18.

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

Lemma 2.16. *Under the conditions of Theorem 2.13, there is a constant A depending only on $\mathcal{I}$ such that $A^2\,\mathrm{KL}(b_0, b) \leq \Delta\|b_0 - b\|_2^2$.*

The proof is essentially the same as that in van der Meulen & van Zanten [83] Lemma 5.1, with minor adjustments to fit the periodic model and non-constant σ used here. Further, all the ideas needed are exhibited in the proof of Lemma 2.17. Thus, we omit the proof.

Lemma 2.17. *Under the conditions of Theorem 2.13, there is a constant $A = A(\mathcal{I})$ so that, for n sufficiently large, $E_{b_0}[\log(p_0/p)^2] \leq \Delta\varepsilon_n^2$ whenever $\|b - b_0\|_2 \leq A\varepsilon_n$.*

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

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

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

$$\frac{p_0(\Delta, x, y)}{p_*(\Delta, x, y)} = E_{\mathbb{W}_\sigma^{(x)}}[\tilde{p}_0(U) \mid U_\Delta = y]. \tag{2.18}$$

The above argument is not rigorous because we condition on an event of probability zero, but the formula (2.18) is true, and is carefully justified in Lemma 2.22 in Appendix 2.A. A corresponding expression holds for $p_b(\Delta, x, y)$, so that

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

Lemma 2.20 in Appendix 2.A allows us to simplify the ratio of conditional expectations. We apply with $\mathbb{P} = \mathbb{W}_\sigma^{(X_0)}$, $\mathbb{Q} = P_{b_0}^{(X_0)}$ and $g = \tilde{p}_b^{(X_0)}/\tilde{p}_0^{(X_0)}$, then further apply conditional

Jensen's inequality and the tower law to find

$$E_{b_0}\Big[\Big(\log\frac{p_0}{p_b}\Big)^2\Big] \le E_{b_0}\Big[h\Big(E_{P_{b_0}^{(X_0)}}\Big[\frac{\tilde p_b}{\tilde p_0}(X)\mid X_\Delta\Big]\Big)\Big] \le E_{b_0}\Big[h\Big(\frac{\tilde p_b}{\tilde p_0}(X)\Big)\Big]$$
$$\le E_{b_0}\Big[\Big(\log\frac{\tilde p_0}{\tilde p_b}(X)\Big)^2\Big] + E_{b_0}\Big[(2e^{-1}\frac{\tilde p_b}{\tilde p_0}(X)-1)\mathbb{1}\Big\{\frac{\tilde p_b}{\tilde p_0}(X)\ge e\Big\}\Big],$$

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

Showing $E_{b_0}\Big[(\log\frac{\tilde p_0}{\tilde p_b})^2\Big] \le \frac{1}{2}\Delta\varepsilon_n^2$: Write $f = \frac{b_0-b}{\sigma}$. Then we apply Girsanov's Theorem (Theorem 2.15) to find

$$E_{b_0}\Big[\Big(\log\frac{\tilde p_0}{\tilde p_b}(X)\Big)^2\Big] = E_{b_0}\Big[\Big(\int_0^\Delta f(X_t)\,\mathrm{d}W_t + \frac{1}{2}\int_0^\Delta f^2(X_t)\,\mathrm{d}t\Big)^2\Big],$$
$$\le 2E_{b_0}\Big[\Big(\int_0^\Delta f(X_t)\,\mathrm{d}W_t\Big)^2\Big] + \frac{1}{2}E_{b_0}\Big[\Big(\int_0^\Delta f^2(X_t)\,\mathrm{d}t\Big)^2\Big],$$

where we have used the Cauchy–Schwarz inequality to control the cross term.

For the first term on the right, we use Itô's isometry ([73] IV.27.5), Fubini's Theorem, periodicity of f and stationarity of μ_0 for the periodised process $\dot X = X \mod 1$ to find

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

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

$$E_{b_0}\Big[\Big(\log\frac{\tilde p_0}{\tilde p_b}(X)\Big)^2\Big] \le 3\Delta\|f\|_{\mu_0}^2 \le \tfrac{1}{2}A^{-2}\Delta\|b_0-b\|_2^2,$$

where Assumptions 1 and 2 allow us to upper bound $\|f\|_{\mu_0}$ by $\|b_0-b\|_2$, up to a constant depending only on $\mathcal{I}$. For $\|b_0-b\|_2 \le A\varepsilon_n$ we then have $E_{b_0}[(\log(\tilde p_b/\tilde p_0))^2] \le \Delta\varepsilon_n^2/2$.

Showing $E_{b_0}[(2e^{-1}\frac{\tilde p_b}{\tilde p_0}(X)-1)\mathbb{1}\{\frac{\tilde p_b}{\tilde p_0}(X)\ge e\}] \le \frac{1}{2}\Delta\varepsilon_n^2$: We have

$$E_{b_0}\Big[\Big(2e^{-1}\frac{\tilde p_b}{\tilde p_0}(X)-1\Big)\mathbb{1}\Big\{\frac{\tilde p_b}{\tilde p_0}(X)\ge e\Big\}\Big] \le 2e^{-1}P_b\Big[\frac{\tilde p_b}{\tilde p_0}\ge e\Big] \le P_b\Big[\log\Big(\frac{\tilde p_b}{\tilde p_0}(X)\Big)\ge 1\Big].$$

By the tower law it suffices to show $P_b^{(x)}[\log(\frac{\tilde{p}_b}{\tilde{p}_0}(X)) \geq 1] \leq \frac{1}{2}\Delta\varepsilon_n^2$ for each $x \in [0,1]$. Applying Girsanov's Theorem (Theorem 2.15) we have, for $f = (b_0 - b)/\sigma$, and for n large enough that $\Delta\|f\|_\infty^2 \leq 1$,

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

Write $M_t = \int_0^t -f(X_s)\,\mathrm{d}W_s$. Then, for $A = \max(1, (2K_0/\sigma_L)^2)$, since A uniformly upper bounds $\|f\|_\infty^2$ for $b \in \Theta$, we see that M is a martingale whose quadratic variation satisfies $|\langle M\rangle_t - \langle M\rangle_s| \leq A|t-s|$. Recalling that $w_1(\delta) = \delta^{1/2}\log(\delta^{-1})^{1/2}$, we apply Lemma 2.9 with $u = w_1(\Delta)^{-1}/2$ to yield that, for n large enough,

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

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

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

$$\Delta \leq \lambda\log(n)^{-2} \implies \Delta \leq \lambda(\log\Delta^{-1})^{-1}\log(n)^{-1} \implies \log(n) \leq \lambda\Delta^{-1}(\log\Delta^{-1})^{-1},$$

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

Lemma 2.18. *Under the conditions of Theorem 2.13, there is a constant A depending only on $\mathcal{I}$ such that $A^2\max\{K(\pi_0,\pi_b), E_{b_0}[\log(\pi_0/\pi_b)^2]\} \leq \|b_0 - b\|_2^2$.*

Proof. In view of the comment after Lemma 8.3 in [32], it suffices to prove that $h^2(\pi_0,\pi_b)\|\pi_0/\pi_b\|_\infty \leq C\|b - b_0\|_2^2$ for some $C = C(\mathcal{I})$, where h is the Hellinger distance between densities defined by $h^2(p,q) = \int(\sqrt{p} - \sqrt{q})^2$. Since π_0, π_b are uniformly bounded above and away from zero, we can absorb the term $\|\pi_0/\pi_b\|_\infty$ into the constant.

We initially prove pointwise bounds on the difference between the densities π_0, π_b. Recall we saw in Section 2.1 that, for $I_b(x) = \int_0^x \frac{2b}{\sigma^2}(y)\,\mathrm{d}y$, we have

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

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

$$D_1 = \frac{e^{I_b(x)}}{\sigma^2(x)}\Big|\frac{1}{H_b} - \frac{1}{H_{b_0}}\Big|\Big(e^{I_b(1)}\int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y\Big),$$
$$D_2 = \frac{|e^{I_b(x)} - e^{I_{b_0}(x)}|}{H_{b_0}\sigma^2(x)}\Big(e^{I_b(1)}\int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y\Big),$$
$$D_3 = \frac{e^{I_{b_0}(x)}}{H_{b_0}\sigma^2(x)}\Big|(e^{I_b(1)} - e^{I_{b_0}(1)})\int_x^1 e^{-I_b(y)}\,\mathrm{d}y\Big|,$$
$$D_4 = \frac{e^{I_{b_0}(x)}}{H_{b_0}\sigma^2(x)}\Big|e^{I_{b_0}(1)}\int_x^1 (e^{-I_b(y)} - e^{-I_{b_0}(y)})\,\mathrm{d}y + \int_0^x (e^{-I_b(y)} - e^{-I_{b_0}(y)})\,\mathrm{d}y\Big|.$$

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

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

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

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

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

2.5 Main contraction results: proofs

We now have the tools we need to apply Theorem 1.3 and derive contraction rates. Recall the definition

$$B_{KL}^{(n)}(\varepsilon) = \left\{ b \in \Theta \text{ s.t. } K(p_0^{(n)}, p_b^{(n)}) \leq (n\Delta + 1)\varepsilon^2, \mathrm{Var}_{b_0}\left(\log \frac{p_0^{(n)}}{p_b^{(n)}}\right) \leq (n\Delta + 1)\varepsilon^2 \right\}.$$

We have the following abstract contraction result, from which we deduce Theorem 2.1.

Theorem 2.19. *Consider data $X^{(n)} = (X_{k\Delta})_{0 \leq k \leq n}$ sampled from a solution X to (2.1) under Assumptions 1 to 4. Let the true parameter be b_0. Let $\varepsilon_n \to 0$ be a sequence of positive numbers and let l_n be a sequence of positive integers such that, for some constant L we have, for all n,*

$$D_{l_n} = 2^{l_n} \leq Ln\Delta\varepsilon_n^2, \quad \text{and} \quad n\Delta\varepsilon_n^2 / \log(n\Delta) \to \infty. \tag{2.19}$$

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

$$b_0 \in \Theta_n \subseteq \{b \in \Theta : \|\pi_{l_n} b - b\|_2 \leq \varepsilon_n\}, \tag{2.20}$$

where π_{l_n} is the L^2–orthogonal projection onto S_{l_n} as described in Section 2.1.1. Let $\Pi^{(n)}$ be a sequence of priors on Θ satisfying

(a) $\Pi^{(n)}(\Theta_n^c) \leq e^{-(2\zeta+8)n\Delta\varepsilon_n^2}$,

(b) $\Pi^{(n)}(B_{KL}^{(n)}(\varepsilon_n)) \geq e^{-\zeta n\Delta\varepsilon_n^2}$,

for some constant $\zeta > 0$. Then $\Pi^{(n)}\left(\{b \in \Theta : \|b - b_0\|_2 \leq M\varepsilon_n\} \mid X^{(n)}\right) \to 1$ in probability under the law P_{b_0} of X, for some constant $M = M(\mathcal{I}, L_0, \zeta, L)$.

In fact, since here we are not targeting a specific rate of convergence to 1 for the P_{b_0}–random variable $\Pi^{(n)}\left(\{b \in \Theta : \|b - b_0\|_2 \leq M\varepsilon_n\} \mid X^{(n)}\right)$, we can relax (a): it can be shown that any exponent $\zeta' > \zeta + 1$ suffices in place of $2\zeta + 8$.

Proof of Theorem 2.19. Given Theorem 2.5, this is immediate from Theorem 1.3, noting that the proof of Theorem 1.3 is unchanged if we replace the parameter n with the effective sample size $n\Delta$ used here. □

Proof of Theorem 2.1. **A:** We apply Theorem 2.19. The key idea which allows us to control the bias and obtain this adaptive result with a sieve prior is *undersmoothing*.

Specifically, when we prove the small ball probabilities, we do so by conditioning on the hyperprior choosing a resolution j_n which corresponds to the minimax rate $(n\Delta)^{-s/(1+2s)}$ rather than corresponding to the slower rate $(n\Delta)^{-s/(1+2s)}\log(n\Delta)^{1/2}$ at which we prove contraction. This logarithmic gap gives us the room we need to ensure we can achieve the bias condition (a) and the small ball condition (b) for the *same* constant ζ. The argument goes as follows.

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

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

where L is a constant to be chosen. Note that (2.19) holds by definition. Recall now from our choice of approximation spaces in Section 2.1.1 that for any $m \in \mathbb{N}$ we have $\|\pi_m b_0 - b_0\|_2 \leq K(s)\|b_0\|_{B^s_{2,\infty}} 2^{-ms}$. For any fixed L we therefore find that for n large enough, writing $K = K(b_0) = K(s)2^s\|b_0\|_{B^s_{2,\infty}}$, we have

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

Similarly, it can be shown that, with $A = A(\mathcal{I})$ the constant of the small ball result (Theorem 2.13) and for n large enough, we have $\|b_0 - \pi_{j_n} b_0\|_2 \leq A\varepsilon_n/2$.

Set $\Theta_n = \{b_0\} \cup (S_{l_n} \cap \Theta)$ and observe that the above calculations show that the bias condition (2.20) holds (since also for $b \in \Theta_n$, if $b \neq b_0$ we have $\|\pi_{l_n} b - b\|_2 = 0$).

Next, for the small ball condition (b), recall Theorem 2.13 tells us that $\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\} \subseteq B^{(n)}_{KL}(\varepsilon_n)$ for all n large enough. Thus it suffices to show, for some $\zeta > 0$ for which we can also achieve (a), that $\Pi(\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\}) \geq e^{-\zeta n\Delta\varepsilon_n^2}$. Using that $\|b - b_0\|_2 \leq \|b - \pi_{j_n} b_0\|_2 + \|\pi_{j_n} b_0 - b_0\|_2 \leq \|b - \pi_{j_n} b_0\|_2 + A\varepsilon_n/2$, and using our assumptions on h and Π_m, we see that

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

for some constant $C = C(\mathcal{I}, \beta_1, \alpha)$. Since $\log(\varepsilon_n^{-1}) = \frac{s}{1+2s}\log(n\Delta) - \frac{1}{2}\log\log(n\Delta) \leq \log(n\Delta)$, we deduce that $\Pi(\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\}) \geq B_1 e^{-C'n\Delta\bar{\varepsilon}_n^2\log(n\Delta)} = B_1 e^{-C'n\Delta\varepsilon_n^2}$,

with a different constant C'. Changing constant again to some $\zeta = \zeta(\mathcal{I}, \beta_1, B_1, \alpha)$, we absorb the B_1 factor into the exponential for large enough n.

For (a), since $\Pi(\Theta^c) = 0$ by assumption, we have $\Pi(\Theta_n^c) \leq \Pi(S_{l_n}^c) = \sum_{m=l_n+1}^{\infty} h(m)$. We have assumed that $h(m) \leq B_2 e^{-\beta_2 D_m}$, which ensures that the sum is at most a constant times $e^{-\beta_2 D_{l_n}} \leq e^{-\frac{1}{2}L\beta_2 n\Delta\varepsilon_n^2}$. For the $\zeta = \zeta(\mathcal{I}, \beta_1, B_1, \alpha)$ for which we proved (b) above, we can therefore choose L large enough to guarantee $\Pi(\Theta_n^c) \leq e^{-(2\zeta+8)n\Delta\varepsilon_n^2}$.

B: Let ε_n and j_n be as in the statement of the theorem and define l_n as above (here we can take $L = 1$). Similarly to before, we apply results from Section 2.1.1 to see

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

Set $\Theta_n = \Theta_s(A_0)$ for all n. Our assumptions then guarantee the bias condition (a) holds for any ζ (indeed, $\Pi^{(n)}(\Theta_n^c) = 0$). Thus it suffices to prove that there exists an ζ such that $\Pi^{(n)}(\{b \in \Theta_s(A_0) : \|b - b_0\|_2 \leq 3\varepsilon_n\}) \geq e^{-\zeta n\Delta\varepsilon_n^2}$, since we can absorb the factor of 3 into the constant M by applying Theorem 2.19 to $\xi_n = 3\varepsilon_n$.

The prior concentrates on $\Theta_s(A_0)$, so that we have $\Pi^{(n)}(\{b : \|\pi_{j_n} b - b\|_2 \leq \varepsilon_n\}) = 1$, and b_0 lies in $\Theta_s(A_0)$, so that $\|\pi_{j_n} b_0 - b_0\|_2 \leq \varepsilon_n$. Thus

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

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

2.5.1 Explicit priors: proofs

Proof of Proposition 2.2. We verify that the conditions of Theorem 2.1A are satisfied. Condition (i) holds by construction. The $B^s_{\infty,1}$ norm can be expressed as

$$\|f\|_{B^s_{\infty,1}} = |f_{-1,0}| + \sum_{l=0}^{\infty} 2^{l(s+1/2)} \max_{0 \leq k < 2^l} |f_{lk}|, \tag{2.21}$$

(see [38] Section 4.3) so that any b drawn from our prior lies in $B^1_{\infty,1}$ and satisfies the bound $\|b\|_{B^1_{\infty,1}} \leq (B+1)(2+\sum_{l\geq 1} l^{-2})$. It follows from standard Besov spaces results (e.g. [38] Proposition 4.3.20, adapted to apply to periodic Besov spaces) that $b \in C^1_{\text{per}}([0,1])$, with a C^1_{per}–norm bounded in terms of B. Thus $\Pi(\Theta) = 1$ for an appropriate choice of

K_0. We similarly see that $b_0 \in \Theta$. It remains to show that (ii) holds. We have

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

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

Proof of Proposition 2.3. We verify the conditions of Theorem 2.1B. Since $s > 1$ similarly to the proof of Proposition 2.2 we see $\Pi^{(n)}(\Theta) = 1$ and $b_0 \in \Theta$ for an appropriate choice of K_0. Observe also that for $A_0 = 2B + 2$ we have $\Pi^{(n)}(\Theta_s(A_0)) = 1$ by construction, and $b_0 \in \Theta_s(A_0)$ by Assumption 5, using the wavelet characterisation (2.2) of $\|\cdot\|_{B^s_{2,\infty}}$. Thus (I) holds and it remains to check (II).

Let $j_n \in \mathbb{N}$ be such that $j_n \le \bar{L}_n$, $2^{j_n} \sim (n\Delta)^{1/(1+2s)}$. Similarly to the proof of Proposition 2.2 we have

$$\Pi^{(n)}(\{b \in \Theta : \|\pi_{j_n} b - \pi_{j_n} b_0\|_2 \le \varepsilon_n\}) \ge \Pi^{(n)}(|u_{lk} - \beta_{lk}| \le \varepsilon_n/2 \ \ \forall l < j_n, \ \forall k < 2^l)$$
$$\ge (\varepsilon_n \alpha/2)^{D_{j_n}},$$

so we're done. □

Proof of Proposition 2.4. We include only the key differences to the previous proofs.

Adapting slightly the proof of Proposition 2.2, we see that H and H_0 both have $B^2_{\infty,1}$–norm bounded by $(B+1)(2 + \sum_{l\ge1} l^{-2})$. Since $\|b\|_{C^1_{\text{per}}} \le \frac{1}{2}\|\sigma^2\|_{C^1_{\text{per}}}(1 + \|H\|_{C^2_{\text{per}}})$ and using [38] Proposition 4.3.20, adapted to apply to periodic Besov spaces, to control $\|H\|_{C^2_{\text{per}}}$ by $\|H\|_{B^2_{\infty,1}}$, we see that for some constant $K_0 = K_0(B)$ we have $b_0 \in \Theta(K_0)$ and $\Pi^{(n)}(\Theta(K_0)) = 1$. From the wavelet characterisation

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

it can be seen that H and H_0 have Sobolev norm $\|\cdot\|_{B^{s+1}_{2,2}}$ bounded by some A_0', hence for some constant $K = K(A_0', s)$ we have $\|H - \pi_m H\|_{B^1_{2,2}} \le K2^{-ms}$ and similarly for H_0. Since the $B^{s+1}_{2,2}$ norm controls the $B^{s+1}_{2,\infty}$ norm, and we have assumed $\sigma^2 \in \Theta_{s+1}$, we

additionally see that $b_0 \in \Theta_s(A_0)$ and $\Pi^{(n)}(\Theta_s(A_0)) = 1$ for an appropriate constant A_0. Note that here we also depend on the assumption $\sigma^2 \in C^s$ to allow us to control $\|b\|_{B^s_{2,\infty}}$: Remark 1 on page 143 of Triebel [79] and Proposition 4.3.20 from [38] together tell us that $\|\sigma^2 H'\|_{B^s_{2,\infty}} \leq c\|\sigma^2\|_{C^s}\|H'\|_{B^s_{2,\infty}}$ for some constant $c = c(s)$, and similarly for H_0.

Observe, for $j_n \in \mathbb{N}$ such that $j_n \leq \bar{L}_n$ and $2^{j_n} \sim (n\Delta)^{1/(1+2s)}$,

$$\begin{aligned}\|\pi_{j_n} b - \pi_{j_n} b_0\|_2 \leq \|b - b_0\|_2 = \|\sigma^2(H' - H_0')/2\|_2 \leq \tfrac{1}{2}\sigma_U^2 \|H - H_0\|_{B^1_{2,2}} \\ \leq \tfrac{\sigma_U^2}{2}\Big(\|H - \pi_{j_n} H\|_{B^1_{2,2}} + \|H_0 - \pi_{j_n} H_0\|_{B^1_{2,2}} + \|\pi_{j_n} H - \pi_{j_n} H_0\|_{B^1_{2,2}}\Big).\end{aligned}$$

Now $\sigma_U^2\|H - \pi_{j_n} H\|_{B^1_{2,2}} \leq \sigma_U^2 K 2^{-j_n s} \leq C(n\Delta)^{-s/(1+2s)} \leq \frac{1}{2}(n\Delta)^{-s/(1+2s)}\log(n\Delta)^{1/2} = \frac{1}{2}\varepsilon_n$ for large enough n, and similarly for H_0.

Thus,

$$\begin{aligned}\Pi^{(n)}\Big(\Big\{b : \|\pi_{j_n} b - \pi_{j_n} b_0\|_2 \leq \varepsilon_n\Big\}\Big) &\geq \Pi^{(n)}\Big(\Big\{b : \|\pi_{j_n} H - \pi_{j_n} H_0\|_{B^1_{2,2}} \leq \tfrac{1}{2}\sigma_U^{-2}\varepsilon_n\Big\}\Big) \\ &\geq \Pi^{(n)}(|u_{lk} - \beta_{lk}| \leq \kappa\varepsilon_n \;\; \forall l < j_n, \; \forall k < 2^l),\end{aligned}$$

where the final inequality can be seen to hold from the wavelet representation of $\|\cdot\|_{B^1_{2,2}}$ (the constant κ can be taken to be $\kappa = \frac{1}{2}\sigma_U^{-2}(1 + (\sum_{k=0}^{\infty} 2^{-2l})^{1/2})^{-1} > \sigma_U^{-2}/6$). The small ball condition (II) follows from our updated assumptions. □

Appendix 2.A Technical lemmas

Lemma 2.20. *Let $\mathbb{Q}, \mathbb{P}$ be mutually absolutely continuous probability measures and write $f = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$. Then, for any measurable g and any sub–σ–algebra $\mathcal{G}$, $E_{\mathbb{Q}}[g \mid \mathcal{G}] = \frac{E_{\mathbb{P}}[fg|\mathcal{G}]}{E_{\mathbb{P}}[f|\mathcal{G}]}$.*

Proof. This follows straightforwardly using the characterisation of conditional expectation in terms of expectations against $\mathcal{G}$–measurable functions. Precisely, we recall that

$$E_{\mathbb{P}}[c(X)v(X)] = E_{\mathbb{P}}[u(X)v(X)] \tag{$\star$}$$

holds for any $\mathcal{G}$–measurable function v if $c(X) = E_{\mathbb{P}}[u(X) \mid \mathcal{G}]$ a.s., and conversely if $c(X)$ is $\mathcal{G}$–measurable and ($\star$) holds for any $\mathcal{G}$–measurable v then $c(X)$ is a version of the conditional expectation $E_{\mathbb{P}}[u(X)]$. For the converse statement it is in fact enough for ($\star$) to hold for all indicator functions $v = \mathbb{1}_A$, $A \in \mathcal{G}$.

Applying ($\star$) repeatedly we find, for $A \in \mathcal{G}$,

$$E_{\mathbb{P}}[E_{\mathbb{Q}}[g \mid \mathcal{G}]E_{\mathbb{P}}[f \mid \mathcal{G}]\mathbb{1}_A] = E_{\mathbb{P}}[fE_{\mathbb{Q}}[g \mid \mathcal{G}]\mathbb{1}_A] = E_{\mathbb{Q}}[E_{\mathbb{Q}}[g \mid \mathcal{G}]\mathbb{1}_A] = E_{\mathbb{Q}}[g\mathbb{1}_A] = E_{\mathbb{P}}[fg\mathbb{1}_A],$$

so that, since also $E_{\mathbb{Q}}[g \mid \mathcal{G}]E_{\mathbb{P}}[f \mid \mathcal{G}]$ is $\mathcal{G}$–measurable, it is (a version of) $E_{\mathbb{P}}[fg \mid \mathcal{G}]$, as required. □

Lemma 2.21. *The variance of the log likelihood ratio tensorises in this model, up to a constant. Precisely,*

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

Proof. We write $\log\left(\frac{p_0^{(n)}(X^{(n)})}{p_b^{(n)}(X^{(n)})}\right) = U + V + W$, where $U = \log\frac{\pi_0(X_0)}{\pi_b(X_0)}$ and

$$V = \sum_{\substack{1\leq k\leq n\\ k \text{ odd}}} \log\frac{p_0(\Delta, X_{(k-1)\Delta}, X_{k\Delta})}{p_b(\Delta, X_{(k-1)\Delta}, X_{k\Delta})}, \qquad W = \sum_{\substack{1\leq k\leq n\\ k \text{ even}}} \log\frac{p_0(\Delta, X_{(k-1)\Delta}, X_{k\Delta})}{p_b(\Delta, X_{(k-1)\Delta}, X_{k\Delta})}.$$

Note now that V and W are both sums are of *independent* terms since $(X_{k\Delta})_{k\leq n}$ is a Markov chain. We thus have

$$\mathrm{Var}_{b_0}(V) = \#\{1 \leq k \leq n : k \text{ odd}\}\,\mathrm{Var}_{b_0}\left(\log\frac{p_0(X_0, X_\Delta)}{p_b(X_0, X_\Delta)}\right),$$

and a corresponding result for W. Using $\mathrm{Var}(R+S+T) = \mathrm{Var}(R) + \mathrm{Var}(S) + \mathrm{Var}(T) + 2\,\mathrm{Cov}(R,S) + 2\,\mathrm{Cov}(S,T) + 2\,\mathrm{Cov}(T,R)$ and $2\,\mathrm{Cov}(R,S) \leq \mathrm{Var}(R) + \mathrm{Var}(S)$, one derives the elementary inequality $\mathrm{Var}(U+V+W) \leq 3(\mathrm{Var}(U) + \mathrm{Var}(V) + \mathrm{Var}(W))$. The result follows. □

Lemma 2.22. *Let $\tilde{p}_0$ be as in (2.17). Let $p^*(\Delta, x, y)$ be the density of transitions from x to y in time Δ for a process $U \sim \mathbb{W}_\sigma^{(x)}$. Then*

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

Proof. Let $U \sim \mathbb{W}_\sigma^{(x)}$ and let $\mathbb{W}_\sigma^{(x,y)}$ denote the law on $C([0,\Delta])$ of U conditional on $U_\Delta = y$. We define the conditional law rigorously via disintegration (e.g. see [69] Chapter 5, Theorem 9, applied to $\lambda = \mathbb{W}_\sigma^{(x)}$, $\mathcal{X} = C([0,\Delta])$ with the sup norm, $T((U_t)_{t\leq\Delta}) = U_\Delta$

and $\mu(\mathrm{d}y) = p^*(\Delta, x, y)\,\mathrm{d}y)$, so that

$$E_{\mathbb{W}_\sigma^{(x)}}[f(U)] = \int_{-\infty}^{\infty} p^*(\Delta, x, y) E_{\mathbb{W}_\sigma^{(x,y)}}[f(U)]\,\mathrm{d}y,$$

for all non-negative measurable functions f. Taking $f(U) = \tilde{p}_0(U)\mathbb{1}\{U_\Delta \in A\}$ for an arbitrary Borel set $A \subseteq \mathbb{R}$, we see

$$P_{b_0}^{(x)}(X_\Delta \in A) = \int_{-\infty}^{\infty} p^*(\Delta, x, y)\mathbb{1}\{y \in A\} E_{\mathbb{W}_\sigma^{(x,y)}}[\tilde{p}_0]\,\mathrm{d}y.$$

The result follows. □

Appendix 2.B Proofs for Section 2.3.1

Proof of Lemma 2.8. Set $Y_t = S(X_t)$, where

$$S(x) = \int_0^x \exp\Big(-\int_0^y \frac{2b}{\sigma^2}(z)\,\mathrm{d}z\Big)\,\mathrm{d}y$$

is the scale function, and let ψ be the inverse of S. Since S'' exists and is continuous, Itô's formula applies to yield

$$\mathrm{d}Y_t = \tilde{\sigma}(Y_t)\,\mathrm{d}W_t, \quad \tilde{\sigma}(y) := S'(\psi(y))\sigma(\psi(y)).$$

Let $A = A(\mathcal{I}) = \max(\sigma_U^2 \exp(4K_0/\sigma_L^2), 1)$ and observe that $\|\tilde{\sigma}^2\|_\infty \le A$. Thus, there are constants $C = C(\mathcal{I})$ and $\lambda = \lambda(\mathcal{I})$ so that for any $u > C\max(\log m, 1)^{1/2}$, the event

$$\mathcal{D} = \Big\{\sup\Big\{\frac{|Y_t - Y_s|}{w_m(|t-s|)} : s, t \in [0, m],\ s \ne t,\ |t-s| \le A^{-1}e^{-2}\Big\} \le u\Big\},$$

occurs with probability at least $1 - 2e^{-\lambda u^2}$, by Lemma 2.9. Now $X_t = \psi(Y_t)$ and ψ is Lipschitz with constant $\|\psi'\|_\infty = \|1/(S' \circ \psi)\|_\infty \le \exp(2K_0\sigma_L^{-2})$. It follows that on $\mathcal{D}$, writing $\tau = A^{-1}e^{-2}$, we have for any $s, t \in [0, m]$, $s \ne t$, $|t-s| \le \tau$,

$$|X_t - X_s| \le \exp(2K_0\sigma_L^{-2})|Y_t - Y_s| \le \exp(2K_0\sigma_L^{-2})w_m(|t-s|)u$$

The result follows by relabelling $(\exp(2K_0/\sigma_L^2)u) \mapsto u$, $\lambda \mapsto \lambda\exp(-4K_0/\sigma_L^2)$ and $C \mapsto C\exp(2K_0/\sigma_L^2)$. □

Proof of Lemma 2.9. Recall $w_m(\delta) := \delta^{1/2}(\log(\delta^{-1})^{1/2} + \log(m)^{1/2})$ for $m \ge 1$ and $w_m(\delta) := w_1(\delta)$ for $m < 1$. We see that we may assume $m \ge 1$ and the result for

$m < 1$ will follow. By the (Dambis–)Dubins-Schwarz Theorem (e.g. (34.1) in Rogers & Williams [73]), we can write $Y_t = Y_0 + B_{\eta_t}$ for B a standard Brownian motion and for $\eta_t = \langle Y \rangle_t$ the quadratic variation of Y. Define the event

$$\mathcal{C} = \Big\{ \sup\Big\{ \frac{|B_{t'} - B_{s'}|}{w_{Am}(|t' - s'|)} : \ s', t' \in [0, Am], \ s' \neq t', \ |t' - s'| \leq e^{-2} \Big\} \leq u \Big\}.$$

By Lemma 2.10, there are universal constants C and λ so that if $u > C \max(\log(Am), 1)^{1/2}$, $\mathcal{C}$ occurs with probability at least $1 - 2e^{-\lambda u^2}$, and note that by allowing C to depend on A we can replace $\max(\log(Am), 1)$ with $\max(\log(m), 1)$. On this event, for $s, t \in [0, m]$ with $|t - s| \leq A^{-1}e^{-2}$ and $s \neq t$ we have

$$\begin{aligned} |Y_t - Y_s| &= |B_{\eta_t} - B_{\eta_s}| \\ &\leq \sup\{|B_{t'} - B_{s'}| : \ s', t' \in [0, Am], \ s' \neq t', \ |t' - s'| \leq A|t - s|\} \\ &\leq u \sup\{w_{Am}(|t' - s'|) : \ s', t' \in [0, Am], \ s' \neq t', \ |t' - s'| \leq A|t - s|\} \\ &\leq w_{Am}(A|t - s|)u, \end{aligned}$$

where we have used that $w_{Am}(\delta)$ is increasing in the range $\delta \leq e^{-2}$ to attain the final inequality. Recalling we assume $A \geq 1$, one sees that $w_{Am}(A\delta) \leq A^{1/2} w_{Am}(\delta)$ provided $\delta \leq A^{-1}$, which holds in the relevant range. Thus, on $\mathcal{C}$, and for s, t and u in the considered ranges,

$$\begin{aligned} |Y_t - Y_s| &\leq A^{1/2} u |t - s|^{1/2} \Big((\log(Am))^{1/2} + (\log|t - s|^{-1})^{1/2} \Big) \\ &\leq A' u |t - s|^{1/2} \Big((\log(m))^{1/2} + (\log|t - s|^{-1})^{1/2} \Big), \end{aligned}$$

where A' is a constant depending on A (note we have absorbed a term depending on $\log(A)$ into the constant, using that $\log(|t - s|^{-1}) \geq 2$). The desired result follows upon relabelling $A'u \mapsto u$ since C and λ are here allowed to depend on A.

For the particular case $\mathrm{d}Y_t = \tilde{\sigma}(Y_t)\,\mathrm{d}W_t$, we simply observe that $|\langle Y \rangle_t - \langle Y \rangle_s| = |\int_s^t \tilde{\sigma}^2(Y_s)\,\mathrm{d}s| \leq \|\tilde{\sigma}^2\|_\infty |t - s|$. □

Proof of Lemma 2.10. Assume $m \geq 1$; the result for $m < 1$ follows. For a Gaussian process B, indexed by T and with intrinsic covariance (pseudo-)metric $d(s, t) = (E[(B_t - B_s)^2])^{1/2}$, Dudley's Theorem ([38] Theorem 2.3.8) says

$$E\Bigg[\sup_{s,t \in T, s \neq t} \frac{|B_t - B_s|}{\int_0^{d(s,t)} \sqrt{\log N(T, d, x)}\,\mathrm{d}x} \Bigg] < \infty,$$

where $N(T,d,x)$ is the number of (closed) balls of $d-$radius x needed to cover T. Inspecting the proof, it is in fact shown that the process

$$C_u = \frac{B_{u_2} - B_{u_1}}{\int_0^{d(u_1,u_2)} \sqrt{\log(N(T,d,x))}\,\mathrm{d}x} \quad \text{on} \quad U = \{u = (u_1,u_2) : u_1, u_2 \in T,\ d(u_1,u_2) \neq 0\},$$

is a Gaussian process on with bounded and continuous sample paths. It follows by [38] Theorem 2.1.20 that

$$\Pr\Big\{\Big|\sup_{v\in V}|C_v| - E\sup_{v\in V}|C_v|\Big| > u\Big\} \leq 2e^{-u^2/2\sigma^2},$$

for any subset V of U, where $\sigma^2 = \sup_{v\in V} E[C_v^2]$. We can upper bound C_v by applying the trivial lower bound for the denominator $\int_0^a \sqrt{\log N(T,d,x)} \geq \frac{a}{2}\sqrt{\log 2}$ for any $a = d(u,v)$ with $u,v \in T$ (this follows from the fact that $N(T,d,x) \geq 2$ if x is less than half the diameter of T). Using also that d is the intrinsic covariance metric, we deduce that $EC_v^2 \leq 4/\log 2$, so we can take $\sigma^2 = 4/\log 2$.

We will apply the result to B a standard Brownian motion on $T = [0,m]$, which has intrinsic covariance metric $d(s,t) = |t-s|^{1/2}$. For this T and d, we have $N(T,d,x) \leq mx^{-2}$. Then, applying Jensen's inequality, we see

$$\begin{aligned}\int_0^{d(s,t)} \sqrt{\log N(T,d,x)}\,\mathrm{d}x &\leq d(s,t)^{1/2}\Big(\int_0^{d(s,t)} \log(N(T,d,x)\Big)^{1/2} \\ &\leq 2^{1/2} d(s,t)[1 + \log(d(s,t)^{-1}) + \log m]^{1/2}.\end{aligned}$$

Set $V = \{u = (s,t) \in U : |t-s| \leq e^{-2}\}$ and observe that for $(s,t) \in V$ we have $1 + \log(d(s,t)^{-1}) = 1 + \frac{1}{2}\log(|t-s|^{-1}) \leq \log(|t-s|^{-1})$. Noting further that $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ for $a,b \geq 0$ and recalling we defined $w_m(\delta) = \delta^{1/2}((\log\delta^{-1})^{1/2} + \log(m)^{1/2})$, we see

$$\int_0^{d(s,t)} \sqrt{\log N(T,d,x)}\,\mathrm{d}x \leq 2^{1/2} w_m(|t-s|).$$

Thus, writing $M = E\Big[\sup\Big\{\frac{|B_t - B_s|}{\int_0^{d(s,t)} \sqrt{\log N(T,d,x)}\,\mathrm{d}x} : s,t \in T, s \neq t, |t-s| \leq e^{-2}\Big\}\Big]$ we see

$$\Pr\Big[\sup\Big\{\frac{|B_t - B_s|}{w_m(|t-s|)} : s,t \in T, s \neq t, |t-s| \leq e^{-2}\Big\} > 2^{1/2}(M+u)\Big] \leq 2e^{-(u^2(\log 2)/8)}.$$

As M is a fixed finite number, we can write $M + u = (1+\varepsilon)u$ with $\varepsilon \to 0$ as $u \to \infty$. Then

$$\Pr\Big[\sup_{\substack{s,t\in T, s\neq t,\\ |t-s|\leq e^{-2}}} \frac{|B_t - B_s|}{w_m(|t-s|)} > u\Big] \leq 2e^{-(u^2(\log 2)/16(1+\varepsilon)^2)}.$$

Thus provided u is larger than M, we have the result with the constant $\lambda = (\log 2)/64$.

Finally we track how M grows with m in order to know when u is large enough for this lemma to apply. Observe that we can write $M = E \max_k M_k$, where

$$M_k = \sup_{\substack{s,t \in T_k, s \neq t, \\ |t-s| \leq e^{-2}}} \frac{|B_t - B_s|}{\int_0^{d(s,t)} \sqrt{\log N(T,d,x)}\,\mathrm{d}x}, \qquad T_k = [ke^{-2}, (k+2)e^{-2}].$$

As $N(T,d,x) \geq N(T_k,d,x)$, defining

$$M'_k = \sup_{s,t \in T_k, s \neq t, |t-s| \leq e^{-2}} \frac{|B_t - B_s|}{\int_0^{d(s,t)} \sqrt{\log N(T_k,d,x)}\,\mathrm{d}x},$$

we see $M_k \leq M'_k$. As with the whole process C we can apply [38] Theorem 2.1.20 to each M'_k to see that $\Pr(|M'_k - EM'_k| > v) \leq 2e^{-v^2/2\sigma^2}$, with $\sigma^2 = 4/\log 2$ as before. That is, each $(M'_k - EM'_k)$ is subgaussian with parameter $12/\sqrt{\log 2}$ (see [38] Lemma 2.3.1). They all have the same constant (i.e. not depending on m) expectation, we can bound their maximum, by standard results for subgaussian variables (e.g. see [38] Lemma 2.3.4):

$$EM = E\Big[EM'_0 + \max_k \{M'_k - EM'_0\}\Big] \leq EM'_0 + 12\sqrt{2\log N/\log 2},$$

where N is the number of M'_k over which we take the maximum and scales linearly with m. It follows that M is of order bounded by $\sqrt{\log(m)}$ as $m \to \infty$. $\square$

Chapter 3

On statistical Calderón problems

Notation

Most of the notation to be used in this chapter is informally gathered here.

$D \subseteq \mathbb{R}^d$, $d \geq 3$, a domain, which is taken to mean a bounded connected open set with smooth boundary ∂D.

$\Subset$ 'compactly contained', i.e. $U \Subset V$ if the closure $\bar{U}$ is a subset of the interior $\operatorname{int} V$.

$\Gamma_{m,D'} = \{\gamma \in C_{\mathbb{R}}(D) : \inf_{x \in D} \gamma(x) \geq m,\ \gamma = 1 \text{ on } D \setminus D'\}$, some $m > 0$ and some domain $D' \Subset D$. $C_{\mathbb{R}}(D)$ denotes the continuous functions from D to $\mathbb{R}$.

$C_u(D)$ the Banach space of uniformly continuous functions from D to $\mathbb{C}$.

$\Gamma^{\alpha}_{m,D'}(M) = \{\gamma \in \Gamma_{m,D'} :\ \|\gamma\|_{H^{\alpha}(D)} \leq M\}$.

$\gamma \in \Gamma_{m,D'}$ a conductivity function, γ_0 its "true" value for some statistical theorems.

m_0, D_0 a lower bound and support set for the "true" γ_0; m_1, D_1 a lower bound and support set for any draw γ from the prior Π of Section 3.2.1.

$\|\cdot\|_{\infty}$ the usual supremum norm on (the bounded subsets of) $C(D)$ or $C(\mathbb{R})$.

$u_{\gamma,f}$ the (weak) solution to the Dirichlet problem (3.1) ($\nabla \cdot (\gamma \nabla u) = 0$ in D, $u = f$ on ∂D).

$B_{\gamma}(u, v) = \int_D \gamma \nabla u \cdot \nabla v^*$ the sesquilinear operator associated to $\nabla \cdot (\gamma \nabla(\cdot))$, where v^* denotes the complex conjugate of v.

H^s an L^2–Sobolev space of complex-valued functions (carefully defined in Section 1.6.1, Appendix 3.A); $H^1_0(D)$ the traceless subset of $H^1(D)$. $H^s_{\mathbb{R}}$ the subspace of H^s consisting of real-valued functions.

$\mathcal{H}_s = \left(H^{\min\{1,s+3/2\}}(D) \cap H^1_{\text{loc}}(D)\right)/\mathbb{C}$ for $H^s_{\text{loc}}(D)$ as in Section 1.6.1.

$H^s_\diamond(\partial D) = \{g \in H^s(\partial D) : \langle g, 1\rangle_{L^2(\partial D)} = 0\}$, $L^2_\diamond(\partial D) = H^0_\diamond(\partial D)$.

$(\phi_k^{(r)})_{k\in\mathbb{N}}$ an orthonormal basis of $H^r(\partial D)$ consisting of eigenfunctions of the Laplace–Beltrami operator on ∂D, with corresponding eigenvalues $\lambda_k \geq 0$ (details in Appendix 3.A).

$\frac{\partial}{\partial\nu}$ the outward normal derivative at the boundary of a domain (i.e. usually on ∂D).

$\Lambda_\gamma : H^{s+1}(\partial D)/\mathbb{C} \to H^s_\diamond(\partial D)$ the Dirichlet-to-Neumann map, taking f to $\gamma\frac{\partial u_{\gamma,f}}{\partial\nu}|_{\partial D}$.

$\tilde{\Lambda}_\gamma = \Lambda_\gamma - \Lambda_1$.

$\|\cdot\|_{A\to B}$ the operator norm between Banach spaces A and B.

$\|\cdot\|_* = \|\cdot\|_{H^{1/2}(\partial D)/\mathbb{C}\to H^{-1/2}(\partial D)}$.

$\mathcal{L}(A,B) = \{T : A \to B \text{ linear s.t. } \|T\|_{A\to B} < \infty\}$.

$\mathcal{L}_2(A,B) = \{T \in \mathcal{L}(A,B) : \|T\|^2_{\mathcal{L}_2(A,B)} := \sum_k \|Te_k^{(A)}\|^2_B < \infty\}$ the space of Hilbert–Schmidt operators from A to B for separable Hilbert spaces A and B, with $(e_k^{(A)})$ an orthonormal basis of A.

$\mathbb{H}_r = \mathcal{L}_2(H^r(\partial D), L^2_\diamond(\partial D))$ for $r \in \mathbb{R}$.

$(b_{kl}^{(r)})_{k,l\in\mathbb{N}}$ an orthonormal basis of $\mathbb{H}_r$, with $b_{kl}^{(r)}(f) \equiv (\phi_k^{(r)}) \otimes \phi_k^{(0)}(f) = \langle f, \phi_k^{(r)}\rangle_{H^r(\partial D)}\phi_l^{(0)}$

$Y = \tilde{\Lambda}_\gamma + \varepsilon\mathbb{W}$ the observed data, where $\mathbb{W}$ is a Gaussian white noise indexed by the Hilbert space $\mathbb{H}_r$, and ε is a noise level which tends to zero for our asymptotic results.

$P_\varepsilon^\gamma = P_{\varepsilon,r}^\gamma$ the law of Y, E_ε^γ the corresponding expectation operator, Var_γ the corresponding variance operator.

$p_\varepsilon^\gamma(Y) = \exp\Big(\frac{1}{\varepsilon^2}\langle Y, \tilde{\Lambda}_\gamma\rangle_{\mathbb{H}_r} - \frac{1}{2\varepsilon^2}\|\tilde{\Lambda}_\gamma\|^2_{\mathbb{H}_r}\Big)$ the probability density of the law of Y with respect to the law of $\varepsilon\mathbb{W}$.

$\ell(\gamma) = \log p_\varepsilon^\gamma$ the log-likelihood function.

$\xi_{\varepsilon,\delta} = (\log(\varepsilon^{-1}))^{-\delta}$ for $\varepsilon, \delta > 0$.

Π a prior on Γ_{m_1,D_1}, built from a Gaussian process prior Π' as described in Section 3.2.1. $\Pi(\,\cdot\mid Y)$ the corresponding posterior. $E^{\Pi_\varepsilon}[\,\cdot\mid Y]$ the posterior expectation. $(\mathcal{H}, \|\cdot\|_{\mathcal{H}}) \subseteq (H^\alpha(D), \|\cdot\|_{H^\alpha(D)})$ the RKHS of Π'.

Φ a 'regular link function' used in the prior construction (see Section 3.2.1).

π_{JK} the $\mathbb{H}_r$–orthogonal projection map onto $\mathrm{span}\{b_{jk}^{(r)} : j \leq J, k \leq K\}$ (see (3.27)).

$K(P,Q) = K(p,q) = E_{X\sim p}\log((p/q)(X))$ for distributions P, Q with densities p, q (the Kullback–Leibler divergence)

$B_{KL}^\varepsilon(\eta) = \{\gamma \in \Gamma_{m_1,D_1} : K(p_\varepsilon^{\gamma_0}, p_\varepsilon^\gamma) \leq (\eta/\varepsilon)^2, \mathrm{Var}_{\gamma_0}(\log(p_\varepsilon^{\gamma_0}/p_\varepsilon^\gamma)^2) \leq (\eta/\varepsilon)^2\}$.

$N(S, \rho, \delta)$ the covering numbers of the set S for metric ρ, i.e. the smallest number of ρ–balls of radius δ needed to cover S.

3.0 Introduction

Let's recapitulate the Calderón problem, as described in Section 1.2, and flesh out some of the details. Consider a domain $D \subset \mathbb{R}^d, d \geq 3$, which recall we understand here to mean a bounded connected open set with smooth boundary ∂D. For $\gamma : D \to (0, \infty)$ a *conductivity coefficient*, consider solutions u to the *Dirichlet problem*

$$\begin{aligned} \nabla \cdot (\gamma \nabla u) &= 0 \quad \text{in } D, \\ u &= f \quad \text{on } \partial D, \end{aligned} \tag{3.1}$$

where ∇ denotes the usual gradient operator and where $f : \partial D \to \mathbb{C}$ prescribes some boundary values. The parameter spaces considered in the sequel are of the form

$$\Gamma_{m,D'} = \left\{ \gamma \in C_{\mathbb{R}}(D) : \inf_{x \in D} \gamma(x) \geq m, \ \gamma = 1 \text{ on } D \setminus D' \right\}, \tag{3.2}$$

$$\Gamma^{\alpha}_{m,D'}(M) = \{ \gamma \in \Gamma_{m,D'} : \|\gamma\|_{H^{\alpha}(D)} \leq M \}, \quad M > 0, \tag{3.3}$$

where $m > 0$ is a fixed constant, D' is a domain compactly supported in D, and $\alpha \geq 0$ measures the regularity of γ in the Sobolev scale, with Sobolev spaces on D as defined in Section 1.6. $C_{\mathbb{R}}(D)$ denotes the space of continuous functions from D to $\mathbb{R}$, whose subspace of bounded functions we equip with the sup-norm $\|\cdot\|_{\infty}$. Unless otherwise stated, all integrals are taken with respect to Lebesgue measure on D and surface measure on ∂D.

The elliptic partial differential equation (PDE) in (3.1) has, for any $\gamma \in \Gamma_{m,D'}$ and any $f \in H^{s+1}(\partial D)/\mathbb{C}$, $s \in \mathbb{R}$, a unique weak solution $u_{\gamma,f}$ in the space

$$\mathcal{H}_s := \left(H^{\min\{1, s+3/2\}}(D) \cap H^1_{\text{loc}}(D) \right) / \mathbb{C}.$$

That is (as shown in Appendix 3.C, Lemma 3.17), for $u \in \mathcal{H}_s$, the equations

$$\begin{aligned} B_{\gamma}(u, v) := \int_D \gamma \nabla u \cdot \nabla v^* &= 0 \quad \forall v \in H^1_0(D), \\ u &= f \quad \text{on } \partial D, \end{aligned} \tag{3.4}$$

hold simultaneously if and only if $u = u_{\gamma,f}$, where the boundary values of u are defined in the trace sense described in Section 1.6.2 (v^* denotes the complex conjugate of v). Here

and below $/\mathbb{C}$ means that we identify functions $f, f+c$ which are equal up to a scalar $c \in \mathbb{C}$. Throughout this chapter, the Sobolev spaces H^s will consist of complex-valued functions, and we will explicitly write $H^s_{\mathbb{R}}$ to denote the corresponding Sobolev space consisting of real-valued functions (this is a subspace of $H^s(D)$, so we continue to write $\|\cdot\|_{H^s(D)}$ for its norm).

Given a solution $u_{\gamma,f}$ to the Dirichlet problem, one can measure the *Neumann (boundary) data*

$$\gamma \frac{\partial u_{\gamma,f}}{\partial \nu}\bigg|_{\partial D} \equiv \frac{\partial u_{\gamma,f}}{\partial \nu}\bigg|_{\partial D}, \quad \gamma \in \Gamma_{m,D'},$$

where $\frac{\partial}{\partial \nu}$ denotes the outward normal derivative on ∂D. It can be shown (see Lemma 3.18) that for any $s \in \mathbb{R}$ and any $f \in H^{s+1}(\partial D)/\mathbb{C}$, with $H^s(\partial D)$ as defined in Appendix 3.A, the Neumann data lies in the space

$$H^s_\diamond(\partial D) := \{g \in H^s(\partial D) : \langle g, 1\rangle_{L^2(\partial D)} = 0\}. \tag{3.5}$$

Thus, we may define the so-called *Dirichlet-to-Neumann map*,

$$\begin{aligned} \Lambda_\gamma : \quad H^{s+1}(\partial D)/\mathbb{C} &\to H^s_\diamond(\partial D), \\ f &\mapsto \gamma \frac{\partial u_{\gamma,f}}{\partial \nu}\bigg|_{\partial D}, \end{aligned} \tag{3.6}$$

which associates to each prescribed boundary value f the Neumann data of the solution of the PDE (3.1). Our choice to quotient the domain of Λ_γ by $\mathbb{C}$ is natural as the Neumann data is invariant with respect to addition of scalars.

The Calderón problem [13] addresses the task of recovering γ from knowledge of Λ_γ. Note that while Λ_γ itself is a linear operator between Hilbert spaces, the 'forward map' $\gamma \to \Lambda_\gamma$ is nonlinear. A landmark injectivity result by Sylvester and Uhlmann shows, however, that recovery is in principle possible.

Theorem (Sylvester & Uhlmann, [78])**.** *If* $\Lambda_{\gamma_1} = \Lambda_{\gamma_2}$ *then* $\gamma_1 = \gamma_2$.

Later Nachman [60] devised an elaborate inversion algorithm that allows recovery of γ if exact knowledge of the entire operator Λ_γ is available. Moreover Alessandrini [4] gave a stability estimate; that is, a quantitative continuity estimate for the inverse map.

The Calderón problem has since been vigorously studied and an excellent survey can be found in Uhlmann [81]. Its importance partly stems from its applications to electric impedance tomography (EIT) – as described in Section 1.2 and again in the section to follow – where discrete boundary measurements of the operator Λ_γ are performed to infer the interior conductivity γ. Any such data comes with error, and the arguably most

natural mathematical description of such approximate measurements is by a *statistical noise model*. As the superposition of many independent errors is well described by a normal distribution (via the central limit theorem), it is further natural to postulate that this noise follows a Gaussian law. In algorithmic practice this has already been widely acknowledged in the general setting of inverse problems, where statistical, and in particular Bayesian, inversion approaches have flourished in the last decade since the influential work of Stuart [77]. In the context of EIT we refer to the articles [46, 45, 49, 74, 30, 25] and the many references therein. Surprisingly little theory is available that gives statistical guarantees for the performance of such Bayesian denoising methodology, particularly for nonlinear problems. Recent progress has been made in some nonlinear settings (see [62, 61, 63, 64, 59]) but in the context of the Calderón problem described above, the only paper known to the authors is Caro & Garcia [14], which addresses estimating the boundary values $\gamma|_{\partial D}$ and so is in some sense orthogonal (and complementary) to the question of interior recovery, with given boundary values, considered here.

We will introduce a natural noise model (3.13) in the next section where one observes Λ_γ corrupted by a Gaussian white noise in an appropriate space of Hilbert–Schmidt operators. The noise is described by the scalar quantity $\varepsilon > 0$ governing its magnitude and a parameter $r \in \mathbb{R}$ determining its "spectral heteroscedasticity". If we denote by $P_\varepsilon^\gamma = P_{\varepsilon,r}^\gamma$ the resulting probability law of the noisy observations Y of Λ_γ, then our main results can be summarised in the following two theorems.

Theorem 3.1. *Let $\alpha > 3 + d$ be an integer, let $m_0 > 0, M > 0$ be given, and let D_0 be a domain in $\mathbb{R}^d$ such that $\bar{D}_0$ is contained in D.*

There exists a function $\hat{\gamma} = \hat{\gamma}_\varepsilon(Y)$ of the observations $Y \sim P_\varepsilon^\gamma$ such that

$$\sup_{\gamma \in \Gamma^\alpha_{m_0,D_0}(M)} P_\varepsilon^\gamma(\|\hat{\gamma} - \gamma\|_\infty > C \log(1/\varepsilon)^{-\delta}) \to 0, \quad \textit{as} \quad \varepsilon \to 0,$$

where $\delta > 0$ depends only on d, and C depends only on α, M, m_0, D, D_0 and r.

The estimator $\hat{\gamma}$ in the previous theorem has a natural Bayesian interpretation as the posterior mean of a suitable Gaussian process based prior for γ. The derivation and implementation of $\hat{\gamma}$ are described in Section 3.2, where the more precise Theorem 3.3, which implies Theorem 3.1, is given. Note that $\hat{\gamma}$ can be calculated without knowledge of the bound M for $\|\gamma_0\|_{H^\alpha(D)}$.

The slow (logarithmic) convergence rate is not surprising in view of the folklore that the Calderón problem is a severely ill-posed inverse problem. The following result

makes this folklore information-theoretically precise – it shows that the convergence rate obtained by the estimator $\hat{\gamma}$ is optimal in the minimax sense, at least up to the precise value of the exponent δ, for the prototypical case where D_0, D are nested balls in $\mathbb{R}^d$.

Theorem 3.2. *Let $D_0 = \{x \in \mathbb{R}^d : \|x\| < 1/2\} \subset D = \{x \in \mathbb{R}^d : \|x\| < 1\}$ in $\mathbb{R}^d$, let α be an integer greater than 2, and let $m_0 \leq 1$ be arbitrary.*

For any $\delta' > \alpha(2d-1)/d$ and all M large enough there exists $c = c(\delta', \alpha, d, m_0, r, M)$ such that

$$\inf_{\tilde{\gamma}} \sup_{\gamma \in \Gamma^{\alpha}_{m_0, D_0}(M)} P^{\gamma}_{\varepsilon}(\|\tilde{\gamma} - \gamma\|_{\infty} > c \log(1/\varepsilon)^{-\delta'}) > 1/7$$

for all ε small enough, where the infimum extends over all measurable functions $\tilde{\gamma} = \tilde{\gamma}(Y)$ of the data $Y \sim P^{\gamma}_{\varepsilon}$.

The particular value of 1/7 in the lower bound is chosen for convenience (cf. Theorem 1.2). We do not pursue the problem of finding the exact exponent δ in the minimax convergence rate: determining the optimal value of δ in the stability estimate underlying our proof is a delicate PDE question in its own right and beyond the scope of this thesis.

This chapter is structured as follows. In Section 3.1 we introduce the measurement model we consider in our theorems, and discuss its precise relationship (in a Le Cam sense) to physical measurement models arising in medical imaging practice. In Section 3.2 we give the construction of the Bayesian algorithm $\hat{\gamma}$ that solves our noisy version of the Calderón problem. All proofs and related background material are relegated to later sections.

3.1 Noise model

We now introduce a rigorous framework for observing a noisy version of the operator Λ_γ from (3.6). Let $\tilde{\Lambda}_\gamma = \Lambda_\gamma - \Lambda_1$ where the fixed (deterministic and known) operator Λ_1 is the Dirichlet-to-Neumann map for the standard Laplace equation, that is, eq. (3.1) with $\gamma = 1$ identically on D. We then equivalently consider measuring a noisy version of $\tilde{\Lambda}_\gamma$.

As described in the Chapter 1, real-world data involving the Calderón problem arises for example in the medical imaging technique of *electrical impedance tomography*, wherein electrodes are attached to a patient (or some other physical medium), and are used both to apply voltages and to record the resulting currents. If we assume the applied voltages are uniform across the surface of any given electrode, and the electrodes measure the average current across their surface, we are led to the observation model

$$Y_{p,q} = \langle \tilde{\Lambda}_\gamma[\psi_p], \psi_q \rangle_{L^2(\partial D)} + \varepsilon g_{p,q}, \quad p, q \leq P, \quad g_{p,q} \overset{iid}{\sim} N(0,1), \ \varepsilon > 0, \tag{3.7}$$

where the ψ_p are, up to scaling factors, indicator functions of some disjoint measurable subsets $(I_p)_{p\leq P}$ of ∂D representing the locations of the electrodes. In principle we might expect the noise level $\varepsilon > 0$ to vary with p and q, but we can accommodate a single noise level by instead varying the scaling factors on ψ_p and ψ_q; in particular, if the ψ_p are $L^2(\partial D)$–orthonormal then the homoscedastic noise model given above is natural.

An alternative noise model, more tractable in the theory that follows, considers Fourier-type measurements. Consider a basis $(\phi_k)_{k\in\mathbb{N}\cup\{0\}} = (\phi_k^{(0)})_{k\in\mathbb{N}\cup\{0\}}$ of $L^2(\partial D)$ comprising eigenfunctions of the Laplace–Beltrami operator on the compact manifold ∂D. By discarding the constant function ϕ_0 we obtain a basis of the spaces $L^2(\partial D)/\mathbb{C}$ and $L^2_\diamond(\partial D) = H^0_\diamond(\partial D)$. Moreover, appropriate rescaling of these basis functions also provides orthonormal bases $(\phi_k^{(r)})_{k\in\mathbb{N}}$ of all $H^r(\partial D)/\mathbb{C}$ and $H^r_\diamond(\partial D)$ spaces, $r \in \mathbb{R}$ – see Appendix 3.A for details. For some $r \in \mathbb{R}$, we then consider the noisy matrix measurement model

$$Y_{j,k} = \langle \tilde{\Lambda}_\gamma[\phi_j^{(r)}], \phi_k^{(0)}\rangle_{L^2(\partial D)} + \varepsilon g_{j,k}, \quad j \leq J, k \leq K, \quad g_{j,k} \overset{iid}{\sim} N(0,1),\ \varepsilon > 0. \tag{3.8}$$

We will work below with (a natural continuous analogue of) this more tractable model, but this does not force us to relinquish the intepretability of our results in the model (3.7), at least when sufficiently many measurements are available ($P \to \infty$): one can approximate Laplace–Beltrami eigenfunctions via linear combinations of indicator functions, and in doing so we approximately recover data from model (3.8) given data from model (3.7). Thus any estimator for γ built in model (3.8) can be approximately constructed from data in model (3.7). The following one-way statistical discrepancy
asymptotic performance of the algorithm. We restate the result precisely in Appendix 3.D (Theorem 3.21), using the notion of Le Cam discrepancy between statistical experiments introduced in Section 1.3.1.

Theorem. *Suppose the parameter $\gamma \in \Gamma_{m,D'}$ has supremum norm bounded by a fixed constant M and suppose the indicator functions $(\psi_p)_{p\leq P}$ are well-spaced within ∂D. Then given data from (3.7), we can construct data from (3.8) with $r = 0$, with, for P large enough depending on J, K and ε, asymptotically vanishing information loss, in the sense of one-way Le Cam discrepancy.*

In particular, given any bounded loss function and any decision rule ρ_2 in model (3.8), we can construct a corresponding rule ρ_1 for model (3.7) whose excess risk relative to ρ_2 tends to zero.

We next argue that model (3.8) is close to a continuous model in which one observes noisy operator-valued data. We first need some definitions. For $(A, \|\cdot\|_A)$ and $(B, \|\cdot\|_B)$ separable Hilbert spaces we equip the space

$$\mathcal{L}(A, B) = \{T : A \to B \text{ linear s.t. } \|T\|_{A\to B} < \infty\}$$

of bounded linear maps from A to B with the usual operator norm

$$\|T\|_{A\to B} = \sup\{\|Tx\|_B : x \in A, \|x\|_A \leq 1\}. \tag{3.9}$$

Define also the space $\mathcal{L}_2(A, B)$ of Hilbert–Schmidt operators $A \to B$,

$$\mathcal{L}_2(A, B) = \Big\{T \in \mathcal{L}(A, B) \text{ s.t. } \|T\|^2_{\mathcal{L}_2(A,B)} := \sum_k \|Te_k^{(A)}\|_B^2 < \infty\Big\}, \tag{3.10}$$

where $(e_k^{(A)})$ is any orthonormal basis of A. This is a Hilbert space with inner product

$$\langle S, T\rangle_{\mathcal{L}_2(A,B)} = \sum_k \langle Se_k^{(A)}, Te_k^{(A)}\rangle_B. \tag{3.11}$$

The preceding definitions are independent of the choice of basis (e_k^A). See Chapter 12 in Aubin [5] for an introduction to spaces of Hilbert–Schmidt operators.

Now define, for $r \in \mathbb{R}$,

$$\mathbb{H}_r := \mathcal{L}_2(H^r(\partial D)/\mathbb{C}, L^2_\diamond(\partial D)), \tag{3.12}$$

and consider observing data Y from probability law $P^\gamma_{\varepsilon,r}$ arising from the equation

$$Y = \tilde{\Lambda}_\gamma + \varepsilon\mathbb{W}, \quad \varepsilon > 0, \tag{3.13}$$

where $\mathbb{W}$ is a Gaussian white noise (*isonormal process*; see, e.g. p19 in [38]) indexed by the Hilbert space $\mathbb{H}_r$. We often suppress the parameter r, and write P^γ_ε for the probability law and E^γ_ε for the corresponding expectation operator.

Using the natural Hilbert space isomorphism between $\mathbb{H}_r$ and the sequence space ℓ^2 given by considering coordinates with respect to the orthonormal basis induced by the $\phi_k^{(r)}$, the model can be interpreted concretely by the action of Y on any $T \in \mathbb{H}_r$: if

$$\langle \mathbb{W}, T\rangle_{\mathbb{H}_r} := \sum_{j,k} g_{jk}\langle T\phi_j^{(r)}, \phi_k^{(0)}\rangle_{L^2(\partial D)}, \text{ for } g_{jk} \overset{iid}{\sim} N(0, 1),$$

then we are given a measurement of the Gaussian process

$$\left(Y(T) = \langle \tilde{\Lambda}_\gamma, T\rangle_{\mathbb{H}_r} + \varepsilon \sum_{j,k} g_{jk} \langle T\phi_j^{(r)}, \phi_k^{(0)}\rangle_{L^2(\partial D)}\right)_{T\in\mathbb{H}_r}. \tag{3.14}$$

What precedes makes sense rigorously only if $\tilde{\Lambda}_\gamma \in \mathbb{H}_r$, and it is proved in Appendix 3.C, Lemma 3.19, that this is indeed the case for any $\gamma \in \Gamma_{m,D'}$ and any $r \in \mathbb{R}$.

The choice of the domain in the definition of $\mathbb{H}_r$ corresponds to the experimental design, so in principle one can choose r freely, and our results are written to accommodate any value. Changing r adjusts how the signal-to-noise ratio varies with frequency: as r increases, the signal at high frequencies (i.e. at larger values of k) decreases compared to the signal at low frequencies. Likely the most realistic choices are $r = 0$, so that the previous theorem relating models (3.7) and (3.8) applies, and $r = 1$, because this ensures that the signal-to-noise ratio is the same across all frequencies: since Λ_γ maps $H^1(\partial D)/\mathbb{C}$ to $L^2_\diamond(\partial D)$ isomorphically (Lemma 3.18), the signal magnitude $\|\Lambda_\gamma \phi_k^{(1)}\|_{L^2(\partial D)}$ is of order 1 for all k. A similar reasoning ($\|\phi_k^{(0)}\|_{L^2(\partial D)} = 1$ for all k) underpins the choice of codomain $L^2_\diamond(\partial D)$.

We will prove our main results Theorems 3.1 and 3.2 in the model (3.13). The following result, given rigorously in Appendix 3.D (Theorem 3.23), justifies focussing our attention on the continuous model (3.13).

Theorem. *Let $r \in \mathbb{R}$. Suppose the parameter $\gamma \in \Gamma_{m,D'}$ has supremum norm bounded by a fixed constant M. Then given data from (3.8), we can construct data from (3.13), and vice versa, with asymptotically vanishing information loss as $\min(J, K) \to \infty$, in the sense of the Le Cam distance.*

We note that in principle, all our results could be directly derived in the model (3.8), but the continuous model is more convenient for the application of PDE techniques and facilitates a clearer exposition in the proofs to follow.

3.2 The Bayesian approach to the noisy Calderón problem

We now construct the estimator $\hat{\gamma}$ featuring in Theorem 3.1. Following the Bayesian approach to inverse problems described in Chapter 1 and proposed already in the context of the EIT inverse problem in [46], we will construct $\hat{\gamma}$ as the posterior mean arising from a certain *Gaussian process* prior for γ. To this end we need to first establish the

existence of a posterior distribution in our measurement setting. In the Gaussian white noise model (3.13), the log-likelihood function can be derived from the Cameron–Martin theorem in a suitable Hilbert space: precisely, the law P_ε^γ of Y is dominated by the law P_ε^1 of $\varepsilon\mathbb{W}$, with log-likelihood function

$$\ell(\gamma) \equiv \log p_\varepsilon^\gamma(Y) := \log \frac{dP_\varepsilon^\gamma}{dP_\varepsilon^1}(Y) = \frac{1}{\varepsilon^2}\langle Y, \tilde{\Lambda}_\gamma\rangle_{\mathbb{H}_r} - \frac{1}{2\varepsilon^2}\|\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r}^2, \tag{3.15}$$

whenever $\tilde{\Lambda}_\gamma \in \mathbb{H}_r$. See Section 7.4 in [61] for a detailed derivation, which requires Borel-measurability (ensured by Lemma 3.6 below) of the map $\gamma \mapsto \tilde{\Lambda}_\gamma$ from the (Polish) space $\Gamma_{m,D'}$ equipped with the $\|\cdot\|_\infty$–topology into the Hilbert space $\mathbb{H}_r$.

Then for any prior (Borel) probability measure Π on $\Gamma_{m,D'}$, the posterior distribution given observations Y is given by

$$\Pi(B \mid Y) = \frac{\int_B p_\varepsilon^\gamma(Y)\,\mathrm{d}\Pi(\gamma)}{\int_{\Gamma_{m,D'}} p_\varepsilon^\gamma(Y)\,\mathrm{d}\Pi(\gamma)}, \quad B \subset \Gamma_{m,D'} \text{ Borel-measurable}, \tag{3.16}$$

see again Section 7.4 in [61] (and also [34], eq (1.1)). We denote by $E^\Pi[\cdot]$ the expectation operator according to the prior, and by $E^\Pi[\,\cdot \mid Y]$ the expectation according to the posterior.

3.2.1 Prior construction

We will construct a Gaussian process prior for the conductivity γ by first drawing a Gaussian random field ρ in D, and then enforcing positivity by a suitable composition map Φ to give $\gamma = \Phi \circ \rho$. In the proofs we will require that the true γ_0 is in the "interior" of the support of the prior, so recalling that Theorem 3.1 is stated uniformly over $\Gamma_{m_0,D_0}^\alpha(M)$, we choose $m_1 < m_0$ and a domain D_1 such that $D_0 \Subset D_1 \Subset D$ (recall $U \Subset V$ means $\bar{U} \subset \operatorname{int} V$) and construct a prior concentrating its mass on Γ_{m_1,D_1}.

For the base prior for ρ we employ the following condition – we refer, e.g., to Sections 2.1 and 2.6 in [38] for the basic definitions of Gaussian measures and processes and their reproducing kernel Hilbert spaces (RKHS).

Assumption A. Let Π' be a centred Gaussian Borel probability measure on the Banach space $C_u(D)$ of uniformly continuous functions on D, and let $\alpha > \beta > 2 + d/2$ be integers. Assume $\Pi'(H_{\mathbb{R}}^\beta(D)) = 1$ and that the RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ of Π' is continuously embedded into the Sobolev space $H_{\mathbb{R}}^\alpha(D)$.

Natural candidates for such priors are restrictions to D of Gaussian processes whose covariances are given by Matérn kernels, see [34], p313 and p575 – in these cases one can

satisfy the assumption for any $2 + d/2 < \beta < \alpha - d/2$ by taking $\mathcal{H}$ to *coincide* with the Sobolev space $H^\alpha_{\mathbb{R}}(D)$. The restriction to integer-valued α, β is convenient to simplify some proofs, but in principle should not be necessary.

Consider a random function $\rho' \sim \Pi'$ and let χ be a smooth cutoff function, identically 1 on D_0 and compactly supported in D_1. Define a new random function

$$\rho(x) = \rho_\varepsilon(x) = \varepsilon^{d/(\alpha+d)}\chi(x)\rho'(x),\ x \in D,\ \rho' \sim \Pi', \tag{3.17}$$

and denote its (Borel) law in $C_u(D)$ by $\Pi_\rho = \Pi_{\rho,\varepsilon}$.

Let $\Phi : \mathbb{R} \to [m_1, \infty)$ be a regular link function in the sense of [64]; that is to say, let Φ be a smooth bijective function satisfying $\Phi(0) = 1$, $\Phi' > 0$ on $\mathbb{R}$, and $\|\Phi^{(j)}\|_\infty < \infty$ for all integers $j \geq 1$. Such a function exists via arguments based on mollifiers similar to those used in the footnote in Section 1.6.1: we refer to [64] Example 3.2 where a regular link function is exhibited, and to [64] Lemma 6.1 for basic properties of such functions. In particular we note that there are constants $C = C(\Phi)$, $C' = C'(\Phi, \alpha)$ and $C'' = C''(\Phi, \alpha, m_0, m_1)$ such that

$$\|\Phi \circ \rho - \Phi \circ \rho_0\|_\infty \leq C\|\rho - \rho_0\|_\infty, \tag{3.18}$$

$$\|\Phi \circ \rho\|_{H^\alpha(D)} \leq C'(1 + \|\rho\|^\alpha_{H^\alpha(D)}), \quad \alpha \in \mathbb{N}, \alpha \geq d/2, \tag{3.19}$$

$$\|\Phi^{-1} \circ \gamma_0\|_{H^\alpha(D)} \leq C''(1 + \|\gamma_0\|^\alpha_{H^\alpha(D)}), \quad \alpha \in \mathbb{N}, \alpha \geq d/2, \tag{3.20}$$

for any bounded functions ρ, ρ_0 and any $\gamma_0 \in \Gamma_{m_0,D_0}$. The first inequality is an immediate consequence of the mean value theorem, the second is given in [64] Lemma 6.1, and the third follows from the arguments of the same lemma, applied to the function Φ^{-1} (this can be seen to be regular on the domain $[m_0, \infty)$ for $m_0 > m_1$ by considering explicit formulas for its derivatives).

The final prior for the conductivity γ is now obtained as the law Π of the random field

$$\gamma(x) = \Phi \circ \rho(x),\ \rho \sim \Pi_\rho,\ x \in D. \tag{3.21}$$

3.2.2 Posterior contraction result

For the following result we define

$$\xi_{\varepsilon,\delta} = \log(1/\varepsilon)^{-\delta}, \quad \varepsilon, \delta > 0. \tag{3.22}$$

Theorem 3.3. *Let Π' be a base prior satisfying Assumption A, let Π be the prior from (3.21), and denote by $\Pi(\,\cdot\mid Y)$ the posterior distribution arising from observations Y in the model (3.13). Suppose that for some $M>0$ the true conductivity γ_0 belongs to the set*

$$\Gamma_{m_0,D_0}\cap\{\Phi\circ\rho:\rho\in\mathcal{H},\|\rho\|_{\mathcal{H}}\le M\}. \tag{3.23}$$

Then there exist constants $\delta=\delta(d)>0$ and $K=K(M,m_0,m_1,D,D_0,D_1,\Phi,\chi,r,\alpha)>0$ such that as $\varepsilon\to 0$,

$$\Pi\Big(\|\gamma-\gamma_0\|_\infty>\tfrac{1}{2}K\xi_{\varepsilon,\delta}\mid Y\Big)\to^{P_\varepsilon^{\gamma_0}}0. \tag{3.24}$$

Moreover, if $E^\Pi[\gamma\mid Y]$ denotes the (Bochner) mean of $\Pi(\,\cdot\mid Y)$, then

$$\sup_{\gamma_0}P_\varepsilon^{\gamma_0}\Big(\|E^\Pi[\gamma\mid Y]-\gamma_0\|_\infty>K\xi_{\varepsilon,\delta}\Big)\to 0\ \text{as}\ \varepsilon\to 0, \tag{3.25}$$

where the supremum extends over all γ_0 in the set (3.23).

Theorem 3.3, whose proof is given in Section 3.3.4, immediately implies Theorem 3.1. Indeed, given an integer $\alpha>3+d$, let Π be a prior from (3.21) whose base prior Π' satisfies Assumption A with RKHS $\mathcal{H}=H^\alpha_{\mathbb{R}}(D)$ (for instance, take the Matérn prior and note that a choice of integer $\beta>2+d/2$ is admissible) and let $\hat\gamma=E^\Pi[\gamma\mid Y]$ be the associated posterior mean. It suffices to show that the conditions of Theorem 3.1 imply those of Theorem 3.3, in particular that for any $\gamma_0\in\Gamma^\alpha_{m_0,D_0}(M)$, there exists an $M'=M'(\alpha,M,m_0,D,D_0)$ such that $\rho_0:=\Phi^{-1}\circ\gamma_0$ has $H^\alpha(D)$ norm bounded by M'. But this is immediate from (3.20).

Remark (Computation of the posterior mean). As noted in Section 1.5, optimisation based methods commonly used in inverse problems (such as the MAP estimates studied in [45, 64]) may not recover global optima in the EIT setting, since the nonlinearity of the map $\rho\mapsto\Lambda_{\Phi(\rho)}\equiv\Lambda_\gamma$ implies that the associated least squares criterion is nonconvex. In contrast, a key advantage of the posterior mean $E^\Pi[\gamma\mid Y]$ is that it can be calculated via MCMC or expectation-propagation methods (naturally in the discretisations (3.7) or (3.8) of our continuous model (3.13)).

For example, the pCN algorithm as described in Section 1.5 allows one to sample from posterior distributions in general inverse problems as long as the forward map $\rho\mapsto\Lambda_{\Phi(\rho)}$ can be evaluated, which in our setting has the basic cost of (numerically) solving the standard elliptic PDE (3.1). Even in the absence of log-concavity of the posterior measure one can give sampling guarantees for this algorithm (see [41]) so that the approximate computation of $E^\Pi[\gamma\mid Y]$ by the sample average $(1/M)\sum_m\gamma_m$ of the pCN Markov chain is provably possible at any given noise level ε. Related work on MCMC-based approaches

in the setting of electric impedance tomography can be found in [46, 74, 25], wherein also many further references can be found. Instead of MCMC methods one can also resort to variational Bayes methods – see for example [30], where computation of the posterior mean is addressed specifically for the EIT problem relevant in the present chapter.

Remark (Non-linearity and Gaussian priors). Dealing with the unboundedness of Gaussian priors for ρ and the nonlinearity of the composite forward map $\rho \mapsto \Lambda_{\Phi(\rho)}$ is a main challenge in proving Theorem 3.3. We show how to adapt the proof template devised in [59] for a very different inverse problem to the case of the Calderón problem – as in [59], this requires the scaling of the base prior Π' in (3.17), and also necessitates the above choice of a *regular* link function Φ, since otherwise the implied priors for the 'regression operators' Λ_γ potentially behave too erratically for our proof method via the stability estimate of [4] to work.

3.3 Proofs

3.3.1 Low rank approximation of $\tilde{\Lambda}_\gamma$

A key idea used in various proofs that follow is that we can project the operator $\tilde{\Lambda}_\gamma$ onto a finite-dimensional subspace and incur only a small error. To define the projection, we introduce the orthonormal basis $(b_{jk}^{(r)})_{j,k\in\mathbb{N}}$ of $\mathbb{H}_r$ consisting of tensor product operators

$$b_{jk}^{(r)}(f) = (\phi_j^{(r)}) \otimes \phi_k^{(0)}(f) := \langle f, \phi_j^{(r)} \rangle_{H^r(\partial D)} \phi_k^{(0)}, \quad f \in H^r(\partial D)/\mathbb{C}, \tag{3.26}$$

where the Laplace-Beltrami eigenfunctions $\phi_j^{(r)}$ were introduced before (3.8) and are described in more detail in Appendix 3.A. For an operator $U \in \mathbb{H}_r$ we remark that the coefficients $u_{jk} := \langle U, b_{jk}^{(r)} \rangle_{\mathbb{H}_r}$ are given by $u_{jk} = \langle U\phi_j^{(r)}, \phi_k^{(0)} \rangle_{L^2(\partial D)}$, and we define the projection map π_{JK} by

$$\pi_{JK} U = \sum_{j\le J} \sum_{k \le K} \langle U\phi_j^{(r)}, \phi_k^{(0)} \rangle_{L^2(\partial D)} b_{jk}^{(r)}. \tag{3.27}$$

Lemma 3.4. *For constants $m, M > 0$ and a domain $D' \Subset D$, let $\gamma \in \Gamma_{m,D'}$ be bounded by M on D. For any $\nu > 0$ there is a constant $C = C(\nu, D, D', r) > 0$ such that*

$$\|\tilde{\Lambda}_\gamma - \pi_{JK}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} \le C\tfrac{M}{m} \min(J, K)^{-\nu}.$$

Proof. Apply Lemma 3.16 from Appendix 3.B with $s = 0$, and, for some $\nu > 0$, $p = r - \nu(d-1)$, $q = \nu(d-1)$, and note $\|\tilde{\Lambda}_\gamma\|_{H^{p-(d-1)}(D)\to H^q(D)} \leq C\frac{M}{m}$ for such a constant C by Lemma 3.18. □

The proofs of the continuity results for the Calderón problem in the next section involve the $H^{1/2}(\partial D)/\mathbb{C} \to H^{-1/2}(\partial D)$ operator norm, which we denote by $\|\cdot\|_*$. To connect this norm to the information-theoretically relevant $\mathbb{H}_r$–norm, the following consequence of Lemma 3.4 will be useful.

Lemma 3.5. *For $m, M_0, M_1 > 0$ and a domain $D' \Subset D$, let $\gamma_0, \gamma_1 \in \Gamma_{m,D'}$ be bounded on D by M_0 and M_1 respectively. Then there are constants C_1 and C_2 depending only on r, D and D_0 such that if $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq 1$ then*

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq C_1\Big(\tfrac{M_1+M_0}{m}\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_*\Big)^{1/2}, \tag{3.28}$$

and if $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq 1$ then

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq C_2\Big(\tfrac{M_1+M_0}{m}\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r}\Big)^{1/2}. \tag{3.29}$$

Proof. For $J > 0$ and $\nu > 0$ to be chosen, by Lemma 3.4 we have

$$\|\tilde{\Lambda}_\gamma - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} \leq C\tfrac{M_1}{m}J^{-\nu},$$

for a constant $C = C(\nu, D, D', r)$, and a corresponding bound holds for $\|\tilde{\Lambda}_{\gamma_0} - \pi_{JJ}\tilde{\Lambda}_{\gamma_0}\|_{\mathbb{H}_r}$. An application of Lemma 3.15 with $s = 0$, $p = d - 1/2$, and $q = -1/2$, also yields (with x_+ denoting $\max(x, 0)$),

$$\begin{aligned}\|\pi_{JJ}\tilde{\Lambda}_\gamma - \pi_{JJ}\tilde{\Lambda}_{\gamma_0}\|_{\mathbb{H}_r} &\leq C'(1 + J^{1/(d-1)})^{1/2+(d-1/2-r)_+}\|\pi_{JJ}(\tilde{\Lambda}_\gamma - \tilde{\Lambda}_{\gamma_0})\|_{\mathcal{L}_2(H^{d-1/2},H^{-1/2})}\\ &\leq C'(1 + J^{1/(d-1)})^{(d+|r|)}\|\tilde{\Lambda}_\gamma - \tilde{\Lambda}_{\gamma_0}\|_{\mathcal{L}_2(H^{d-1/2},H^{-1/2})}\\ &\leq c'(1 + J^{1/(d-1)})^{(d+|r|)}\|\tilde{\Lambda}_\gamma - \tilde{\Lambda}_{\gamma_0}\|_*,\end{aligned}$$

for constants C', c' depending on D, r, where we use Lemma 3.16 to obtain the final inequality. Since $\Lambda_\gamma - \Lambda_{\gamma_0} = \tilde{\Lambda}_\gamma - \tilde{\Lambda}_{\gamma_0}$, we deduce, for a constant C'' that

$$\begin{aligned}\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} &\leq \|\tilde{\Lambda}_\gamma - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} + \|\tilde{\Lambda}_{\gamma_0} - \pi_{JJ}\tilde{\Lambda}_{\gamma_0}\|_{\mathbb{H}_r} + \|\pi_{JJ}\tilde{\Lambda}_\gamma - \pi_{JJ}\tilde{\Lambda}_{\gamma_0}\|_{\mathbb{H}_r}\\ &\leq C''\Big(\Big(\tfrac{M_1+M_0}{m}\Big)J^{-\nu} + J^{(d+|r|)/(d-1)}\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_*\Big).\end{aligned}$$

Since $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq 1$, we can choose an integer J to balance the two terms up to a constant (take $J = \lfloor(\frac{m}{M_0+M_1}\|\tilde{\Lambda}_\gamma - \tilde{\Lambda}_{\gamma_0}\|_*)^{-(d-1)/(\nu(d-1)+d+|r|)}\rfloor$). This yields, for a constant

C''',

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq C'''\Big(\tfrac{M_1+M_0}{m}\Big)\Big(\Big(\tfrac{M_1+M_0}{m}\Big)^{-1}\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_*\Big)^{\nu(d-1)/(\nu(d-1)+d+|r|)}.$$

Choosing $\nu = (d+|r|)/(d-1)$ yields (3.28).

For (3.29), given that $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq \|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathcal{L}_2(H^{1/2},H^{-1/2})}$, which follows from the fact that $\|\cdot\|_{A\to B} \leq \|\cdot\|_{\mathcal{L}_2(A,B)}$ for any separable Hilbert spaces A and B, and the observation that the proof of Lemma 3.4 equally applies with the $\mathcal{L}_2(H^{1/2}, H^{-1/2})$ norm in place of the $\mathbb{H}_r$ norm, an almost identical argument to the above yields

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq C\Big(\Big(\tfrac{M_1+M_0}{m}\Big)J^{-\nu} + J^{(r-1/2)_+/(d-1)}\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r}\Big).$$

Choosing J to balance the terms yields

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq C\Big(\tfrac{M_1+M_0}{m}\Big)\Big(\Big(\tfrac{M_1+M_0}{m}\Big)\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r}\Big)^{\nu(d-1)/(\nu(d-1)+(r-1/2)_+)},$$

and the result follows from noting that the exponent is at least $1/2$ for $\nu > (r-1/2)_+$. □

3.3.2 Continuity and stability results

We now prove the following continuity estimates for the maps $\gamma \mapsto \Lambda_\gamma, \Lambda_\gamma \to \gamma$.

Lemma 3.6. *For $m, M_0, M_1 > 0$ and a domain $D' \Subset D$, let $\gamma, \gamma_0 \in \Gamma_{m,D'}$ be bounded on D by M_1 and M_0 respectively. Then there exist constants $C = C(r, D, D')$, $\tau = \tau(D)$ such that*

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq C\tfrac{M_1M_0}{m^2}\|\gamma - \gamma_0\|_\infty^{1/2},$$

whenever $\|\gamma - \gamma_0\|_\infty \leq \tau\frac{m^2}{M_0M_1}$.

Lemma 3.7. *For some $\beta > 2+d/2$, some $m, M > 0$ and some domain $D' \Subset D$, suppose $\gamma, \gamma_0 \in \Gamma^\beta_{m,D'}(M)$. Then there exist constants C and τ depending only on M, D, D', m, β and r such that, for a constant $\delta = \delta(d,\beta) \in (0,1)$,*

$$\|\gamma - \gamma_0\|_\infty \leq C|\log\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r}|^{-\delta},$$

whenever $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq \tau$.

Note we calculate the explicit form of the dependence on the bounds M_1 and M_0 in Lemma 3.6 because this is required in the proofs of the main theorems (see in particular the proof of Lemma 3.11).

Proof of Lemma 3.6. We initially show, for some $C = C(D)$, that

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_* \leq C\tfrac{M_0(m+M_1)}{m^2}\|\gamma - \gamma_0\|_\infty. \tag{3.30}$$

The result then follows from Lemma 3.5, noting that necessarily $m < 1 < \min(M_0, M_1)$ so that $(M_0 + M_1)M_0(m + M_1)/m^3 \leq 4M_0^2M_1^2/m^4$.

For γ as given and $f \in H^{1/2}(\partial D)/\mathbb{C}$, recall we write $u_{\gamma,f}$ for the unique solution in $\mathcal{H}_1 \equiv H^1(D)/\mathbb{C}$ to the Dirichlet problem on D with conductivity γ and boundary data f, whose existence is guaranteed by Lemma 3.17. The equivalence class of functions $u_{\gamma,f} - u_{\gamma_0,f}$ has a representative $w \in H_0^1(D)$, which is easily seen to solve the PDE

$$\begin{aligned} \nabla\cdot(\gamma\,\nabla\,w) &= \nabla\cdot((\gamma_0-\gamma)\,\nabla\,u_{\gamma_0,f}) && \text{in } D,\\ w &= 0 && \text{on } \partial D. \end{aligned} \tag{3.31}$$

We have the dual representation

$$\Big\|\frac{\partial w}{\partial\nu}\Big\|_{H^{-1/2}(\partial D)} = \sup\Big\{\Big|\Big\langle\frac{\partial w}{\partial\nu},\sigma\Big\rangle_{L^2(\partial D)}\Big| : \sigma\in H^{1/2}(\partial D), \|\sigma\|_{H^{1/2}(\partial D)} = 1\Big\}.$$

For $\sigma \in H^{1/2}(\partial D)$, Theorem 1.8 tells us that there exists $\Sigma \in H^1(D)$ such that $\Sigma|_{\partial D} = \sigma$ and $\|\Sigma\|_{H^1(D)} \leq C\|\sigma\|_{H^{1/2}(\partial D)}$ for a constant $C = C(D)$. Repeatedly applying the divergence theorem (recalling that $\gamma = \gamma_0 = 1$ on ∂D) and the Cauchy–Schwarz inequality, we deduce

$$\begin{aligned} \Big|\int_{\partial D}\sigma^*\frac{\partial w}{\partial\nu}\Big| &= \Big|\int_D \Sigma^*\,\nabla\cdot(\gamma\,\nabla\,w) + \int_D\gamma\,\nabla\,\Sigma^*\cdot\nabla\,w\Big|\\ &\leq \Big|\int_D\Sigma^*\,\nabla\cdot\Big((\gamma_0-\gamma)\,\nabla\,u_{\gamma_0,f}\Big)\Big| + \|\gamma\|_\infty\|\Sigma\|_{H^1(D)}\|\nabla\,w\|_{L^2(D)}\\ &\leq \Big|\int_D(\gamma_0-\gamma)\,\nabla\,\Sigma^*\cdot\nabla\,u_{\gamma_0,f}\Big| + CM_1\|\sigma\|_{H^{1/2}(\partial D)}\|\nabla\,w\|_{L^2(D)}\\ &\leq C\|\sigma\|_{H^{1/2}(\partial D)}\Big(\|\gamma_0-\gamma\|_\infty\|\nabla\,u_{\gamma_0,f}\|_{L^2(D)} + M_1\|\nabla\,w\|_{L^2(D)}\Big), \end{aligned}$$

hence

$$\Big\|\frac{\partial w}{\partial\nu}\Big\|_{H^{-1/2}(\partial D)} \leq C\Big(\|\gamma_0-\gamma\|_\infty\|\nabla\,u_{\gamma_0,f}\|_{L^2(D)} + M_1\|\nabla\,w\|_{L^2(D)}\Big). \tag{3.32}$$

Next, again by the divergence theorem, we have for any $v \in H_0^1(D)$

$$\int_D\gamma\,\nabla\,w\cdot\nabla\,v^* = \int_D(\gamma_0-\gamma)\,\nabla\,u_{\gamma_0,f}\cdot\nabla\,v^*.$$

In particular this applies with $v = w$, hence, since $\gamma \geq m$ on D, we apply the Cauchy–Schwarz inequality to deduce

$$m\|\nabla w\|^2_{L^2(D)} \leq \|\gamma_0 - \gamma\|_\infty \|\nabla u_{\gamma_0,f}\|_{L^2(D)} \|\nabla w\|_{L^2(D)},$$

which, returning to (3.32), shows that

$$\left\|\frac{\partial w}{\partial \nu}\right\|_{H^{-1/2}(\partial D)} \leq C\|\nabla u_{\gamma_0,f}\|_{L^2(D)} \|\gamma_0 - \gamma\|_\infty \left(1 + \tfrac{M_1}{m}\right).$$

Applying Theorem 1.8 to each representative of the equivalence class $f \in H^{1/2}(D)/\mathbb{C}$ as for σ and optimising, there exists $F \in H^1(D)/\mathbb{C}$ such that $F|_{\partial D} = f$ and $\|F\|_{H^1(D)/\mathbb{C}} \leq C\|f\|_{H^{1/2}(\partial D)/\mathbb{C}}$ for a constant $C = C(D)$. Recall by definition of a weak solution (3.4),

$$\int_D \gamma_0 \nabla u_{\gamma_0,f} \cdot \nabla (u_{\gamma_0,f} - F)^* = 0,$$

and arguing as with w we deduce

$$\|\nabla u_{\gamma_0,f}\|_{L^2(D)} \leq C\tfrac{M_0}{m}\|f\|_{H^{1/2}(\partial D)/\mathbb{C}}.$$

Overall we have shown

$$\left\|(\Lambda_\gamma - \Lambda_{\gamma_0})f\right\|_{H^{-1/2}(\partial D)} \leq C\Big(1 + \frac{M_1}{m}\Big)\frac{M_0}{m}\|\gamma - \gamma_0\|_\infty \|f\|_{H^{1/2}(\partial D)/\mathbb{C}}.$$

Taking the supremum over all f with $H^{1/2}(\partial D)/\mathbb{C}$ norm equal to 1, (3.30) follows. □

Proof of Lemma 3.7. Theorem 1 in Alessandrini [4] (see the lecture notes [75] for an alternative exposition) states that there exist constants $\delta = \delta(d)$ and $C = C(M, m, D, D', \beta)$ such that there is a (monotone) function ω satisfying

$$\|\gamma - \gamma_0\|_\infty \leq C\omega(\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_*), \quad \omega(t) \leq \log(1/t)^{-\delta} \text{ for } t < e^{-1}. \tag{3.33}$$

Appealing to Lemma 3.5, noting that M upper bounds $\|\gamma\|_\infty$ and $\|\gamma_0\|_\infty$ by a Sobolev embedding (recall Theorem 1.12), we see for a constant C' depending on M, m, D, D' and r that

$$\begin{aligned}\omega(\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_*) &\leq \omega(C'\|\Lambda_\gamma - \Lambda_{\gamma_0}\|^{1/2}_{\mathbb{H}_r}) \\ &\leq (\tfrac{1}{4}\log(\|\Lambda_\gamma - \Lambda_{\gamma_0}\|^{-1}_{\mathbb{H}_r}))^{-\delta},\end{aligned}$$

provided $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} < \min(e^{-2}(C')^{-2}, (C')^{-4}, 1)$. The result follows. □

3.3.3 Concentration of an estimator and prior support properties

In this section we prove two main auxiliary results, Lemmas 3.8 and 3.11, that will allow us to apply Theorem 1.5. We start with the following proof that there exists an estimator of Λ_γ exhibiting adequate concentration properties. Recall Γ_{m_1,D_1} is a superset of Γ_{m_0,D_0} on which the prior (3.21) concentrates its mass.

Lemma 3.8. *Let $\eta_\varepsilon > 0$ satisfy $\eta_\varepsilon \varepsilon^{-(1-\delta)} \to \infty$ as $\varepsilon \to 0$ for some $0 < \delta < 1$. There exists an estimator $\hat{\Lambda}$ for which, given $\kappa > 0, M > 0$ there is a constant $C = C(\kappa, m_1, D_1, M, D)$ so that for all ε small enough,*

$$\sup\{P_\varepsilon^\gamma(\|\hat{\Lambda} - \tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} > C\eta_\varepsilon) : \gamma \in \Gamma_{m_1,D_1}, \|\gamma\|_\infty \le M\} \le e^{-\kappa(\eta_\varepsilon/\varepsilon)^2}. \tag{3.34}$$

Proof. Define the estimator $\hat{\Lambda}$ by $\hat{\Lambda} = \sum_{j,k\le J} \hat{\Lambda}_{jk} b_{jk}^{(r)}$, where $J = J_\varepsilon = \lfloor \eta_\varepsilon/\varepsilon \rfloor$ and

$$\hat{\Lambda}_{jk} = \langle Y, b_{jk}^{(r)} \rangle_{\mathbb{H}_r} = \langle \tilde{\Lambda}_\gamma \phi_j^{(r)}, \phi_k^{(0)} \rangle_{L^2(\partial D)} + \varepsilon g_{jk}, \ Y \sim P_\varepsilon^\gamma, \tag{3.35}$$

where we note $g_{jk} = \langle \mathbb{W}, b_{jk}^{(r)} \rangle_{\mathbb{H}_r} \overset{iid}{\sim} N(0,1)$. Then we have the bias-variance decomposition

$$P_\varepsilon^\gamma(\|\hat{\Lambda} - \tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} > C\eta_\varepsilon) \le \mathbb{1}\{\|\tilde{\Lambda}_\gamma - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} > \tfrac{1}{2}C\eta_\varepsilon\} + P_\varepsilon^\gamma(\|\hat{\Lambda} - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} > \tfrac{1}{2}C\eta_\varepsilon). \tag{3.36}$$

Recall, by Lemma 3.4, for any $\nu > 0$ there is a constant $C_1 = C_1(\nu, r, M_1, m_1, D_1, D)$ such that

$$\|\tilde{\Lambda}_\gamma - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} \le C_1 J^{-\nu}, \tag{3.37}$$

hence the indicator in (3.36) is bounded by $\mathbb{1}\{C_1 J^{-\nu} > \frac{1}{2}C\eta_\varepsilon\}$. Choosing $\nu > (1-\delta)/\delta$, one finds that the assumption $\eta_\varepsilon \varepsilon^{-(1-\delta)} \to \infty$ ensures this term vanishes for ε small enough.

For the variance term in (3.36), observe that by Parseval's identity

$$\|\hat{\Lambda} - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r}^2 = \sum_{j,k\le J} (\hat{\Lambda}_{jk} - \langle \tilde{\Lambda}_\gamma \phi_j^{(r)}, \phi_k^{(0)} \rangle_{L^2(\partial D)})^2 = \varepsilon^2 \sum_{j,k\le J} g_{jk}^2.$$

One now applies a standard tail inequality (e.g. Theorem 3.1.9 in [38]) to the effect that

$$\Pr\Big(\sum_{j,k\le J} g_{jk}^2 \ge J^2 + 2J\sqrt{x} + 2x \Big) \le e^{-x}. \tag{3.38}$$

For a constant $\kappa > 0$, taking $x = \kappa(\eta_\varepsilon/\varepsilon)^2$, and for our choice $J = \lfloor \eta_\varepsilon/\varepsilon \rfloor$, we see that for C large enough depending only on κ, we have

$$P_\varepsilon^\gamma(\|\hat{\Lambda} - \pi_{JJ}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r}^2 > \tfrac{1}{2}C\eta_\varepsilon^2) \leq e^{-\kappa(\eta_\varepsilon/\varepsilon)^2},$$

hence the result. □

To proceed recall $K(p,q) = E_{X\sim p}\log\frac{p}{q}(X)$ denotes the Kullback–Leibler divergence between distributions with densities p and q, and recall the definition of the probability densities p_ε^γ from (3.15). Also denote by Var_γ the variance operator associated to the probability measure P_ε^γ. The following is then a standard result for a white noise model on a Hilbert space.

Lemma 3.9. *Let $\gamma_0, \gamma_1 \in \Gamma_{m_1,D_1}$. Then we have*

$$K(p_\varepsilon^{\gamma_0}, p_\varepsilon^{\gamma_1}) = \frac{1}{2}\varepsilon^{-2}\|\Lambda_{\gamma_0} - \Lambda_{\gamma_1}\|_{\mathbb{H}_r}^2, \text{ and } \mathrm{Var}_{\gamma_0}\Big(\log\frac{p_\varepsilon^{\gamma_0}}{p_\varepsilon^{\gamma_1}}\Big) = \varepsilon^{-2}\|\Lambda_{\gamma_0} - \Lambda_{\gamma_1}\|_{\mathbb{H}_r}^2.$$

Proof. Using the explicit formula (3.15) for the log-likelihoods, we see that under γ_1,

$$\begin{aligned}\ell(\gamma_0) - \ell(\gamma_1) &= \varepsilon^{-2}\langle Y, \tilde{\Lambda}_{\gamma_0} - \tilde{\Lambda}_{\gamma_1}\rangle_{\mathbb{H}_r} - \tfrac{1}{2}\varepsilon^{-2}\|\tilde{\Lambda}_{\gamma_0}\|_{\mathbb{H}_r}^2 + \tfrac{1}{2}\varepsilon^{-2}\|\tilde{\Lambda}_{\gamma_1}\|_{\mathbb{H}_r}^2 \\ &= \tfrac{1}{2}\varepsilon^{-2}\|\tilde{\Lambda}_{\gamma_0} - \tilde{\Lambda}_{\gamma_1}\|_{\mathbb{H}_r}^2 + \varepsilon^{-1}\langle \mathbb{W}, \tilde{\Lambda}_{\gamma_1} - \tilde{\Lambda}_{\gamma_2}\rangle_{\mathbb{H}_r},\end{aligned}$$

which is normally distributed with mean $\frac{1}{2}\varepsilon^{-2}\|\tilde{\Lambda}_{\gamma_0} - \tilde{\Lambda}_{\gamma_1}\|_{\mathbb{H}_r}^2$ and variance $\varepsilon^{-2}\|\tilde{\Lambda}_{\gamma_0} - \tilde{\Lambda}_{\gamma_1}\|_{\mathbb{H}_r}^2$. Noting that $\tilde{\Lambda}_{\gamma_0} - \tilde{\Lambda}_{\gamma_1} = \Lambda_{\gamma_0} - \Lambda_{\gamma_1}$, we deduce the result. □

A variant of the "small balls" of (1.12) appropriate to the setting here defines $B_{KL}^\varepsilon(\eta)$ as

$$B_{KL}^\varepsilon(\eta) = \{\gamma \in \Gamma_{m_1,D_1} : K(p_\varepsilon^{\gamma_0}, p_\varepsilon^\gamma) \leq (\eta/\varepsilon)^2, \mathrm{Var}_{\gamma_0}(\log(p_\varepsilon^{\gamma_0}/p_\varepsilon^\gamma)^2) \leq (\eta/\varepsilon)^2\}. \tag{3.39}$$

Then the following is an immediate consequence of Lemma 3.9.

Corollary 3.10. *For any $\eta > 0$, $\{\gamma \in \Gamma_{m_1,D_1} : \|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq \eta\} \subseteq B_{KL}^\varepsilon(\eta)$.*

With the preceding preparations, we can now prove the following support result for the prior Π of Theorem 3.3, using a result of Li & Linde [52].

Lemma 3.11. *Let $\eta_\varepsilon = \varepsilon^{\alpha/(\alpha+d)}$. Under the conditions of Theorem 3.3, there exists a constant $\zeta = \zeta(\alpha, m_1, M, D, D_1, \Phi, r) > 0$ such that $\Pi(B_{KL}^\varepsilon(\eta_\varepsilon)) \geq e^{-\zeta(\eta_\varepsilon/\varepsilon)^2}$ for all ε small enough, uniformly across across γ_0 in the set (3.23).*

Proof. By (3.18) and a Sobolev embedding, there is a constant M_0 depending only on Φ, α and M such that $\|\gamma_0\|_\infty \leq M_0$. By Lemma 3.6, we deduce, for a constant $C = C(r, D, D_1)$, that

$$\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq C\frac{M_0\|\gamma\|_\infty}{m_1^2}\|\gamma - \gamma_0\|_\infty^{1/2}$$

provided $\|\gamma - \gamma_0\|_\infty$ is small enough. It follows from this calculation and Corollary 3.10 that for η_ε small enough and some constant $C' > 0$ we have

$$\{\|\gamma - \gamma_0\|_\infty \leq C'\eta_\varepsilon^2\} \subseteq \{\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq \eta_\varepsilon\} \subseteq B_{KL}^\varepsilon(\eta_\varepsilon).$$

Defining $\rho_0 = \Phi^{-1} \circ \gamma_0$ and appealing again to (3.18), we further deduce that

$$\{\rho \in C_u(D) : \|\rho - \rho_0\|_\infty \leq A\eta_\varepsilon^2\} \subseteq B_{KL}^\varepsilon(\eta_\varepsilon),$$

for a constant $A = A(\alpha, M, m_1, D, D_1, r, \Phi)$. Recalling the definition of Π_ρ from (3.17), it therefore suffices to lower bound $\Pi_\rho\Big(\|\rho - \rho_0\|_\infty \leq A\eta_\varepsilon^2\Big)$. Note that Π_ρ has RKHS $\mathcal{H}_\varepsilon = \{\rho'\chi : \rho' \in \mathcal{H}\}$, with norm $\|\cdot\|_{\mathcal{H}_\varepsilon}$ satisfying the bound $\|\rho\|_{\mathcal{H}_\varepsilon} \leq \varepsilon^{-d/(\alpha+d)}\|\rho'\|_{\mathcal{H}} = (\eta_\varepsilon/\varepsilon)\|\rho'\|_{\mathcal{H}}$, for any ρ' such that $\rho'\chi = \rho$. Since $\rho_0 = \rho_0\chi$, we deduce that $\|\rho_0\|_{\mathcal{H}_\varepsilon} \leq (\eta_\varepsilon/\varepsilon)\|\rho_0\|_{\mathcal{H}} \leq M\eta_\varepsilon/\varepsilon$. By Corollary 2.6.18 in [38], we then have

$$\begin{aligned}\Pi_\rho(\|\rho - \rho_0\|_\infty \leq A\eta_\varepsilon^2) &\geq e^{-\frac{1}{2}\|\rho_0\|_{\mathcal{H}_\varepsilon}^2}\Pi_\rho(\|\rho\|_\infty \leq A\eta_\varepsilon^2)\\ &\geq e^{-\frac{1}{2}M^2(\eta_\varepsilon/\varepsilon)^2}\Pi'\Big(\|\rho'\|_\infty \leq A\tfrac{\eta_\varepsilon^3}{\varepsilon}\Big).\end{aligned}$$

Next, since $\mathcal{H}$ embeds continuously into $H^\alpha(I_d)$ for some large enough cube I_d (by a standard extension argument for Sobolev spaces), the unit ball $B_{\mathcal{H}}$ of $\mathcal{H}$ has covering numbers with respect to the supremum norm $N = N(B_{\mathcal{H}}, \|\cdot\|_\infty, \delta)$ satisfying

$$N(B_{\mathcal{H}}, \|\cdot\|_\infty, \delta) \leq K\delta^{-d/\alpha} \tag{3.40}$$

for some constant $K = K(\alpha, D)$ (see [38], equations (4.184) and (4.185)). We can thus apply [52] Theorem 1.2, to see

$$\Pi'\Big(\|\rho\|_\infty \leq A\tfrac{\eta_\varepsilon^3}{\varepsilon}\Big) \geq e^{-A'(\frac{\eta_\varepsilon^3}{\varepsilon})^{-s}},$$

for some constant $A' = A'(A, K)$, where s is such that $\frac{d}{\alpha} = \frac{2s}{2+s}$, i.e. $s = \frac{2d}{2\alpha - d}$.

Overall, we have shown $\Pi(B^{\varepsilon}_{KL}(\eta_\varepsilon)) \geq e^{-\frac{1}{2}M^2(\eta_\varepsilon/\varepsilon)^2} e^{-A'(\eta_\varepsilon^3/\varepsilon)^{-2d/(2\alpha-d)}}$, for a constant A' depending only on $D, \alpha, M, m_1, D_1, r, \Phi$. For $\eta_\varepsilon = \varepsilon^{\alpha/(\alpha+d)}$ we find $(\eta_\varepsilon^3/\varepsilon)^{-2d/(2\alpha-d)} = (\eta_\varepsilon/\varepsilon)^2$, and the result follows. □

3.3.4 Posterior contraction proofs

Posterior regularity and contraction about Λ_{γ_0}

Following ideas for Bayesian nonparametric statistics with Gaussian priors as in van der Vaart–van Zanten [85], we prove the following prior regularity result as a final auxiliary result before proceeding as in [59] to prove Theorem 3.3.

Lemma 3.12. *Under the assumptions of Theorem 3.3 and for η_ε, ζ as in Lemma 3.11, there exists $M' > 0$ such that*

$$\Pi(\|\gamma\|_{H^\beta(D)} > M') \leq e^{-(2\zeta+8)(\eta_\varepsilon/\varepsilon)^2}. \tag{3.41}$$

Proof. Recalling (3.19) and the definition of the prior (3.21),

$$\Pi(\|\gamma\|_{H^\beta(D)} > M') \leq \Pi'\Big(\|\rho'\|_{H^\beta(D)} > \frac{\eta_\varepsilon}{\varepsilon}\|\chi\|^{-1}_{H^\beta(D)}(M'/C' - 1)^{1/\beta}\Big).$$

Since $\eta_\varepsilon/\varepsilon \to \infty$ and since $\Pi'(H^\beta) = 1$ by hypothesis, we can apply a version of Fernique's theorem, more specifically Theorem 2.1.20 in [38], to deduce that for any $c > 0$ there exists a $M' = M'(c, C', \beta, \chi)$ such that the last probability does not exceed $e^{-c(\eta_\varepsilon/\varepsilon)^2}$. Taking $c = 2\zeta + 8$ concludes the proof. □

Proof of Theorem 3.3. This follows immediately from Theorem 1.5 as follows. With $\eta_\varepsilon, \zeta, M'$ as in Lemma 3.12, we make the notational changes $\{n \to \infty\} \leftrightarrow \{\varepsilon \to 0\}$, $\varepsilon_n \leftrightarrow \eta_\varepsilon$, $n\varepsilon_n^2 \leftrightarrow (\eta_\varepsilon/\varepsilon)^2$, $\Theta \leftrightarrow \Gamma_{m_1,D_1}$, $\theta \leftrightarrow \gamma$, $G(\theta) \leftrightarrow \Lambda_\gamma$, $B^n_{KL} \leftrightarrow B^\varepsilon_{KL}(\eta_\varepsilon)$, and $\xi_n \leftrightarrow K\xi_{\varepsilon,\delta}$ (for a constant K to be chosen). Further, we define

- $\tilde{d}$ as the $\mathbb{H}_r$ norm distance, d as the $\|\cdot\|_\infty$ norm distance;
- $\Theta_n \leftrightarrow \Gamma_\varepsilon := \{\gamma \in \Gamma_{m_1,D_1} : \|\gamma\|_{H^\beta(D)} \leq M'\}$, which we note is $\|\cdot\|_\infty$–bounded by a Sobolev embedding;
- $\tilde{\Theta} \leftrightarrow \tilde{\Gamma} := \Gamma_{m_0,D_0} \cap \{\Phi \circ \rho : \|\rho\|_{\mathcal{H}} \leq M\}$, which we note is $\|\cdot\|_{H^\beta(D)}$–bounded and $\|\cdot\|_\infty$–bounded, by (3.19) and a Sobolev embedding.

For the conditions of Theorem 1.5, we see that Lemma 3.12 yields (a), Lemma 3.8 yields (b), and Lemma 3.11 yields (c), uniformly across $\gamma_0 \in \tilde{\Gamma}$. For (d) let C be such that

we achieve (b) with constant C; then for $\gamma \in \Gamma_\varepsilon$ and $\|\Lambda_\gamma - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq C\eta_\varepsilon$, we see by Lemma 3.7 that there exists a constant K such that $\|\gamma - \gamma_0\|_\infty \leq K\xi_{\varepsilon,\delta}$, at least if ε is small enough. Finally, note that $\xi_{\varepsilon,\delta}^{-1} e^{-(\eta/\varepsilon)^2} \to 0$ as $\varepsilon \to 0$, and note that since $E^{\Pi'}\|\rho'\|_\infty^2$ is finite (see exercise 2.1.2 in [38]), (3.18) implies that $E^{\Pi}\|\gamma\|_\infty^2$ is bounded. □

3.3.5 Proof of the lower bound Theorem 3.2

Recall the shorthand (3.22), and recall by Theorem 1.2 that it suffices to find $\gamma_0, \gamma_1 \in \Gamma^\alpha_{m_0,D_0}(M)$ such that, for some μ small enough,

(1) $\|\gamma_1 - \gamma_0\|_\infty \geq \xi_{\varepsilon,\delta'}$,

(2) $K(p_\varepsilon^{\gamma_1}, p_\varepsilon^{\gamma_0}) \leq \mu$.

We appeal to Corollary 1 in [56], which says that for any integer $k \geq 2$, any $q > 0$, some $B > 0$ and any $\xi > 0$ sufficiently small there exist γ_0, γ_1 such that $\operatorname{supp}(\gamma_j - 1) \subseteq D_0$, $\gamma_j \geq 1$ on D for $j = 0, 1$, and

a. $\|\gamma_1 - \gamma_0\|_\infty \geq \xi$,

b. $\|\Lambda_{\gamma_1} - \Lambda_{\gamma_0}\|_{H^{-q}(\partial D)/\mathbb{C} \to H^q_\diamond(\partial D)} \leq \exp\Big(-\xi^{-\frac{d}{(2d-1)k}}\Big)$,

c. $\max(\|\gamma_1\|_{C^k(D)}, \|\gamma_0\|_{C^k(D)}) \leq B$,

where $C^k(D)$ is the usual space consisting of functions with bounded continuous partial derivatives up to order k. (Note that [56] states this with full norm $H^{-q}(\partial D)$ in place of the quotient norm, but since Λ_{γ_j} maps constant functions to 0 for $j = 0, 1$, the two norms coincide.) For $k = \alpha$ and $B = M$, noting $H^\alpha(D) \supset C^\alpha(D)$, we deduce there exist such $\gamma_0, \gamma_1 \in \Gamma^\alpha_{m_0,D_0}(M)$. Taking $\xi = \xi_{\varepsilon,\delta'}$ we note that (1) holds by definition.

For (2), applying Lemma 3.16 with $p = \min(d-1, r)$ and $q = (d-1-r)_+ \equiv \max(d-1-r, 0) = d-1-p$ we see that, for a constant $C = C(d, r)$,

$$\|\Lambda_{\gamma_1} - \Lambda_{\gamma_0}\|_{\mathbb{H}_r} \leq C\|\Lambda_{\gamma_1} - \Lambda_{\gamma_0}\|_{H^{-q} \to H^q}.$$

Thus, appealing to Lemma 3.9, we can bound the Kullback–Leibler divergences $K(p_\varepsilon^{\gamma_1}, p_\varepsilon^{\gamma_0})$ by

$$\varepsilon^{-2}\|\Lambda_{\gamma_1} - \Lambda_{\gamma_0}\|^2_{\mathbb{H}_r} \leq C^2\varepsilon^{-2}\|\Lambda_{\gamma_1} - \Lambda_{\gamma_0}\|^2_{H^{-q} \to H^q} \leq C^2 \exp\Big[2\log(1/\varepsilon) - (\log(1/\varepsilon))^{-\frac{\delta' d}{(2d-1)\alpha}}\Big].$$

Since $\delta' > \alpha(2d-1)/d$ by assumption, the final expression tends to zero as $\varepsilon \to 0$ and is smaller than any μ required in condition (2) for ε small enough.

Appendix 3.A Laplace–Beltrami eigenfunctions

Here we define the Sobolev spaces $H^s(\partial D)$ in terms of a orthonormal basis $(\phi_k)_{k\in\mathbb{N}\cup\{0\}} = (\phi_k^{(0)})_{k\in\mathbb{N}\cup\{0\}}$ of $L^2(\partial D)$ consisting of eigenfunctions of the Laplace–Beltrami operator $\Delta_{\partial D}$, and outline some important properties of such a basis. Basic properties can be found, for example, in Chavel [17], Chapter I. Let $\lambda_k > 0$ be the corresponding eigenvalues, which we assume to have been sorted in increasing order:

$$-\Delta_{\partial D}\,\phi_k = \lambda_k \phi_k.$$

Definition $(H^r(\partial D))$**.** For $r \geq 0$, we define

$$H^r(\partial D) = \{f \in L^2(\partial D) \text{ s.t. } \sum_{k=0}^{\infty}(1+\lambda_k)^r |\langle f, \phi_k\rangle_{L^2(\partial D)}|^2 =: \|f\|^2_{H^r(\partial D)} < \infty\},$$

where the space $L^2(\partial D)$ is defined relative to the surface element on ∂D (and with complex scalars).

For $r < 0$, we define $H^r(\partial D)$ as the completion of $L^2(\partial D)$ with respect to the norm $\|\cdot\|_{H^r(\partial D)}$.

Remarks. i. It is immediate from the definitions that $\{\phi_k\}$ is an orthogonal spanning set of $H^r(\partial D)$, and that setting $\phi_k^{(r)} = (1+\lambda_k)^{-r/2}\phi_k$ yields an orthonormal basis of $H^r(\partial D)$.

ii. This definition of $H^r(\partial D)$ coincides with the previous definitions. For example, for $r = 1$ the calculation

$$\int_{\partial D} \nabla\,\phi_k \cdot \nabla\,\phi_l^* = -\int_{\partial D} \phi_k\,\Delta_{\partial D}\,\phi_l^* = \lambda_l^* \int_{\partial D} \phi_k \phi_l^* = \lambda_l \delta_{kl},$$

derived via the divergence theorem for a manifold (e.g. see [17] eq (35); note that the manifold ∂D is compact) implies that our definition of $\|\cdot\|_{H^1(\partial D)}$ is equivalent to the standard definition $\|f\|^2_{H^1(\partial D)} = \|f\|^2_{L^2(\partial D)} + \|\nabla\,f\|^2_{L^2(\partial D)}$, and inductively the same is true for $H^r(\partial D)$, $r \in \mathbb{N}$.

For the equivalence of this definition with some other definitions of negative or non-integer Sobolev spaces, see [53] Chapter I Section 7.3 (p34-37). In particular note that $H^{-s}(\partial D)$ is the topological dual space of $H^s(\partial D)$ for any $s \in \mathbb{R}$.

iii. Note that ϕ_0 is a constant function, hence the $H^r(\partial D)/\mathbb{C}$ norm, defined by $\|[f]\|_{H^r(\partial D)/\mathbb{C}} = \inf_{z\in\mathbb{C}}\|f - z\|$ for $[f]$ the equivalence class over $\mathbb{C}$ of a function

$f \in H^r(\partial D)$, can also be characterised by

$$\|[f]\|^2_{H^r(\partial D)/\mathbb{C}} = \sum_{k=1}^{\infty}(1+\lambda_k)^r|\langle f, \phi_k\rangle_{L^2(\partial D)}|^2. \tag{3.42}$$

Recall also we defined $H^s_\diamond(\partial D) = \{g \in H^s(\partial D) : \langle g, 1\rangle_{L^2(\partial D)} = 0\}$. Note that each $[f] \in H^s(\partial D)/\mathbb{C}$ has a representative $g \in H^s_\diamond(\partial D)$, and $\|f\|_{H^s(\partial D)/\mathbb{C}} = \|g\|_{H^s(\partial D)}$. We thus use the norm (3.42) on spaces $H^s(\partial D)/\mathbb{C}$ and on $H^s_\diamond(\partial D)$ without further mention. We also typically write f for the equivalence class $[f]$ and only comment further on this where necessary.

This "spectral" definition of $H^r(\partial D)$ is useful particularly because Weyl's law allows us to understand the scaling of λ_k with k fairly explicitly.

Lemma 3.13 (Weyl's law on a compact closed manifold, e.g. [17] eq.(49))**.** *Suppose M is a closed compact manifold of dimension d. Then*

$$N(\lambda) = (2\pi)^{-d}\lambda^{d/2}\omega_d \operatorname{Vol}(M) + o(\lambda^{d/2}),$$

where $N(\lambda)$ is the number of eigenvalues (counted with multiplicity) no bigger than λ and ω_d is the volume of a unit disc in $\mathbb{R}^d$.

Corollary 3.14. *For constants C_1, C_2 depending only on D, the eigenvalues of the Laplace–Beltrami operator $\Delta_{\partial D}$ satisfy $C_1 k^{2/(d-1)} \le \lambda_k \le C_2 k^{2/(d-1)}$. Hence, the eigenfunctions satisfy*

$$C_3(1+k^{\frac{1}{d-1}})^{s-r} \le \|\phi_k^{(r)}\|_{H^s(\partial D)} \le C_4(1+k^{\frac{1}{d-1}})^{s-r}, \ \ s, r \in \mathbb{R}, \tag{3.43}$$

for constants C_3 and C_4 depending only on ∂D and on the difference $s-r$. For $k > 0$ the same expression holds with the quotient norm $\|\phi_k^{(r)}\|_{H^s(\partial D)/\mathbb{C}}$ in place of $\|\phi_k^{(r)}\|_{H^s(\partial D)}$.

Proof. We apply Weyl's law on the manifold ∂D, which has dimension $d-1$. Writing $N(\lambda^-)$ for $\lim_{x\uparrow\lambda} N(x)$ and $N(\lambda^+)$ for $\lim_{x\downarrow\lambda} N(x)$, we thus have

$$N(\lambda_k^-) \le k \le N(\lambda_k^+).$$

It follows that $C\lambda_k^{(d-1)/2} + o(\lambda_k^{(d-1)/2}) \le k \le C\lambda_k^{(d-1)/2} + o(\lambda_k^{(d-1)/2})$ for the constant $C = C(D) = (2\pi)^{-(d-1)}\omega_{d-1}\operatorname{Area}(\partial D)$ and hence we deduce the scaling of the eigenvalues. Then (3.43) follows from the first remark after the definition of $H^r(\partial D)$. □

Appendix 3.B Comparison results between Hilbert–Schmidt operators

For separable Hilbert spaces A and B, we recall the notations $\mathcal{L}(A, B)$ for the space of bounded linear maps $A \to B$ equipped with the operator norm $\|\cdot\|_{A\to B}$, and $\mathcal{L}_2(A, B)$ for the space of Hilbert–Schmidt operators $A \to B$. Generalising (3.26), define the orthonormal basis $(b_{jk}^{(p,q)})$ of $\mathcal{L}_2(H^p, H^q)$ by

$$b_{jk}^{(p,q)}(f) = (\phi_j^{(p)}) \otimes \phi_k^{(q)}(f) = \langle f, \phi_j^{(p)}\rangle_{H^p}\phi_k^{(q)}, \quad j, k \in \mathbb{N}$$

(in this section we omit explicit reference to the domain, writing H^p for either $H^p(\partial D)/\mathbb{C}$ or $H^p_\diamond(\partial D)$; as noted in the remark in Appendix 3.A, both spaces can be identified with $\operatorname{span}\{\phi_k^{(p)} : k \geq 1\}$, hence the omission should not cause confusion). The compatibility between our bases of $H^p(\partial D)$ for different $p \in \mathbb{R}$ means that the subspaces spanned by $(b_{jk}^{(p,q)})_{j\leq J, k\leq K}$ coincide for all p and q, and the $\mathcal{L}_2(H^p, H^q)$ projections onto this subspace coincide with π_{JK} (as defined in (3.27)). Corollary 3.14 implies the following results controlling Hilbert–Schmidt norms for different domains and codomains in terms of each other, and in terms of operator norms.

Lemma 3.15. *Let $T \in \operatorname{span}\{b_{jk}^{(r)} : 1 \leq j \leq J,\ 1 \leq k \leq K\}$, where we recall that $b_{jk}^{(r)} = (\phi_j^{(r)}) \otimes \phi_k^{(0)}$. Then, for a constant C depending only on D and on the differences $r - p$, $s - q$, we have*

$$\|T\|_{\mathcal{L}_2(H^r,H^s)} \leq C(1 + J^{1/(d-1)})^{(p-r)_+}(1 + K^{1/(d-1)})^{(s-q)_+}\|T\|_{\mathcal{L}_2(H^p,H^q)},$$

where $x_+ = \max(x, 0)$ for $x \in \mathbb{R}$.

Proof. The coefficients $a_{jk}^{(r,s)}$ of T with respect to the basis $(b_{jk}^{(r,s)})$ are given by

$$a_{jk}^{(r,s)} = \langle T, b_{jk}^{(r,s)}\rangle_{\mathcal{L}_2(H^r,H^s)} = \langle T\phi_j^{(r)}, \phi_k^{(s)}\rangle_{H^s}$$

and we see from Corollary 3.14 that

$$|a_{jk}^{(r,s)}| \leq C(1 + j^{1/(d-1)})^{p-r}(1 + k^{1/(d-1)})^{s-q}|a_{jk}^{(p,q)}| \tag{3.44}$$

for a constant C depending only on D and the differences $r - p, s - q$.

Upper bounding $(1+j^{1/(d-1)})^{(p-r)} \leq (1+J^{1/(d-1)})^{(p-r)_+}$ for $j \leq J$, and similarly for k, we find that

$$\|T\|^2_{\mathcal{L}_2(H^r,H^s)} = \sum_{j\leq J, k\leq K} |a_{jk}^{(r,s)}|^2 \leq C(1+J^{1/(d-1)})^{2(p-r)_+}(1+K^{1/(d-1)})^{2(s-q)_+}\|T\|^2_{\mathcal{L}_2(H^p,H^q)},$$

hence the result. □

Lemma 3.16. *For some $p \leq r$ and $q \geq s$, let $T \in \mathcal{L}(H^{p-(d-1)}, H^q)$. Then we have $T \in \mathcal{L}_2(H^r, H^s)$ and, for constant C depending only on D and the differences $r-p$, $q-s$,*

$$\|T - \pi_{JK}T\|_{\mathcal{L}_2(H^r,H^s)} \leq C\|T\|_{H^{p-(d-1)}\to H^q} \max\Big((1+J^{1/(d-1)})^{p-r}, (1+K^{1/(d-1)})^{s-q}\Big).$$

In the special case $J = K = 0$, we have

$$\|T\|_{\mathcal{L}_2(H^r,H^s)} \leq C\|T\|_{H^{p-(d-1)}\to H^q}.$$

Proof. Firstly, as a consequence of Corollary 3.14, we have, for a constant $C = C(D,d)$,

$$\|T\phi_j^{(p)}\|^2_{H^q} \leq \|T\|^2_{H^{p-(d-1)}\to H^q}\|\phi_j^{(p)}\|^2_{H^{p-(d-1)}} \leq C\|T\|^2_{H^{p-(d-1)}\to H^q}(1+j^{1/(d-1)})^{-2(d-1)}, \tag{3.45}$$

which is summable over j, hence $T \in \mathcal{L}_2(H^p, H^q)$ and $\|T\|_{\mathcal{L}_2(H^p,H^q)} \leq C'\|T\|_{H^{p-(d-1)}\to H^q}$. Note by monotonicity of H^α norms, we also have $T \in \mathcal{L}_2(H^r, H^s)$.

Since the $\mathcal{L}_2(H^r, H^s)$–orthogonal projection maps coincide for all r and s, defining $a_{jk}^{(r,s)}$ as in the previous proof, we have from (3.44) that for a constant C,

$$\begin{aligned}\|T - \pi_{JK}T\|^2_{\mathcal{L}_2(H^r,H^s)} &= \sum_{j>J \text{ or } k>K} |a_{jk}^{(r,s)}|^2 \\ &\leq C \sum_{j>J \text{ or } k>K} (1+j^{1/(d-1)})^{2(p-r)}(1+k^{1/(d-1)})^{2(s-q)}|a_{jk}^{(p,q)}|^2\end{aligned}$$

Since $p \leq r$ and $q \geq s$, we see that

$$\begin{aligned}\sum_{j>J}\sum_k (1+j^{1/(d-1)})^{2(p-r)}(1+k^{1/(d-1)})^{2(s-q)}|a_{jk}^{(p,q)}|^2 &\leq (1+J^{1/(d-1)})^{2(p-r)} \sum_{j>J}\sum_k |a_{jk}^{(p,q)}|^2 \\ &\leq (1+J^{1/(d-1)})^{2(p-r)}\|T\|^2_{\mathcal{L}_2(H^p,H^q)}.\end{aligned}$$

Arguing similarly for the sum over all j and over $k > K$, we deduce that

$$\|T - \pi_{JK}T\|^2_{\mathcal{L}_2(H^r,H^s)} \leq 2C\|T\|^2_{\mathcal{L}_2(H^p,H^q)} \max\Big((1+J^{1/(d-1)})^{2(p-r)}, (1+K^{1/(d-1)})^{2(s-q)}\Big).$$

The result follows. □

Appendix 3.C Mapping properties of Λ_γ and $\tilde{\Lambda}_\gamma$

In this appendix we prove the following mapping properties of Λ_γ and $\tilde{\Lambda}_\gamma$ which were used throughout the main body of this chapter.

Lemma 3.17. *Let $m > 0$ and let D' be a domain compactly contained in D. For $\gamma \in \Gamma_{m,D'}$ and $f \in H^{s+1}(\partial D)/\mathbb{C}$, there is a unique weak solution $u_{\gamma,f} \in \mathcal{H}_s = \left(H^{\min\{1,s+3/2\}}(D) \cap H^1_{\mathrm{loc}}(D)\right)/\mathbb{C}$ to the Dirichlet problem (3.1). Moreover, if $u_{1,f}$ is the unique solution when $\gamma = 1$, then for any other $\gamma \in \Gamma_{m,D'}$ bounded by M on D, $u_{\gamma,f} - u_{1,f}$ lies in $H^1_0(D)/\mathbb{C}$ and satisfies the estimate*

$$\|u_{\gamma,f} - u_{1,f}\|_{H^1(D)/\mathbb{C}} \leq C\tfrac{M}{m}\|f\|_{H^{s+1}(\partial D)/\mathbb{C}}, \tag{3.46}$$

for some constant $C = C(D, D', s)$.

Lemma 3.18. *For some $m > 0$ and some domain D' compactly contained in D, let $\gamma \in \Gamma_{m,D'}$. For each $s \in \mathbb{R}$, Λ_γ is a continuous linear map from $H^{s+1}(\partial D)/\mathbb{C}$ to $H^s_\diamond(\partial D)$, and it is continuously invertible. For each $s, t \in \mathbb{R}$, the shifted operator $\tilde{\Lambda}_\gamma = \Lambda_\gamma - \Lambda_1$ is a continuous map from $H^s(\partial D)/\mathbb{C}$ to $H^t_\diamond(\partial D)$.*

Moreover, if γ also satisfies the bound $\|\gamma\|_\infty \leq M$, then we have the explicit bounds

$$\|\Lambda_\gamma\|_{H^{s+1}\to H^s} \leq C_1\frac{M}{m}, \tag{3.47}$$

$$\|\tilde{\Lambda}_\gamma\|_{H^s\to H^t} \leq C_2\frac{M}{m}, \tag{3.48}$$

for constants $C_1 = C(D, D', s)$ and $C_2 = C_2(D, D', s, t)$.

Given Lemma 3.18, the following is an immediate consequence of Lemma 3.16. Recall $\mathbb{H}_r$ was defined in (3.12).

Lemma 3.19. *For any $r \in \mathbb{R}$ and any $\gamma \in \Gamma_{m,D'}$, $\tilde{\Lambda}_\gamma \in \mathbb{H}_r$.*

A key to proving Lemmas 3.17 and 3.18 is the following basic fact about harmonic functions. For convenience of the reader, we include a proof (following Lemma A.1 in [42]). Note that as we have assumed $\gamma = 1$ on a neighbourhood $D \setminus D'$ of ∂D, our solutions $u_{\gamma,f}$ are harmonic on this neighbourhood.

Lemma 3.20 (Interior smoothness of harmonic functions)**.** *Let U_0, U be domains such that $U \Subset U_0$. Then for any $s, t \in \mathbb{R}$, there is a constant $C = C(s, t, U, U_0)$ such that for any harmonic function $v \in H^s(U)$,*

$$\|v\|_{H^s(U)/\mathbb{C}} \leq C\|v\|_{H^t(U_0)/\mathbb{C}}.$$

Proof. By monotonicity of H^t norms it suffices to prove the result for $s = t + k$ for $k \in \mathbb{N}$. Let $v \in H^t(U_0)$ represent the equivalence class and choose a domain U_1 such that $\bar{U} \subseteq U_1 \subseteq \bar{U}_1 \subseteq U_0$. Let ϕ be a smooth cutoff function, identically one on U_1 and compactly supported in U_0. For $z \in \mathbb{C}$ we observe that $\tilde{v} := (v - z)\phi$ satisfies

$$\begin{aligned} \Delta\, \tilde{v} &= F \quad \text{in } U_0, \\ \tilde{v} &= 0 \quad \text{on } \partial U_0, \end{aligned}$$

where $F = 2\,\nabla\,\phi \cdot \nabla\, v + (v - z)\,\Delta\,\phi$. Then

$$\|v\|_{H^{t+1}(U_1)/\mathbb{C}} \leq \|v - z\|_{H^{t+1}(U_1)} \leq \|\tilde{v}\|_{H^{t+1}(U_0)} \leq C\|F\|_{H^{t-1}(U_0)},$$

by Theorem 1.11. Note

$$\|F\|_{H^{t-1}(U_0)} \leq C(\phi)(\|v\|_{H^t(U_0)/\mathbb{C}} + \|v - z\|_{H^{t-1}(U_0)}),$$

and optimising across $z \in \mathbb{C}$ yields

$$\|v\|_{H^{t+1}(U_1)/\mathbb{C}} \leq C\|v\|_{H^t(U_0)/\mathbb{C}}. \tag{3.49}$$

Finally, we choose a finite sequence of domains $(U_j)_{1 \leq j \leq k}$ such that $U_k = U$ and, for $1 \leq j \leq k$, $\bar{U}_j \subset U_{j-1}$; applying (3.49) successively on each pair (U_j, U_{j-1}), we deduce the result. □

Proof of Lemma 3.17. Recall the remark in Section 1.6.3, which noted that a convenient way to show existence of a weak solution $u = u_{\gamma,f}$ is to decouple the task of satisfying the boundary values from the task of satisfying the PDE. Here, to access the wide body of theory for the Laplacian, we write $u = u_{1,f} + w$, where $u_{1,f}$ will be the solution with $\gamma = 1$ and $w \in H_0^1(D)$ will ensure the PDE is satisfied.

From standard theory for the Laplacian as in Theorem 1.11 (and Remark 2 thereafter), for $f \in H^{s+1}(\partial D)/\mathbb{C}$ there exists a solution $u_{1,f} \in H^{s+3/2}(D)/\mathbb{C}$ to

$$\begin{aligned} \Delta u &= 0 \quad \text{in } D, \\ u &= f \quad \text{on } \partial D, \end{aligned}$$

and this solution satisfies

$$\|u_{1,f}\|_{H^{s+3/2}(D)/\mathbb{C}} \leq C\|f\|_{H^{s+1}(\partial D)/\mathbb{C}} \tag{3.50}$$

for a constant $C = C(D, s)$. Also note that, as a harmonic function, $u_{1,f} \in H^1_{\text{loc}}(D)/\mathbb{C}$ by Lemma 3.20.

Now we show the existence of a weak solution $w \in H^1_0(D)$ to $L_\gamma[w] = -L_\gamma[u_{1,f}]$, where $L_\gamma[u] = \nabla \cdot (\gamma \nabla u)$. The operator L_γ is easily seen to be uniformly elliptic with ellipticity constant m. Recalling that $B_\gamma(w, v)$ denotes $\int_D \gamma \nabla w \cdot \nabla v^*$, the weak formulation of this PDE is

$$B_\gamma(w, v) = A(v) := -\int_D \gamma \nabla u_{1,f} \cdot \nabla v^*, \quad \forall v \in H^1_0(D).$$

In order to apply the general existence result Theorem 1.10, it remains to verify that the conjugate linear map A has bounded operator norm.

Since $\gamma = 1$ on $D \setminus D'$, for $v \in H^1_0(D)$ an application of the divergence theorem yields

$$\begin{aligned} -\int_{D\setminus D'} \gamma \nabla u_{1,f} \cdot \nabla v^* &= -\int_D \nabla u_{1,f} \cdot \nabla v^* + \int_{D'} \nabla u_{1,f} \cdot \nabla v^* \\ &= \int_D v^* \Delta u_{1,f} - \int_{\partial D} v^* \frac{\partial u_{1,f}}{\partial \nu} + \int_{D'} \nabla u_{1,f} \cdot \nabla v^* \\ &= \int_{D'} \nabla u_{1,f} \cdot \nabla v^* \end{aligned}$$

It follows, since $\|1 - \gamma\|_\infty \leq \|\gamma\|_\infty \leq M$, that

$$|A(v)| = \left| \int_{D'} (1 - \gamma) \nabla u_{1,f} \cdot \nabla v^* \right| \leq M\|v\|_{H^1(D)} \|\nabla u_{1,f}\|_{L^2(D')}.$$

By interior smoothness of harmonic functions (Lemma 3.20) and recalling (3.50), there are constants C and C' depending only on D, D' and s such that

$$\|\nabla u_{1,f}\|_{L^2(D')} \leq \|u_{1,f}\|_{H^1(D')/\mathbb{C}} \leq C'\|u_{1,f}\|_{H^{s+3/2}(D)/\mathbb{C}} \leq C\|f\|_{H^{s+1}(\partial D)/\mathbb{C}}.$$

Thus, $|A(v)| \leq CM\|f\|_{H^{s+1}(\partial D)}\|v\|_{H^1(D)}$. The existence of a unique solution $w \in H^1_0$, which satisfies (3.46), then follows by Theorem 1.10.

It remains to show that the (equivalence class of) function(s) u so constructed is the unique solution in $\mathcal{H}_s$ to (3.4). Since we have shown uniqueness of w, it is enough to show that the difference h between two $\mathcal{H}_s$ solutions lies in H^1_0, since then it must be the zero function. (We are considering h as a function, rather than an equivalence class of functions, which we can do by for example choosing a representative with average zero.) This is clear for $s \geq -1/2$ as then $\mathcal{H}_s \subset H^1$, and can be shown also for $s < -1/2$ as in [42], Theorem A.2 (the idea being that we know $h \in H^1_{\text{loc}}(D)$ so it suffices to prove $h \in H^1(D \setminus \Omega)$ for some $\Omega \Subset D$; this can be proved using harmonicity arguments as above to show the trace of h on $\partial\Omega$ is smooth enough that Theorem 1.11 gives the result). □

Proof of Lemma 3.18. We first remark that, by the divergence theorem,

$$\left\langle \tfrac{\partial u}{\partial \nu}, 1 \right\rangle_{L^2(\partial D)} = \int_{\partial D} \gamma \tfrac{\partial u}{\partial \nu} = \int_D \nabla \cdot (\gamma \nabla u) = 0$$

for a solution u to the Dirichlet problem (3.1), so that it suffices to prove (3.47), (3.48), and the continuity of $\Lambda_\gamma^{-1} : H^s_\diamond(\partial D) \to H^{s+1}(\partial D)/\mathbb{C}$.

We first prove (3.48), by adapting the proof of Theorem A.3 from [42] and tracking the constants. Given $f \in H^{s+1}(\partial D)/\mathbb{C}$ let $u_{\gamma,f} \in \mathcal{H}_s$ be the unique solution to the Dirichlet problem (3.1) and let $w \in H^1_0$ be a representative of the function class $u_{\gamma,f} - u_{1,f}$. Choose a domain Ω such that $D' \Subset \Omega \Subset D$. Choose also domains U, U_0 such that

$$\partial\Omega \subset U \Subset U_0 \Subset D \setminus D'.$$

Noting that w is harmonic on $D \setminus \bar{\Omega}$, the trace theorem (Theorem 1.8) applied to $w - z$ and optimised across $z \in \mathbb{C}$ yields

$$\|\partial w/\partial \nu\|_{H^t(\partial D)} \leq C\|w\|_{H^{t+3/2}(D\setminus\bar{\Omega})/\mathbb{C}}. \tag{3.51}$$

Applying Theorem 1.11 we see

$$\|w\|_{H^{t+3/2}(D\setminus\bar{\Omega})/\mathbb{C}} \leq C\Big(\|\operatorname{tr} w\|_{H^{t+1}(\partial D)/\mathbb{C}} + \|\operatorname{tr} w\|_{H^{t+1}(\partial\Omega)/\mathbb{C}}\Big) = C\|\operatorname{tr} w\|_{H^{t+1}(\partial\Omega)/\mathbb{C}}.$$

Again applying the trace theorem, this time on a subset of U bounded on one side by $\partial\Omega$, and applying Lemma 3.20, we see

$$\|\operatorname{tr} w\|_{H^{t+1}(\partial\Omega)/\mathbb{C}} \leq C\|w\|_{H^{t+3/2}(U)/\mathbb{C}} \leq C'\|w\|_{H^1(U_0)/\mathbb{C}}.$$

The constants in the above depend on D and (via Ω, U and U_0) on D', but are otherwise independent of γ. Recalling (3.46), which, because the smoothness of f here is s, tells us $\|w\|_{H^1(U_0)/\mathbb{C}} \leq C\left(\frac{M}{m}\right)\|f\|_{H^s(D)/\mathbb{C}}$, we overall have $\|\partial w/\partial\nu\|_{H^t(\partial D)} \leq C\left(\frac{M}{m}\right)\|f\|_{H^s(\partial D)/\mathbb{C}}$, so that we have proved (3.48).

Now we prove (3.47); given (3.48), it suffices to show $\|\Lambda_1\|_{H^{s+1}(\partial D)/\mathbb{C}\to H^s(\partial D)} \leq C\frac{M}{m}$ for an appropriate constant C. Since $u_{1,f}$ is harmonic on D, Theorem 1.8 yields

$$\|\partial u_{1,f}/\partial\nu\|_{H^s(\partial D)} \leq C\|u_{1,f}\|_{H^{s+3/2}(D)/\mathbb{C}}$$

for a constant $C = C(D, s)$, while Theorem 1.11 yields

$$\|u_{1,f}\|_{H^{s+3/2}(D)/\mathbb{C}} \leq C\|f\|_{H^{s+1}(\partial D)/\mathbb{C}}$$

for a constant $C = C(D, s)$, and (3.47) follows.

Finally we remark that the same arguments (see Theorem A.3 in [42]) applied to the inverse of Λ_γ, which is the *Neumann-to-Dirichlet map*, show that this is continuous from $H^s_\diamond(\partial D)$ to $H^{s+1}(\partial D)/\mathbb{C}$ as required. □

Appendix 3.D Statistical equivalence results for the noisy Calderón problem

In this appendix we rigorously state and prove the asymptotic equivalence results of Section 3.1. The notion of equivalence is in terms of Le Cam discrepancies, as defined Section 1.3.1, and, recalling that $K(P, Q) = K(p, q)$ denotes the Kullback–Leibler divergence between measures P, Q with densities p, q, we restate the following lemma.

Lemma (Lemma 1.1, restated)**.** *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be experiments with a common parameter set Θ: write $\mathcal{E}_j = (\mathcal{X}_j, \mathcal{F}_j, \{P_{j,\theta}\}_{\theta\in\Theta})$.*

a. Suppose further that the experiments are defined on a common probability space, i.e. that $\mathcal{X}_1 = \mathcal{X}_2$ and $\mathcal{F}_1 = \mathcal{F}_2$. Then

$$\Delta(\mathcal{E}_1, \mathcal{E}_2) \leq \sup_{\theta\in\Theta}\|P_{1,\theta} - P_{2,\theta}\|_{\mathrm{TV}} \leq \sup_{\theta\in\Theta}\sqrt{K(P_{1,\theta}, P_{2,\theta})/2}.$$

b. Let $F : \mathcal{X}_1 \to \mathcal{X}_2$ be any (deterministic) measurable map. Then

$$\delta(\mathcal{E}_1, \mathcal{E}_2) \leq \sup_{\theta\in\Theta}\|P_{1,\theta} \circ F^{-1} - P_{2,\theta}\|_{\mathrm{TV}}.$$

c. *Let $F : \mathcal{X}_1 \to \mathcal{X}_2$ be a measurable map. Suppose that $P_{1,\theta} \circ F^{-1} = P_{2,\theta}$ for each $\theta \in \Theta$ and suppose that $F(X)$ is a sufficient statistic for $X \sim P_{1,\theta}$. Then $\Delta(\mathcal{E}_1, \mathcal{E}_2) = 0$.*

Recall that for fixed positive noise level $\varepsilon > 0$, in model (3.7) we are given data

$$Y_{p,q} = \langle \tilde{\Lambda}_\gamma[\psi_p], \psi_q \rangle_{L^2(\partial D)} + \varepsilon g_{p,q}, \quad p, q \leq P, \quad g_{p,q} \overset{iid}{\sim} N(0,1),$$

where $\psi_p = c_p \mathbb{1}_{I_p}$ for some disjoint (measurable) sets $I_p \subseteq \partial D$, with c_p such that the ψ_p are $L^2(\partial D)$ orthonormal; for some $r \in \mathbb{R}$, in model (3.8) we are given data

$$Y_{j,k} = \langle \tilde{\Lambda}_\gamma \phi_j^{(r)}, \phi_k^{(0)} \rangle_{L^2(\partial D)} + \varepsilon g_{j,k}, \quad j \leq J, k \leq K, \quad g_{j,k} \overset{iid}{\sim} N(0,1),$$

for a Laplace–Beltrami basis $(\phi_k^{(r)})_{k \in \mathbb{N}}$ of $H^r(\partial D)/\mathbb{C}$, and in model (3.13) we are given data

$$Y = \tilde{\Lambda}_\gamma + \varepsilon \mathbb{W}, \quad \mathbb{W} \text{ a Gaussian white noise indexed by } \mathbb{H}_r,$$

or equivalently data

$$\Big(Y(T) = \langle \tilde{\Lambda}_\gamma, T \rangle_{\mathbb{H}_r} + \varepsilon \sum_{j,k} g_{jk} \langle T \phi_j^{(r)}, \phi_k^{(0)} \rangle_{L^2(\partial D)} : T \in \mathbb{H}_r \Big).$$

Let $\mathcal{E}_0$, $\mathcal{E}_1^{(r)}$ and $\mathcal{E}_2^{(r)}$ denote the experiments corresponding to these data models respectively, in each case taking the parameter space to be $\{\gamma \in \Gamma_{m,D'} : \|\gamma\|_\infty \leq M\}$ for some constants $m, M > 0$ and some domain D' compactly contained in D. We have the following.

Theorem 3.21. *Suppose $\cup_{p \leq P} I_p = \partial D$ and $\operatorname{diam}(I_p) \leq (A/P)^{1/(d-1)}$ for a constant A independent of P. Then the one-way Le Cam deficiency $\delta(\mathcal{E}_0, \mathcal{E}_1^{(0)})$ satisfies*

$$\delta(\mathcal{E}_0, \mathcal{E}_1^{(0)}) \leq C \Big(\max(J,K)^{(5d-2)/(2d-2)} + \varepsilon^{-1} \max(J,K)^{3d/(2d-2)} \Big) P^{-1/(d-1)},$$

for some constant $C = C(A, D', D, M, m)$, and hence vanishes asymptotically if P is large enough compared to ε, J and K.

Remarks. i. The conditions on $(I_p)_{p \leq P}$ are only used to prove that we can approximate any Laplace–Beltrami eigenfunction at a rate $P^{-1/(d-1)}$ with respect to the $L^2(\partial D)$ distance (Lemma 3.22). If $(I_p)_{p \leq P}$ are such that we can approximate Laplace–Beltrami eigenfunctions at a rate $f(P)$ then we achieve the result with $f(P)$ in place of $P^{-1/(d-1)}$.

ii. The given conditions are naturally satisfied by "evenly spaced" sets $(I_p)_{p\leq P}$ partitioning the boundary ∂D, with a constant A depending only on the domain D. This can be seen by considering the covering numbers $N(\partial D, d_{\partial D}, \delta)$ (the smallest number of $d_{\partial D}$ balls of radius δ needed to cover ∂D) for $d_{\partial D}$ the geodesic distance. Theorem 4.5 in Geller & Pensenson [31] applied to the current setting yields

$$N(\partial D, d_{\partial D}, \delta) \leq A\delta^{-(d-1)}$$

for a constant $A = A(D)$, for any $\delta > 0$. Taking $\delta = 2(A/P)^{1/(d-1)}$ we deduce that there exist P balls of radius $\delta/2$ covering D. To construct P disjoint subsets of diameter at most δ, we simply assign each $x \in \partial D$ to exactly one of the balls containing it.

iii. The proof idea can also be used to show a one-way discrepancy result in the other direction. However, this requires that $\min(J, K)$ is large compared to P, hence the two results do not combine to give a (two-way) asymptotic equivalence result.

The idea of the proof is to approximate Laplace–Beltrami eigenfunctions via linear combinations of the indicator functions. The following lemma allows us to control the error in this approximation.

Lemma 3.22. *Under the hypotheses of Theorem 3.21, let ϕ_j^P denote the L^2–orthogonal projection of $\phi_j^{(0)}$ onto* $\operatorname{span}\{\psi_p : p \leq P\}$. *Then there is a constant C depending only on the constant A of Theorem 3.21 and on D such that*

$$\|\phi_j^{(0)} - \phi_j^P\|^2_{L^2(\partial D)} \leq C \max(J, K)^{(2+d)/(d-1)} P^{-2/(d-1)}. \tag{3.52}$$

Proof. Since ϕ_j^P as the L^2–orthogonal projection minimises the L^2 distance to $\phi_j^{(0)}$ of any function in $\operatorname{span}\{\psi_p : p \leq P\}$, for any points $x_p \in I_p$ we see

$$\begin{aligned}
\|\phi_j^{(0)} - \phi_j^P\|^2_{L^2(\partial D)} &\leq \|\phi_j^{(0)} - \sum_{p=1}^{P} \phi_j^{(0)}(x_p)\mathbb{1}_{I_p}\|^2_{L^2(\partial D)} \\
&\leq \max_{p\leq P}(\operatorname{diam}(I_p)^2) \sum_{p=1}^{P} \int_{I_p} \frac{|\phi_j^{(0)}(x) - \phi_j^{(0)}(x_p)|^2}{|x - x_p|^2}\,\mathrm{d}x \\
&\leq (A/P)^{2/(d-1)} \|\phi_j^{(0)}\|^2_{\mathrm{Lip}} \operatorname{Area}(\partial D).
\end{aligned}$$

Using a Sobolev embedding for the compact manifold ∂D, we may estimate the Lipschitz constant of $\phi_j^{(0)}$, denoted $\|\phi_j^{(0)}\|_{\mathrm{Lip}}$ in the above, by a constant times $\|\phi_j^{(0)}\|_{H^\kappa(\partial D)}$ for any

$\kappa > 1 + (d-1)/2$. In particular, taking $\kappa = 1 + d/2$, we see that the final expression is bounded by $C \max(J,K)^{(2+d)/(d-1)} P^{-2/(d-1)}$ for some $C = C(A,D)$ by Corollary 3.14. $\square$

Proof of Theorem 3.21. Let (Y_{pq}) be the data from experiment $\mathcal{E}_0$. Let ϕ_j^P be as in Lemma 3.22, and write $a_{jp} = \langle \phi_j^{(0)}, \psi_p \rangle_{L^2(\partial D)}$, so that $\phi_j^P = \sum_{p=1}^P a_{jp}\psi_p$. Define $F : \mathbb{R}^{P\times P} \to \mathbb{R}^{J\times K}$ via

$$F((u_{pq})_{p,q\le P})_{jk} = \sum_{p,q\le P} a_{jp}a_{kq}u_{pq}.$$

Let $\mathcal{E}_0'$ denote the experiment with data

$$Y_{jk}' = F((Y_{pq})_{p,q\le P})_{jk} = \langle \tilde{\Lambda}_\gamma \phi_j^P, \phi_k^P \rangle_{L^2(\partial D)} + \varepsilon g_{jk}', \tag{3.53}$$

where we define $g_{jk}' = \sum_{p,q} a_{jp}a_{kq}g_{pq}$, and let $\mathcal{E}_1'$ denote the experiment with data (3.53) but for i.i.d. Gaussian noise. Then, by the triangle inequality and Lemma 1.1b,

$$\delta(\mathcal{E}_0, \mathcal{E}_1) \le \delta(\mathcal{E}_0, \mathcal{E}_0') + \delta(\mathcal{E}_0', \mathcal{E}_1) \le \Delta(\mathcal{E}_0', \mathcal{E}_1') + \Delta(\mathcal{E}_1', \mathcal{E}_1).$$

We control each the terms on the right.

$\Delta(\mathcal{E}_0', \mathcal{E}_1')$**:** The covariance of $(g_{jk}')_{jk}$ is given by

$$\mathrm{Cov}(g_{jk}', g_{lm}') = \langle \phi_j^P, \phi_l^P \rangle_{L^2(\partial D)} \langle \phi_k^P, \phi_m^P \rangle_{L^2(\partial D)}.$$

Writing

$$\langle \phi_j^P, \phi_l^P \rangle_{L^2(\partial D)} = \langle \phi_j^{(0)}, \phi_l^{(0)} \rangle_{L^2(\partial D)} + \langle \phi_j^{(0)}, \phi_l^P - \phi_l^{(0)} \rangle_{L^2(\partial D)} + \langle \phi_j^P - \phi_j^{(0)}, \phi_l^P \rangle_{L^2(\partial D)},$$

and applying the Cauchy–Schwarz inequality and Lemma 3.22, noting that $\|\phi_l^P\|_{L^2(\partial D)} \le \|\phi_l^{(0)}\|_{L^2(\partial D)} = 1$, we see for a constant $C = C(A,D)$ that

$$|\langle \phi_j^P, \phi_l^P \rangle_{L^2(\partial D)} - \delta_{jl}| \le C \max(J,K)^{(1+d/2)/(d-1)} P^{-1/(d-1)}.$$

Thus, controlling the Le Cam distance between Gaussian experiments with equal means by $\sqrt{2}$ times the Frobenius distance between the covariance matrices (as in the

proof of Theorem 3.1 in [71]) yields

$$\begin{aligned}\Delta(\mathcal{E}_0', \mathcal{E}_1') &\leq \Big(\sum_{j,l\leq J,}\sum_{k,m\leq K}(\mathrm{Cov}(g_{jk}', g_{lm}') - \delta_{jl}\delta_{km})^2\Big)^{1/2} \\ &\leq CJK\max(J,K)^{(1+d/2)/(d-1)}P^{-1/(d-1)} \\ &\leq C\max(J,K)^{(5d-2)/(2d-2)}P^{-1/(d-1)}.\end{aligned}$$

$\Delta(\mathcal{E}_1', \mathcal{E}_1)$**:** Explicitly calculating the Kullback–Leibler divergence between multivariate normals with the same covariance matrix and using Lemma 1.1a yields

$$\Delta(\mathcal{E}_1', \mathcal{E}_1) \leq \varepsilon^{-1} \times \sup_{\gamma\in\Gamma_{m,D'}:\|\gamma\|_\infty\leq M_0}\Big\|\Big(\langle\tilde{\Lambda}_\gamma\phi_j^{(0)}, \phi_k^{(0)}\rangle_{L^2(\partial D)} - \langle\tilde{\Lambda}_\gamma\phi_j^P, \phi_k^P\rangle_{L^2(\partial D)}\Big)_{j\leq J,k\leq K}\Big\|_{\mathbb{R}^{J\times K}},$$

where the norm on the right is the usual Frobenius or Hilbert–Schmidt norm on the space of $J\times K$ matrices. By Lemma 3.18, $\|\tilde{\Lambda}_\gamma\|_{L^2(\partial D)\to L^2(\partial D)}$ is bounded by a constant $C = C(D, D', M, m)$, hence applying also Lemma 3.22 we have for a different constant $C' = C'(A, D, D', M, m)$,

$$\begin{aligned}&(\langle\tilde{\Lambda}_\gamma\phi_j^{(0)}, \phi_k^{(0)}\rangle_{L^2(\partial D)} - \langle\tilde{\Lambda}_\gamma\phi_j^P, \phi_k^P\rangle_{L^2(\partial D)})^2 \\ &\quad= (\langle\tilde{\Lambda}_\gamma(\phi_j^{(0)} - \phi_j^P), \phi_k^{(0)}\rangle_{L^2(\partial D)} + \langle\tilde{\Lambda}_\gamma\phi_j^P, \phi_k^{(0)} - \phi_k^P\rangle_{L^2(\partial D)})^2 \\ &\quad\leq C'\max(J,K)^{(2+d)/(d-1)}P^{-2/(d-1)}.\end{aligned}$$

Summing over j and k we deduce $\Delta(\mathcal{E}_1', \mathcal{E}_1) \leq C'\varepsilon^{-1}\max(J,K)^{3d/(2d-2)}P^{-1/(d-1)}$.

□

Theorem 3.23. *For any $r\in\mathbb{R}$ and any $\nu > 0$ there is a constant $C = C(\nu, r, D, D_0, M, m)$ such that the Le Cam distance $\Delta(\mathcal{E}_1^{(r)}, \mathcal{E}_2^{(r)})$ satisfies*

$$\Delta(\mathcal{E}_1^{(r)}, \mathcal{E}_2^{(r)}) \leq C\varepsilon^{-1}\min(J,K)^{-\nu}.$$

Proof. We introduce the experiments $\mathcal{E}_i^{(r)}$, $i = 3, 4$ with parameter space $\{\gamma\in\Gamma_{m,D'} : \|\gamma\|_\infty \leq M\}$ corresponding to observations

$$\begin{aligned}&\Big(\pi_{JK}\tilde{\Lambda}_\gamma + \varepsilon\mathbb{W}\Big)(U)_{U\in\mathbb{H}_r}, \quad \text{and} \\ &\Big(\pi_{JK}\tilde{\Lambda}_\gamma + \varepsilon\mathbb{W}\Big)(\pi_{JK}U)_{U\in\mathbb{H}_r},\end{aligned}$$

where we recall the projection π_{JK} was defined in (3.27). By the triangle inequality for the Le Cam distance, we decompose $\Delta(\mathcal{E}_1^{(r)}, \mathcal{E}_2^{(r)}) \leq \Delta(\mathcal{E}_2^{(r)}, \mathcal{E}_3^{(r)}) + \Delta(\mathcal{E}_3^{(r)}, \mathcal{E}_4^{(r)}) + \Delta(\mathcal{E}_4^{(r)}, \mathcal{E}_1^{(r)})$. We control each of the terms on the right.

$\Delta(\mathcal{E}_2^{(r)}, \mathcal{E}_3^{(r)})$**:** Lemmas 1.1a, 3.4 and 3.9 yield

$$\Delta(\mathcal{E}_2^{(r)}, \mathcal{E}_3^{(r)}) \leq \tfrac{1}{2}\varepsilon^{-1} \times \sup_{\gamma \in \Gamma_{m,D'} : \|\gamma\|_\infty \leq M_0} \|\tilde{\Lambda}_\gamma - \pi_{JK}\tilde{\Lambda}_\gamma\|_{\mathbb{H}_r} \leq C\varepsilon^{-1} \min(J,K)^{-\nu},$$

for a constant $C = C(\nu, r, D, D', M, m)$.

$\Delta(\mathcal{E}_3^{(r)}, \mathcal{E}_4^{(r)})$**:** We note that $\big(\pi_{JK}\tilde{\Lambda}_\gamma + \varepsilon\mathbb{W}\big)(\pi_{JK}U)_{U \in \mathbb{H}_r}$ is a sufficient statistic for $\big(\pi_{JK}\tilde{\Lambda}_\gamma + \varepsilon\mathbb{W}\big)(U)_{U \in \mathbb{H}_r}$, so that $\Delta(\mathcal{E}_3^{(r)}, \mathcal{E}_4^{(r)}) = 0$ by Lemma 1.1c.

$\Delta(\mathcal{E}_4^{(r)}, \mathcal{E}_1^{(r)})$**:** Using Lemma 1.1b as in the proof of Theorem 3.21, one shows that the experiment $\mathcal{E}_1^{(r)}$ is equivalent to observing

$$\Big(\textstyle\sum_{j \leq J, k \leq K}\Big(\langle \tilde{\Lambda}_\gamma \phi_j^{(1)}, \phi_k^{(0)} \rangle_{L^2(\partial D)} u_{jk} + \varepsilon g_{jk} u_{jk}\Big)\Big)_{u \in \ell^2}.$$

Since $g'_{jk} := \langle \mathbb{W}, b_{jk}^{(r)} \rangle_{\mathbb{H}_r} \overset{d}{=} g_{jk}$, and, for $U = \sum_{j,k} u_{jk} b_{jk}^{(r)}$,

$$\textstyle\sum_{j \leq J, k \leq K}\Big(\langle \tilde{\Lambda}_\gamma \phi_j^{(1)}, \phi_k^{(0)} \rangle_{L^2(\partial D)} u_{jk}\Big) = \langle \pi_{JK}\tilde{\Lambda}_\gamma, \pi_{JK}U \rangle_{\mathbb{H}_r},$$

we deduce $\Delta(\mathcal{E}_4^{(r)}, \mathcal{E}_1^{(r)}) = 0$ after recalling the concrete formulation (3.14) of model (3.13). $\square$

References

[1] Kweku Abraham. Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data. *Bernoulli*, 25(4A):2696–2728, 11 2019.

[2] Kweku Abraham and Richard Nickl. On statistical Calderón problems, preprint, 2019. arXiv:1906.03486.

[3] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Elsevier, Amsterdam, 2003.

[4] Giovanni Alessandrini. Stable determination of conductivity by boundary measurements. *Applicable Analysis*, 27(1-3):153–172, 1988.

[5] J.P. Aubin. *Applied Functional Analysis*. John Wiley & Sons, Ltd, 2011.

[6] Yannick Baraud. A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli*, 16(4):1064–1085, 2010.

[7] Richard F Bass. *Stochastic Processes*. Cambridge University Press, New York, 1st edition, 2011.

[8] Rabi Bhattacharya, Manfred Denker, and Alok Goswami. Speed of convergence to equilibrium and to normality for diffusions with multiple periodic scales. *Stochastic Processes and their Applications*, 80(1):55–86, 1999.

[9] Rabi N. Bhattacharya and Edward C. Waymire. *Stochastic Processes with Applications*, volume 61 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009. Reprint of the 1990 original.

[10] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[11] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[12] Simon Phillip Blomberg. Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *bioRxiv*, 2017.

[13] Alberto-P. Calderón. On an inverse boundary value problem. In *Seminar on Numerical Analysis and its Applications to Continuum Physics (Rio de Janeiro, 1980)*, pages 65–73. Soc. Brasil. Mat., Rio de Janeiro, 1980.

[14] Pedro Caro and Andoni Garcia. The Calderón problem with corrupted data. *Inverse Problems*, 33(8):085001, 17, 2017.

[15] Ismaël Castillo and Richard Nickl. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.*, 41(4):1999–2028, 2013.

[16] Ismaël Castillo and Richard Nickl. On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Annals of Statistics*, 42(5):1941–1969, 2014.

[17] Isaac Chavel. *Eigenvalues in Riemannian geometry*, volume 115 of *Pure and Applied Mathematics*. Academic Press, Inc., Orlando, FL, 1984.

[18] Jakub Chorowski. *Statistics for diffusion processes with low and high-frequency observations*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2016.

[19] Jakub Chorowski and Mathias Trabs. Spectral estimation for diffusions with random sampling times. *Stochastic Process. Appl.*, 126(10):2976–3008, 2016.

[20] Alberto J. Coca. Adaptive nonparametric estimation for compound Poisson processes robust to the discrete-observation scheme, preprint, 2018. arXiv:1803.09849.

[21] Fabienne Comte, Valentine Genon-Catalot, and Yves Rozenholc. Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli*, 13(2):514–543, 2007.

[22] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: Modifying old algorithms to make them faster. *Statist. Sci.*, 28(3):424–446, 2013.

[23] Arnak Dalalyan. Sharp adaptive estimation of the drift function for ergodic diffusions. *The Annals of Statistics*, 33(6):2507–2528, 2005.

[24] Persi Diaconis. Bayesian numerical analysis. In *Statistical decision theory and related topics, IV, Vol. 1 (West Lafayette, Ind., 1986)*, pages 163–175. Springer, New York, 1988.

[25] Matthew M. Dunlop and Andrew M. Stuart. The Bayesian formulation of EIT: Analysis and algorithms. *Inverse Probl. Imaging*, 10(4):1007–1036, 2016.

[26] Richard Durrett. *Stochastic calculus: A practical introduction*. CRC Press, 1996.

[27] L C Evans. *Partial Differential Equations*. American Mathematical Society, 1998.

[28] Christian L E Franzke, Terence J O'Kane, Judith Berner, Paul D Williams, and Valerio Lucarini. Stochastic climate theory and modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 6(1):63–78, 2014.

[29] C Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Berlin Heidelberg, 2013.

[30] Matthias Gehre and Bangti Jin. Expectation propagation for nonlinear inverse problems – with an application to electrical impedance tomography. *Journal of Comp. Phys.*, 259:513–535, 2014.

[31] Daryl Geller and Isaac Z Pesenson. Band-Limited Localized Parseval Frames and Besov Spaces on Compact Homogeneous Manifolds. *Journal of Geometric Analysis*, 21(2):334–371, apr 2011.

[32] Subhashis Ghosal, Jayanta Ghosh, and Aad van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531, 2000.

[33] Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.

[34] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, Cambridge, 2017.

[35] Ĭ. Ī. Gīhman and A. V. Skorohod. *Stochastic differential equations*. Springer-Verlag, New York-Heidelberg, 1972. Translated from the Russian by Kenneth Wickwire, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72.

[36] David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[37] Evarist Giné and Richard Nickl. Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.*, 39(6):2883–2911, 2011.

[38] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, Cambridge, 2016.

[39] Emmanuel Gobet, Marc Hoffmann, and Markus Reiß. Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics*, 32(5):2223–2253, 2004.

[40] Shota Gugushvili and Peter Spreij. Nonparametric Bayesian drift estimation for multidimensional stochastic differential equations. *Lithuanian Mathematical Journal*, 54(2):127–141, 2014.

[41] Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.*, 24(6):2455–2490, 2014.

[42] Martin Hanke, Nuutti Hyvönen, and Stefanie Reusswig. Convex backscattering support in electric impedance tomography. *Numerische Mathematik*, 117(2):373–396, feb 2011.

[43] Marc Hoffmann. Adaptive estimation in diffusion processes. *Stochastic Processes and their Applications*, 79(1):135–163, 1999.

[44] Victor Isakov. *Inverse problems for partial differential equations*, volume 127 of *Applied Mathematical Sciences*. Springer, Cham, third edition, 2017.

[45] Bangti Jin and Peter Maass. An analysis of electrical impedance tomography with applications to Tikhonov regularization. *ESAIM Control Optim. Calc. Var.*, 18(4):1027–1048, 2012.

[46] Jari P. Kaipio, Ville Kolehmainen, Erkki Somersalo, and Marko Vauhkonen. Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse Problems*, 16(5):1487–1522, 2000.

[47] Alexander Katchalov, Yaroslav Kurylev, and Matti Lassas. *Inverse boundary spectral problems*, volume 123 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. Chapman & Hall/CRC, Boca Raton, FL, 2001.

[48] Mathieu Kessler. Estimation of an Ergodic Diffusion from Discrete Observations. *Scandinavian Journal of Statistics*, 24(2):211–229, 1997.

[49] Ville Kolehmainen, Matti Lassas, Petri Ola, and Samuli Siltanen. Recovering boundary shape and conductivity in electrical impedance tomography. *Inverse Probl. Imaging*, 7(1):217–242, 2013.

[50] Yury A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer London, London, 2004.

[51] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, New York, 01 1986.

[52] Wenbo V Li and Werner Linde. Approximation, Metric Entropy and Small Ball Estimates for Gaussian Measures. *Ann. Probab.*, 27(3):1556–1578, 1999.

[53] J. L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1972.

[54] Robert S. Liptser and Albert N. Shiryaev. *Statistics of Random Processes : I. General Theory*. Springer Berlin Heidelberg, first edition, 1977.

[55] Beatriz Lobo, Cecilia Hermosa, Ana Abella, and Federico Gordo. Electrical impedance tomography. *Annals of Translational Medicine*, 6(2), 2017.

[56] Niculae Mandache. Exponential instability in an inverse problem for the Schrödinger equation. *Inverse Problems*, 17(5):1435–1444, aug 2001.

[57] Ester Mariucci. Le Cam theory on the comparison of statistical models. *Grad. J. Math.*, 1(2):81–91, 2016.

[58] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993.

[59] François Monard, Richard Nickl, and Gabriel P. Paternain. Consistent Inversion of Noisy Non-Abelian X-Ray Transforms, preprint, 2019. arXiv:1905.00860.

[60] Adrian I. Nachman. Reconstructions from boundary measurements. *Ann. of Math. (2)*, 128(3):531–576, 1988.

[61] Richard Nickl. Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *Journal of the European Mathematical Society*, to appear, 2019.

[62] Richard Nickl and Jakob Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.*, 45(4):1664–1693, 2017.

[63] Richard Nickl and Jakob Söhl. Bernstein–von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat.*, 13(2):3513–3571, 2019.

[64] Richard Nickl, Sara van de Geer, and Sven Wang. Convergence rates for penalised least squares estimators in PDE-constrained regression problems, preprint, 2018. arXiv:1809.08818.

[65] Omiros Papaspiliopoulos, Yvo Pokern, Gareth O. Roberts, and Andrew M. Stuart. Nonparametric estimation of diffusions: A differential equations approach. *Biometrika*, 99(3):511–531, 2012.

[66] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:32 pp., 2015.

[67] Henri Poincaré. *Calcul des probabilités*. Gautier-Villars, Paris, 1912.

[68] Y Pokern, A M Stuart, and J H van Zanten. Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stochastic Processes and their Applications*, 123(2):603–628, 2013.

[69] David Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2001.

[70] Kolyan Ray. Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7(1):2516–2549, 2013.

[71] Markus Reiß. Asymptotic Equivalence for Nonparametric Regression with Multivariate and Random Design. *The Annals of Statistics*, 36(4):1957–1982, 2008.

[72] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 1*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Foundations, Reprint of the second (1994) edition.

[73] L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. Itô calculus, Reprint of the second (1994) edition.

[74] Lassi Roininen, Janne M. J. Huttunen, and Sari Lasanen. Whittle-Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography. *Inverse Probl. Imaging*, 8(2):561–586, 2014.

[75] Mikko Salo. Lecture notes on the Calderón problem, Spring 2008. http://users.jyu.fi/~salomi/lecturenotes/calderon_lectures.pdf.

[76] Jakob Söhl and Mathias Trabs. Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift. *ESAIM: PS*, 20:432–462, 2016.

[77] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.

[78] John Sylvester and Gunther Uhlmann. A global uniqueness theorem for an inverse boundary value problem. *Ann. of Math. (2)*, 125(1):153–169, 1987.

[79] Hans Triebel. *Theory of function spaces*. Monographs in mathematics. Birkhäuser Verlag, 1983.

[80] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer, New York, 2009.

[81] G. Uhlmann. Electrical impedance tomography and Calderón's problem. *Inverse Problems*, 25(12):123011, 39, 2009.

[82] Frank van der Meulen, Moritz Schauer, and Jan van Waaij. Adaptive nonparametric drift estimation for diffusion processes using Faber–Schauder expansions. *Statistical Inference for Stochastic Processes*, 2017.

[83] Frank van der Meulen and Harry van Zanten. Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli*, 19(1):44–63, 2013.

[84] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[85] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[86] Jan van Waaij and Harry van Zanten. Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electron. J. Statist.*, 10(1):628–645, 2016.

[87] Harry van Zanten. Nonparametric Bayesian methods for one-dimensional diffusion models. *Mathematical Biosciences*, 243(2):215–222, 2013.

[88] Tao Zhu, Rui Feng, Jin-qi Hao, Jian-guo Zhou, Hua-lin Wang, and Shuo-qin Wang. The Application of Electrical Resistivity Tomography to Detecting a Buried Fault: A Case Study. *Journal of Environmental and Engineering Geophysics*, 14(3):145–151, 2009.