Theory and Practice of Bayesian Methods in Inverse Problems and Related Nonparametric Models

by

Youngsoo Baek

Department of Statistical Science
Duke University

Defense Date: June 13, 2024

Approved:

Sayan Mukherjee, Supervisor

Surya Tokdar

Samuel I. Berchuck

Wilkins Aquino

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2024

ABSTRACT

Theory and Practice of Bayesian Methods in Inverse Problems and Related Nonparametric Models

by

Youngsoo Baek

Department of Statistical Science
Duke University

Defense Date: June 13, 2024

Approved:

Sayan Mukherjee, Supervisor

Surya Tokdar

Samuel I. Berchuck

Wilkins Aquino

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2024

Abstract

This dissertation collects results on novel applications of Bayesian inference and computation to areas using high-to-infinite-dimensional mathematical models. The two areas covered here are inverse problems based on mathematical modeling through partial differential equations and probabilistic machine learning through deep neural networks. Chapter 2 presents fundamental results on the frequentist coverage of Bayes posteriors in nonlinear inverse problems based on PDE models. Chapters 3 presents theoretical and methodological results on using generalized Bayes posteriors in these inverse problems under model misspecification. Chapter 4 presents results on the generalizability of posterior inference on new examples in a 2-layer neural network setting. The main novelty of the theoretical results is to characterize frequentist coverage of Bayes posterior high probability sets corresponding to reasonable Gaussian process priors in each problem. Methodologically, this work aims to contribute inferential and computational tools for using Bayesian inference in inverse problems with mathematical models that are potentially misspecified.

Contents

List of Tables

List of Figures

Acknowledgements

I thank my advisor, Sayan Mukherjee, who has been a formative influence to the development of my academic pursuit over the course of last five years. I thank the University of Leipzig and Max Planck Institute of Mathematical Sciences for their funding and hospitality during my stay as a guest researcher in Leipzig over the summer of 2023. I thank the Department of Statistical Science and the Duke Graduate School for their generosity in funding my travels towards the end of my dissertation phase. My collaborators Sam Berchuck, Wilkins Aquino and Katerina Papagiannouli deserve all due credit for their helpful comments and feedback while I was working on the materials included in this thesis. Tuhin Roy and Murthy Guddati deserve a special mention for their assistance as I was carrying out the simulated experiment included in Chapter 3. I am also indebted to Simon Mak, Galen Reeves, Matthew M. Engelhard, Boyao Li, David Page and Alexander J. Thomson, who all have provided stimulating conversations as I was preparing the materials contained in Chapters 3 and 4.

I thank my colleagues Joe Mathews, Raphaël Morsomme, Phuc Nguyen, Frances Hung, Federica Stolz and countless others who offered invaluable friendship through some hard times. Finally, I thank my parents, my brother and Yelim for their love and support throughout the entirety of my stay in graduate school.

1. Introduction

This work collects new progress on the theory of Bayesian inference in inverse problems along with materials previously published by the author (Baek, Aquino, and Mukherjee, 2023; Baek, Berchuck, and Mukherjee, 2024) covering inverse problems and high-dimensional machine learning. The majority of this work (Chapters 2 and 3) concerns itself with the theory and practice of *inverse problems*, a broad class of problems in mathematical modeling which violate in some form the well-posedness properties defined by Hadamard, 1902. That implies, roughly, that a mathematical formulation of a certain scientific problem may lack meaningful notion of a solution, that such a solution may not be uniquely defined, or that a solution may not depend continuously on the inputs to the model. Modern theory of postulating solutions to such ill-posed problems through appropriate use of regularization dates back at least to the foundational work of Tikhonov and Arsenin, 1977. A standard reference in this field is Heinz W Engl, Kunisch, and Neubauer, 1989.

Over the last decade or so, more attention was paid to the *statistical* nature of inverse problems arising in practice. Such statistical properties include more than just the fact that the "noise" corrupting the finite number of observations made can be modelled stochastically. Quantification of uncertainty in the context of inverse problems is increasingly demanded by many applications (A M Stuart, 2010), a demand to which Bayesian statistics provides a useful viewpoint (Cotter, Dashti, et al., 2009). In a Bayesian framework, one prescribes a prior distribution summarizing relative uncertainties about possible solutions to the inverse problem. After observing noisy data, one updates the probabilities to obtain a posterior distribution of the possible solutions.

Let us mathematize a typical structure of inverse problems one faces in scientific applications before any further discussion. The data comprise N points $(Y_i, X_i)_{i=1}^N$ related to each other in a regression model:

$$Y_i = G(\theta)(X_i) + \varepsilon_i,\ i = 1, 2, \ldots, N.$$

The parameter $\theta \in \Theta$ is of interest and often describes the underlying physical properties

of the system. The "forward map" G defined on Θ connects this parameter to the observed field. We will mainly consider those maps arising as nonlinear implicit functions that map a parameter θ of a partial differential equation (PDE) to the solution $G(\theta)$ of this PDE. The "covariates" X_is correspond to the design points at which the field $G(\theta)$ is evaluated, while the "response" variate Y_is are corrupted with some stochastic noise ε_i. At this point, the model is fairly general and various assumptions of increasing complexity can be added to give more structure to the estimation of θ. A major challenge is the ill-posed nature of the problem: the inverse of G, if it does exist, is often not continuous, so that the estimate of θ is not stable in slight changes to the observed data (Y_i, X_i)s. This highlights the need for appropriate regularization enforcing stability, which goal plays quite nicely with encoding prior knowledge through Bayesian inference. The regression notation formulation above will be heavily used in Chapter 2, but somewhat revised in Chapter 3 to accommodate somewhat different perspectives on the same class of problems.

Chapter 2 concerns itself with some foundational questions about the quality of our estimate. We will assume there exists some $\theta_0 \in \Theta$ such that the observed data are evaluations of $G(\theta_0)$ at distinct points, up to an additive noise that does not depend on θ_0. Further, this additive noise component is assumed to be independent with each other and has a simple, identical probability distribution. The asymptotic properties of the estimate we obtain through the Bayes posterior can then be analyzed using the language of non-parametric Bayesian statistics. Such properties would include *consistency* (do we recover θ_0 with increasing N?) and the *rate of convergence* (how fast do we recover θ_0?). Both of these points have been intensively studied by previous authors, some of which we cite below. The Bayes formulation, however, provides a much richer structure to the derived estimate, as it does not offer a single most likely "point estimator" for θ, but rather a measure of how likely each possible solution in the permitted parameter space is. Statistically, the desirable property such a probability measure must possess is its *frequentist coverage*: Will the high probability region under my posterior on Θ contain the unknown θ_0 most of the times, across hypothetical replication of the same experiment? Chapter 2 discusses the novel findings in

this line of work. Following the pioneering study of Monard, Nickl, and Paternain, 2021 in this regard, we study in depth the asymptotic coverage of Bayes posteriors when estimating linear functionals of θ and the prior is induced by a Gaussian process. Our results are general enough to cover the cases when the so-called Bernstein von-Mises theorem does not hold and clearly characterize the class of functionals for which the theorem does hold.

Chapter 3 complicates the foregoing account by introducing the notion of *model misspecification*. Its major concern is an applied one, namely, that in real life one is never fully assured of the agreement between the posited scientific model and the actual procedure by which the observed data are generated. For instance, the practicability of Bayesian methods to concrete problems relies on the ability to both evaluate the density of, and easily sample from, the *likelihood* of the data $(Y_i, X_i)_{i=1}^N$ indexed by differing parameters θ. However, it is difficult to specify the data-generating process in nonlinear inverse problems due to two main sources of model uncertainty: forward model uncertainty, with respect to the underlying system dynamics; uncertainty, or lack of knowledge, with respect to the distribution of noise. The unrealistic nature of guessing the "true likelihood," even in a broad class of models, raises a serious concern about using Bayesian methods. We thus consider an alternative application of *Gibbs posterior* or *generalized Bayes* framework that has already been proposed by various authors for different problems. The framework similarly requires a prior distribution and outputs a probability update conditional on the data. Gibbs posteriors do not rely on the knowledge of likelihood. They are derived as a solution to a variational problem on the space of probability measures on the space of solutions. They require the choice of a loss function that measures the mismatch between the model and the data. The practicability of such generalized Bayes methods in inverse problems relies not only on establishing various desirable properties already satisfied by Bayes posteriors, such as stability and some notion of consistency, but also on the need to choose appropriately an additional regularization parameter that is absent from the Bayesian formulation. Chapter 3 covers a gamut of results in both respects, presenting some theoretical results along with a new computational approach to choosing the regularization parameter based on a cross-validation

scheme implemented through sequential Monte Carlo methods.

Chapter 4 is somewhat removed from the rest of the work in that it does not explicitly address the same class of problems as covered by Chapters 2 and 3. Instead, it analyzes the so-called "random features (RF) models" that have garnered some attention in the deep learning community through the lens of Bayesian methods, due to their ability to mimic some surprising phenomena exhibited by training deep neural networks in practical applications. The surprise here is the generalizability of overparameterized models that can perfectly interpolate the data, seemingly refuting the classical notions of "bias-variance trade-offs" on unseen data. This Chapter, covering previously published material, does not attempt to bridge it in any way to the study of inverse problems in previous Chapters. Nevertheless, some conceptual connections merit a discussion here, as they are not all so tenuous. First, machine learning methods can assist scientific modeling by enhancing its scope of application. "Deep" models like neural networks can presumably approximate highly complicated functions that lie outside reproducing kernel Hilbert spaces (Parhi and Nowak, 2022) and thus hold a promise for modeling very complex dynamics that is not covered by classical smoothness scales or kernel-type statistical learning methods. Even when one has access to a physical model that describes the data generating mechanism to a high level of accuracy, machine learning methods can still be of use if they can be trained faster with reasonable approximation than this physical model. Second, technical findings in machine learning and inverse problems can illuminate each other when unified under the lens of (Bayesian) nonparametric statistics. This is not so surprising, given both areas are concerned with learning high-to-infinite-dimensional parameters. For instance, recent literature testifies the concept of regularization is implicitly embedded in machine learning tasks that do not appear to be adding any penalty to the objective functional minimized. The "double descent" results studied by Mei and Montanari, 2022 and also mentioned in Chapter 4 may be viewed as one proof. Furthermore, the theoretical results in Chapter 4 will highlight some surprising similarities between the behaviors of Bayes posterior credible regions when training RF models and a simpler problem studied by Freedman, 1999, which in itself may be viewed as a vastly

simplified inverse problem.

1.1 Previous Works

The relation between the regularized least-squares problem proposed by Tikhonov and Arsenin, 1977 and the maximum a posteriori (MAP) estimation problem in Bayesian statistics has been known for some time. Bayesian methods for inverse problems have been successfully adopted in diverse domains, nicely summarized by Kaipio and Somersalo, 2005. Recent literature (Cotter, Dashti, et al., 2009; A M Stuart, 2010; Cotter, Roberts, et al., 2013) has extended the Bayesian framework with Gaussian likelihood to infinite-dimensional settings. Consistency and contraction rate of Bayes posteriors have been analyzed by Vollmer, 2013 and Giordano and Nickl, 2020. The frequentist coverage of confidence sets and the related property of Fisher efficiency are classical concepts in asymptotic statistics, although nonparametric estimation problems cause significant technical challenges to applying these concepts. Pioneering studies in a Bayesian context include Cox, 1993; Freedman, 1999; Johnstone, 2010. A more recent literature, including the pioneering studies of Castillo and Nickl, 2013; Castillo and Rousseau, 2015; Nickl, 2020,has applied the analysis to PDE-based inverse problems and derived Bernstein von-Mises theorem for specific inverse problems.

The Gibbs posterior framework (P G Bissiri, C C Holmes, and S G Walker, 2016; Jiang and Tanner, 2008; Martin, Mess, and S G Walker, 2017) is not new, and its application in inverse problems was studied by Zou et al., 2019; Dunlop and Yang, 2021. Similar concepts have been studied by P. D. Grünwald and Langford, 2007; P. D. Grünwald and Ommen, 2017; Miller and Dunson, 2019; Bhattacharya, Pati, and Yang, 2019, among others, for improving the robustness of Bayesian inference under model misspecification. The novel model selection theory we develop in this paper can be viewed as an analog of the theory of Bayesian model selection and Bayesian cross-validation under model misspecification (Bernardo and Smith, 2009). Computationally, we rely on sequential Monte Carlo and particle filters algorithms. These algorithms have gained recent attention for potential use in Bayesian inverse problems. (Kantas, Beskos, and Jasra, 2014; Beskos et al., 2015) have used particle filters to solve

parabolic and elliptic inverse problems. Zou et al., 2019 have proposed a combination of particle filter and reduced order models for improved computational efficiency.

For RF models, the "double descent" curve, referring to the test error first increasing then decreasing with model complexity, has been both empirically and theoretically validated for linear (Trevor Hastie et al., 2022) and nonlinear models (Ghorbani et al., 2021; Mei and Montanari, 2022; Hu and Lu, 2022). Mei and Montanari, 2022 showed that the generalization error of the random features (RF) model proposed by Rahimi and Recht, 2007 does demonstrate double descent. Perhaps more surprisingly, they also showed that vanishingly small regularization can yield optimal generalization in a nearly noiseless learning task. These findings highlight the recent trend of explaining the success of machine learning through the prism of beneficial interpolation and overparameterization (Belkin, 2021).

There exists, however, another approach to overparameterized learning problems, which is the Bayesian posterior predictive distribution. In Bayesian practice, one can define the training objective as the negative log-likelihood of a probabilistic model. The ridge regularized predictor for the RF model studied by Mei and Montanari, 2022 is the maximum a posteriori (MAP) estimator of this probabilistic model. A "truly Bayesian" approach, on the other hand, is to derive the posterior predictive distribution which averages over different predictors and summarizes one's uncertainty in prediction. The posterior predictive distribution is also referred to as the "Bayesian model average" in the literature (Ovadia et al., 2019; Fortuin et al., 2021). A fundamental question in Bayesian statistics is whether Bayesian credible sets, defined as high probability regions of the posterior, are also valid confidence sets in the frequentist sense (Kleijn and A. v. d. Vaart, 2012).

2. Frequentist Coverage of Bayes Posteriors in Inverse Problems

In this Chapter, we show some fundamental results on the statistical theory of uncertainty quantification for Bayes posteriors when used in PDE-based inverse problems. The novel results focus on the frequentist coverage of certain credible regions, which in themselves presume that the posterior distribution asymptotically charges a large amount of probability mass in small neighborhoods of some θ_0 indexing the unknown data likelihood. We follow Monard, Nickl, and Paternain, 2021 and develop some new technical tools that allow us to cover the cases where Bernstein von-Mises theorem established by the previous authors fail to hold. The take-away is that Bayes posteriors can be at least conservative when certain smoothness conditions are met, both for the prior and the unknown parameter, but its exact asymptotic coverage may not agree with the nominal one. We then apply the general theorems to the concrete case of estimating diffusitivity field in a divergence form elliptic equation model. Along the way, some novel results are also derived on the rate of contraction of Bayes posteriors and the improvement of existing Bernstein von-Mises theorems for PDE-based inverse problems.

2.1 Problem Setup

2.1.1 Notation

For two sequences a_N and b_N, we write $a_N = O(b_N)$ if there exists a constant C such that $|a_N| \leq Cb_N$ for all sufficiently large N. We often use the shorthand $a_N \lesssim b_N$. We say $a_N \asymp b_N$ when $a_N \lesssim b_N$ and $a_N \gtrsim b_N$. $a_N = o(b_N)$ when for sufficiently large N, $|a_N| \leq \varepsilon b_N$ for any constant $\varepsilon > 0$. We often use the shorthand $a_n \ll b_N$.

Let $(\Omega, \mathscr{S}, P)$ be a measure space. We write by $L^p_\lambda(\Omega)$ the space of P-integrable real-valued maps on S. Let $X : \Omega \to \mathbb{R}$ be a random variable when P is a probability measure and $\mathbb{R}$ is equipped with the standard Borel σ-algebra. We use the shorthand $\mathscr{L}(X)$ to denote the probability law generated by X. For two sequences of random variables $X_N : \Omega \to \mathbb{R}$ and $Y_N : \Omega \to \mathbb{R}$, we write $X_N = O_P(Y_N)$ if X_N/Y_N is a stochastically bounded sequence, and $X_N = o_P(Y_N)$ if X_N/Y_N converges in probability to zero (Resnick, 2019). Convergence of a

sequence X_N to zero in probability is also written $X_N \xrightarrow{P} 0$.

The topological dual space of a normed linear space X is denoted X^*. A ball in X centered at zero with radius r is denoted $B_X(r)$. A ball centered at a nonzero vector x_0 is denoted $B_X(r;x_0)$. For two sets A and B in a linear space, $A+B$ denotes their Minkowski sum $\{a+b : a \in A,\, b \in B\}$. When B is a singleton $\{b\}$, we simply write $A+b$.

Let Ω be a open bounded connected subset of $\mathbb{R}^d$. $C^r(\Omega)$ for r positive is the standard Hölder space of $\lfloor r \rfloor$-times continuously differentiable functions whose partial derivatives of order $\lfloor r \rfloor$ are Hölder continuous with an exponent $r - \lfloor r \rfloor$. $H^m(\Omega)$ for m a positive integer order is the standard Sobolev space of $\mathbb{R}$-valued functions on Ω for which all m-th order weak derivatives exist and belong to $L^2(\Omega)$. Non-integer order spaces are defined through interpolation (Lions and Magenes, 2012). $C_0^\infty(\Omega)$ is the space of all infinitely differentiable functions with compact support contained in $\mathscr{X}$. $H_0^s(\Omega)$, $s \geq 0$ is the closure of $C_0^\infty(\Omega)$ in $H^s(\Omega)$, being equal to $H^s(\Omega)$ only if $s \leq \frac{1}{2}$. $H^{-s}(\Omega)$, $s > 0$ is the dual space $(H_0^s(\Omega))^*$. We omit the indication of the domain Ω for the Sobolev scales and their norms when it is made clear by the context.

$\mathscr{N}(S,d,\varepsilon)$ indicates the log-covering (entropy) number of a set S in a metric space $(\mathscr{F},d)$ using metric balls of radius ε.

2.1.2 Observation Model

Let $\mathscr{Z}, \mathscr{X} \subset \mathbb{R}^d$ for a fixed d be open bounded domains with smooth boundaries. We begin by assuming the observed data arise from a model

$$Y_i = G(\theta)(X_i) + \varepsilon_i, \quad i = 1, \ldots, N. \tag{2.1}$$

The unknown parameter of interest in an inverse problem is θ that belongs to the parameter space Θ, which is assumed to be a separable linear subspace of $L^\infty(\mathscr{Z}) \subset L^2_\zeta(\mathscr{Z})$, ζ being the Lebesgue measure restricted to $\mathscr{Z}$. A canonical "observation space" is $L^2_\lambda(\mathscr{X})$, λ being the uniform probability measure on $\mathscr{X}$. X_is are "design points" taking values in $\mathscr{X}$. $G : \Theta \to L^2(\mathscr{X})$ is an operator that encodes the mathematical model. In this work, we will focus on nonlinear operators that arise as "forward models" defined by a mapping from a coefficient

parameter of an elliptic PDE to the solution of a PDE. Throughout, G is assumed to be injective. We also simply write $L^2_\zeta(\mathscr{Z}) \equiv L^2(\mathscr{Z})$ and $L^2_\lambda(\mathscr{X}) \equiv L^2(\mathscr{X})$ with the understanding that ζ, λ are fixed.

We assume the use of a random design in which X_is are uniformly distributed over a bounded subset $\mathscr{X}$ with smooth boundary $\partial\mathscr{X}$. The observation errors ε_is are assumed to be i.i.d. standard normal random variables. Together, a pair of i.i.d. random variables (Y_i, X_i) has a probability distribution on $\mathbb{R} \times \mathscr{X}$ which we denote as P_θ, parameterized by $\theta \in \Theta$. The goal is to infer some fixed, unknown "inverse solution" θ_0, having either sufficient Hölder or Sobolev smoothness, based on i.i.d. noisy observations.

2.1.3 Gaussian Prior and Smoothness Scales

This Section briefly summarizes the important properties of a Gaussian prior used in this work. To do so, we first define an operator equivalent to a lower power of the covariance operator, which in turn defines an appropriate smoothness scale with respect to which regularity conditions of the inverse problem will be stated. This perspective, previously adopted by Vollmer, 2013, maintains close connections to the extant literature on regularization in Hilbert scales and provides a unifying language for discussing the theoretical properties of most commonly used Gaussian priors.

Let Λ^{-1} be an injective, strictly positive, self-adjoint, compact operator on $L^2(\mathscr{Z})$. It has a discrete spectral decomposition that can be used to define powers Λ^{-p}, $p \in (0, \infty)$. Throughout this work, we assume Λ^{-p} is trace-class for any power p greater than $\frac{d}{2}$. A primary example of such an operator is the compact inverse of a Dirichlet Laplacian for a bounded domain with smooth boundary. Another example is obtained by considering sample paths of a Gaussian process on $\mathbb{R}^d$ with a Whittle-Matérn covariance kernel, possessing sufficient regularity and restricted to Ω, as defining a Gaussian measure on $L^2(\mathscr{Z})$.

The inverse Λ of Σ is an unbounded operator, densely defined on the domain

$$D(\Lambda) := \left\{ x \in L^2(\mathscr{Z}) : \sum_j \lambda_j \langle x, e_j \rangle^2_{L^2(\mathscr{Z})} \right\} = R(\Lambda^{-1}),$$

for positive real eigenvalues $(\lambda_j)_{j=1}^{\infty}$ and eigenvectors $(e_j)_{j=1}^{\infty}$ forming a complete orthonormal basis of $L^2(\mathscr{Z})$. Furthermore, there exists a constant $c > 0$ satisfying for every $x \in D(\Lambda)$

$$\|\Lambda x\|_{L^2(\mathscr{Z})} \geq c\|x\|_{L^2(\mathscr{Z})}. \tag{2.2}$$

Based on the results from Section 8.4 (Heinz Werner Engl, Hanke, and Neubauer, 1996), we can deduce that

$$\widetilde{H}^{\infty} := \bigcap_{p\in\mathbb{R}} D(\Lambda^p)$$

is dense in $L^2(\mathscr{Z})$. We define the associated Hilbert scales $(\widetilde{H}^p)_{p\in\mathbb{R}}$ as the completion of $\widetilde{H}^{\infty}$ with respect to the norm $\|\cdot\|_{\widetilde{H}^p}$ induced by the inner product

$$\langle f, g\rangle_{\widetilde{H}^p} := \langle \Lambda^{p/2} f, \Lambda^{p/2} g\rangle_{L^2(\mathscr{Z})}. \tag{2.3}$$

The following result for Hilbert scales can be found in Proposition 8.19 (Heinz Werner Engl, Hanke, and Neubauer, 1996).

Theorem 2.1.1. *The Hilbert scales $(\widetilde{H}^p)_{p\in\mathbb{R}}$ induced by the operator Λ has the following properties.*

(i) For $-\infty < p < q < \infty$, $\widetilde{H}^q$ is dense and continuously embedded in $\widetilde{H}^p$.

(ii) Let $p, q \in \mathbb{R}$. The operator Λ^{p-q} defined on $L^2(\mathscr{Z})$ has a unique extension to $\widetilde{H}^p$ that is an isomorphism from $\widetilde{H}^p$ onto $\widetilde{H}^q$. If $p > q$, then the extension restricted to $\widetilde{H}^p$ in $\widetilde{H}^q$ is self-adjoint and strictly positive. For appropriate extensions, we have $\Lambda^{p-q} = \Lambda^p\Lambda^{-q}$ and $(\Lambda^{-q})^{-1} = \Lambda^{-q}$.

(iii) If $p \geq 0$, then $\widetilde{H}^p = D(\Lambda^p)$ and $\widetilde{H}^{-p} = (\widetilde{H}^p)^*$.

(iv) If $-\infty < p < q < r < \infty$, then we have the interpolation inequality

$$\|f\|_{\widetilde{H}^q} \leq \|f\|_{\widetilde{H}^p}^{\frac{r-q}{r-p}} \|f\|_{\widetilde{H}^r}^{\frac{q-p}{r-p}}.$$

Remark. The fact that there exists some power Λ^{-p} that is trace-class or has a discrete spectrum is not essential for the theory of Hilbert scales. For us, such power is important as it should correspond to a valid covariance operator of a Radon Gaussian measure defined on a separable Hilbert space.

We now define a Gaussian prior for Bayesian inference and connect it to the Hilbert scales $(\widetilde{H}^s)$. Our primary reference here is Section 2 of Bogachev, 1998. Similar, more succinct accounts can be found in Section 6 (Andrew M Stuart, 2010).

Let Π_0 be a Gaussian measure defined on a separable Hilbert space $L^2(\mathscr{Z})$ per Definition 2.2.1 (Bogachev, 1998). Separability implies this measure is a Radon, in particular Borel, measure, defining an $L^2(\mathscr{Z})$-valued random element θ in the sense of Definition 11.2 (Subhashis Ghosal and Aad W van der Vaart, 2017). Furthermore, since $L^2(\mathscr{Z})$ is a Hilbert space, by Riesz isomorphism it can be uniquely characterized by the Fourier transform

$$\vartheta \mapsto \exp\left(i\langle \vartheta, m\rangle_{L^2} - \frac{1}{2}\langle K\vartheta, \vartheta\rangle_{L^2}\right),$$

where m is the mean vector in $L^2(\mathscr{Z})$ and $K : L^2(\mathscr{Z}) \to L^2(\mathscr{Z})$ is the symmetric, non-negative, trace-class covariance operator. We choose $m \equiv 0$ and $K = \Lambda^{-\alpha}$ for Λ above and $\alpha > p'$. The Cameron-Martin space of Π_0 for a centered Gaussian measure on $L^2(\mathscr{Z})$ has a simple characterization of being $\widetilde{H}^\alpha \equiv \Lambda^{-\alpha/2}(L^2(\mathscr{Z}))$, which by Theorem 2.1.1 above is $\equiv \Lambda^{-\alpha}(\widetilde{H}^{-\alpha}(\mathscr{Z}))$. The Cameron-Martin space is in itself is a separable Hilbert space equipped with the inner product (2.3) for $p = \alpha$. The Cameron-Martin theorem (Proposition 2.4.2, Bogachev, 1998) asserts that whenever $h \in \widetilde{H}^\alpha$, the pushforward $T_{h\#}\Pi_0$ of Π_0 under the shift map $T_h : x \mapsto x + h$ is equivalent to Π_0, the Radon-Nikodym derivative given as

$$\frac{d(T_{h\#}\Pi_0)}{d\Pi_0}(\vartheta) = \exp\left(\langle H, \vartheta\rangle_{L^2(\mathscr{Z})} - \frac{1}{2}\|h\|_{\widetilde{H}^\alpha}\right). \tag{2.4}$$

Here, $H := \Lambda^\alpha h \in \widetilde{H}^{-\alpha}$ and the above density is indeed well-defined, as it is integrable with respect to the measure Π_0.

The topological support of Π_0 constructed above is known to be the closure of its Cameron-Martin space in $L^2(\mathscr{X})$, which, by the density result of Theorem 2.1.1, is just $L^2(\mathscr{X})$ itself. A smaller Hilbert space on which Π_0 has full measure can be characterized, using the fact that the natural embedding from $\widetilde{H}^\alpha$ to $\widetilde{H}^s$ is Hilbert-Schmidt whenever $s < \alpha - \frac{d}{2}$. Such $\widetilde{H}^s$ then has full measure under Π_0, as can be seen through the proof of

Theorem 3.9.6. (Bogachev, 1998) or Theorem 3.20 (Neerven et al., 2010). When the support of Π_0 is embedded in the usual Sobolev space, the interpolation inequality of Theorem 2.1.1 can be used to prove sufficient regularity of the random elements drawn from the posterior distribution. Hence the following assumption, which will be seen to be met by the examples we consider in Section 2.3.

Assumption 1. *$\widetilde{H}^p(\mathscr{L})$ continuously embeds into $H^p(\mathscr{L})$ with $\widetilde{H}^p(\mathscr{L}) \cap \Theta \neq \varnothing$ for $p > \frac{d}{2}$.*

Remark. We refer the reader to the results Sections 2.3-4 (Bogachev, 1998) for a detailed discussion of the equivalence between our above characterizations and the usual definitions of the Cameron-Martin space. We also note that the Cameron-Martin space is often referred to as the "reproducing kernel Hilbert space (RKHS)" in the mathematical statistics literature (Subhashis Ghosal and Aad W van der Vaart, 2017; Giné and Nickl, 2021). We employ a different terminology to maintain consistency with Bogachev, 1998, wherein the term RKHS refers to $\widetilde{H}^{-\alpha}$ in our setting, rather than $\widetilde{H}^{\alpha}$.

For the purpose of the results in this work, we will have to consider priors that are dependent on N, the growing number of observations. We consider a family of priors obtained by "rescaling" the fixed prior Π_0, which has its precedent in the results of A. v. d. Vaart and J. v. Zanten, 2007 and studied in the context of inverse problems by both Giordano and Nickl, 2020 and Monard, Nickl, and Paternain, 2021.

Assumption 2. *Let Π_N be a Gaussian measure defined by the relation*

$$\Pi_N(A) = \Pi_0(\tau_N \theta \in A) \tag{2.5}$$

for every Borel subset $A \subset \Theta$ with a fixed sequence (τ_N) satisfying $N\tau_N^2 \to \infty$ and $\tau_N \lesssim N^k$ for some $k > 0$.

Combining the observation model (2.1) with the prior constructed as under Assumption 2, we can write the "posterior distribution" corresponding to an N-dependent prior Π_N as

$$\Pi_N(\theta \in A | X_1, \ldots, X_N, Y_1, \ldots, Y_N) := \frac{\int_A e^{-l_N(\theta)} \, d\Pi_N(\theta)}{\int_\Theta e^{-l_N(\theta)} \, d\Pi_N(\theta)}, \tag{2.6}$$

for every A Borel measurable, with l_N defining the "empirical loss" term

$$l_N(\theta) = \sum_{i=1}^{N} l_i(\theta),\ l_i(\theta) = \frac{1}{2}|Y_i - G(\theta)(X_i)|^2.$$

Henceforth we will abbreviate the notation for the posterior by writing
$\mathscr{D}^N = (Y_1, \ldots, Y_N, X_1, \ldots, X_N)$ and $\Pi_N(A|Y_1, \ldots, Y_N, X_1, \ldots, X_N) = \Pi_N(A|\mathscr{D}^N)$.

2.2 Main Results

In this section, we study the frequentist coverage properties of a credible interval induced by the posterior distribution of a linear functional of θ. We limit ourselves to continuous linear functionals on $L^2(\mathscr{Z})$, each corresponding to a Riesz representer ψ:

$$\Psi : \theta \mapsto \langle \theta, \psi \rangle_{L^2(\mathscr{Z})},\ \psi \in L^2(\mathscr{Z}).$$

We will consider credible intervals on the real line induced by the first and second posterior moments for $\Psi(\theta)$:

$$I_N = \{x \in \mathbb{R} : |x - \Psi(\overline{\theta}_N)| \leq z_{1-\gamma/2}\widehat{s}_N\}, \tag{2.7}$$

where $\overline{\theta}_N$ is the posterior mean $\mathbb{E}[\theta|\mathscr{D}^N]$, defined as a Bochner integral against measure (2.6), and $\widehat{s}_N^2$ is the posterior variance, defined as the second moment of a random variable $\Psi(\theta - \overline{\theta}_N)$ whose law is induced by the posterior distribution $\Pi_N(\cdot|\mathscr{D}^N)$. γ denotes the desired level of coverage with a corresponding standard normal quantile $z_{1-\gamma/2}$.

2.2.1 Structural Assumptions

To discuss the results of this section, we follow Monard, Nickl, and Paternain, 2021 and introduce several conditions needed to guarantee the posterior is "nearly Gaussian" around a sufficiently small region around θ_0. Some assumptions are also standard in the existing theory of stability estimates for inverse estimation.

Assumption 3. *There exists a $\theta_0 \in B_{\widetilde{H}^\beta}(M) \cap \Theta$ for some $M > 0$ and $\beta > \frac{d}{2}$ such that $(Y_i, X_i) \overset{iid}{\sim} P_{\theta_0}$. Furthermore, the prior* (2.5) *satisfies*

$$\Pi_N(\theta : \|G(\theta) - G(\theta_0)\|_{L^2(\mathscr{X})} \leq \varepsilon_N) \geq e^{-AN\varepsilon_N^2}$$

for some $A > 0$ and a sequence ε_N satisfying $\varepsilon_N \to 0$, $\varepsilon_N \gg \sqrt{\frac{\log(N)}{N}}$ and $N\varepsilon_N^2 \to \infty$.

For the following assumptions, let $\frac{d}{2} < s < (\alpha - \frac{d}{2})$ be a fixed number such that $\widetilde{H}^s$ is of full measure. The choice of s can be arbitrary and depend on α. Then write $\beta' = s \wedge \beta$.

Assumption 4. *For all $M > 0$ and whenever $\theta_1, \theta_2 \in \Theta \cap B_{\widetilde{H}^{\beta'}}(M)$, there exist $U = U(M)$ such that*

$$\|G(\theta)\|_{L^\infty(\mathscr{X})} \leq U$$

and $L = L(M) > 0$, $\kappa \geq 0$, $\eta \in (0,1]$ such that

$$\|G(\theta_1) - G(\theta_2)\|_{L^2(\mathscr{X})} \leq L\|\theta_1 - \theta_2\|_{\widetilde{H}^{-\kappa_1}(\mathscr{Z})},$$

$$\|G(\theta_1) - G(\theta_2)\|_{L^\infty(\mathscr{X})} \leq L\|\theta_1 - \theta_2\|^\eta_{L^\infty(\mathscr{Z})}.$$

Assumption 5. *Let T be a measurable linear subspace of $L^\infty(\mathscr{Z})$ such that $\theta_0 + |s|T \subset \Theta$ for small enough $s \in \mathbb{R}$. There exists a continuous linear operator*

$$\mathbb{I}_{\theta_0} : \left(T, \langle\cdot,\cdot\rangle_{L^2(\mathscr{Z})}\right) \to L^2(\mathscr{X}),$$

satisfying, for some $\rho \in (1,2]$,

$$R_{\theta_0}(h) := \|G(\theta_0 + h) - G(\theta_0) - \mathbb{I}_{\theta_0}(h)\|_{L^2(\mathscr{X})} = O(\|h\|^\rho_{L^\infty(\mathscr{Z})}) \tag{2.8}$$

as $\|h\|_{L^\infty(\mathscr{Z})} \to 0$ for every $h \in T$.

Assumption 6. *Suppose $h \in \overline{T} \cap B_{\widetilde{H}^\alpha}(M)$ for $\overline{T}$ the closure of T in $L^2(\mathscr{Z})$. There exists $\eta^{\text{lin}} \in (0,1)$ such that for all $M > 0$ there exists a positive $c = c(M)$ satisfying*

$$\|h\|_{L^2(\mathscr{Z})} \leq c(M)\|\mathbb{I}_{\theta_0}(h)\|^{\eta^{\text{lin}}}_{L^2(\mathscr{X})}.$$

Assumption 7. *Define the event, for sequences $\overline{\varepsilon}_N \to 0$ satisfying $N\overline{\varepsilon}_N^2 \to \infty$ and a constant $M > 0$,*

$$\Theta_\infty(M, \overline{\varepsilon}_N) := \{\theta : \|\theta - \theta_0\|_{L^\infty} \leq \overline{\varepsilon}_N,\ \|\theta\|_{\widetilde{H}^s} \leq M\}. \tag{2.9}$$

The posterior distribution (2.6) *satisfies*

$$P^N_{\theta_0}(\Pi_N(\Theta \setminus \Theta_\infty(M, \overline{\varepsilon}_N)|\mathscr{D}^N) \geq e^{-LN\varepsilon_N^2}) \to 0,$$

for $L = 2(2U^2+1)+A$ with A in Assumption 3, $U = U(M)$ in Assumption 4 and ε_N in Assumption 3.

Whenever this Assumption is met, we say the posterior contracts at a rate of $\overline{\varepsilon}_N$.

Assumption 8. *Let the sequence $\sigma_N \geq \overline{\varepsilon}_N^{\rho}$ for all $N \geq 1$ and ρ from Assumption 5. Let the sequence J_N satisfy, for all $N \geq 1$,*

$$J_N \equiv J_N(t_1,t_2) \geq \int_0^{t_1} \sqrt{\log 2\mathscr{N}(\Theta_\infty(M,\overline{\varepsilon}_N), \|\cdot\|_{L^\infty(\mathscr{Z})}, t_2\delta)}\, d\delta,\ t_1,t_2>0. \tag{2.10}$$

For the exponent η from Assumption 4, we have

$$\max\left\{N(\sigma_N^2+\sigma_N\overline{\varepsilon}_N^{\eta}),\ \sqrt{N}\overline{\varepsilon}_N^{2\eta}J_N(1,\overline{\varepsilon}_N^{2\eta}),\ \sqrt{N}J_N(\sigma_N,1),\ \frac{\overline{\varepsilon}_N^{\eta}\sqrt{\log N}}{\sigma_N^2}J_N^2(\sigma_N,1)\right\} \to 0.$$

Remark. Assumptions 3 and 4, together with the regularity conditions assumed on the Gaussian prior in the previous Section, indeed imply the Bochner integrability of a random variable with the posterior measure (2.6) given almost every realization of the observed data $\mathscr{D}^N \in (\mathbb{R}\times V)^N$ (Monard, Nickl, and Paternain, 2021). This justifies the use of expressions such as $\overline{\theta}_N$ at the beginning of this Section. Furthermore, any continuous linear functional $\Psi(\theta)$ is measurable in the usual sense. For completeness, the proof of this claim is attached to the Supplement.

Many of the above assumptions have been introduced in some form by Monard, Nickl, and Paternain, 2021 to exercise an appropriate control of the approximation error induced by linearization of "pointwise" observations of $G(\theta)$. Some minor modifications were necessary to handle our motivating example, estimation for the divergence form elliptic equation, which was not covered by the mentioned authors' theorem. A novel, key assumption introduced here is Assumption 6. A similar assumption has been used by Nickl and S. Wang, 2022, where an appropriate choice of the tangent space T and a weaker norm than $\|\cdot\|_{L^2(\mathscr{Z})}$ were used to deduce in-probability lower bound for the smallest eigenvalue of the Hessian of a sieve-type likelihood. That our assumption only considers those tangent vectors belonging to $B_{\tilde{H}^\alpha}(M)$ reflects the fact that when efficiency theory fails, the choice of a prior will have a real effect on the frequentist coverage of Bayes credible sets. We note that conditions similar to Assumption 6 have played an important role in convergence rate analysis

of regularization in Hilbert scales (Heinz Werner Engl, Hanke, and Neubauer, 1996; Hohage and Pricop, 2008).

2.2.2 General Theorems

We consider two separate cases, depending on whether ψ belongs to the range of $\mathbb{I}^*_{\theta_0}$ or not. If it does, one can expect some version of Bernstein-von Mises theorem should hold, given that this is a necessary condition for the existence of a semiparametric efficient estimator. On the other hand, as shown by Nickl and Paternain, 2022, the geometric obstructions inherent to linearizing the map G can rule out a wide class of smooth functionals from $R(\mathbb{I}^*_{\theta_0})$. Thus, a general theorem stating that at least the Bayes credible set is conservative, but not overconfident, can be useful. The first theorem thus picks up where Nickl and Paternain, 2022 left off and gives sufficient conditions under which a Gaussian posterior approximation is asymptotically valid even when $\psi \notin R(\mathbb{I}^*_{\theta_0})$. As we will see, the frequentist coverage expected from the resulting credible set does not exactly meet the stated nominal level.

Define, for the following theorems, the sequence of random variables

$$Z_N \equiv Z_N(\psi, \theta_0) := \Psi(\theta - \theta_0) - \widehat{\Psi}_N, \tag{2.11}$$

and its "centering term"

$$\widehat{\Psi}_N := \sum_{i=1}^{N} \varepsilon_i \mathbb{I}_{\theta_0}(\overline{\psi}_N)(X_i). \tag{2.12}$$

Let d_{weak} be an arbitrarily chosen metric that metrizes the weak topology of probability measures.

Theorem 2.2.1. *Suppose Assumptions 1-8 hold. Let $\psi \notin R(\mathbb{I}^*_{\theta_0})$ and define $\omega := \frac{\eta^{\text{lin}}}{2-\eta^{\text{lin}}}$.*

(i) If there exists $a \in (0, 1 - \frac{d}{2\alpha})$ such that

$$\tau_N \ll N^{\frac{\omega a}{2(1-\omega a)}} \varepsilon_N^{\frac{1}{1-\omega a}}; \tag{2.13}$$

and

(ii) either

(a) $\theta_0 = 0$*; or,*

(b) $\theta_0 \in \widetilde{H}^\beta$ *and the sequence* τ_N *satisfies*

$$\tau_N \gg N^{\frac{\omega\nu}{2(1-\omega\nu)}},\ \nu = \frac{\alpha-\beta}{\alpha\vee\beta}; \tag{2.14}$$

then, we have

$$d_{weak}(\mathscr{L}(s_N^{-1}Z_N|\mathscr{D}^N), \text{Normal}(0,1)) \overset{P_{\theta_0}^N}{\to} 0.$$

In particular, condition (2.14) *is met for* $\tau_N \equiv 1$ *if and only if* $\beta > \alpha$*.*

The next theorem handles the case when ψ does belong to the range of $\mathbb{I}_{\theta_0}^*$. In this case, Assumption 6 plays no role and we obtain the so-called Bernstein-von Mises theorem in its familiar form.

Theorem 2.2.2. *Suppose Assumptions 1-5 and 7 hold. Let* $\psi \in R(\mathbb{I}_{\theta_0}^*)$ *with* $\mathbb{I}_{\theta_0}^*\varphi = \psi$*. If Assumption 8 holds with* s_N (A.3) *replaced by*

$$s_N = \frac{1}{\sqrt{N}}\|\varphi\|_{L^2(\mathscr{X})} = \frac{1}{\sqrt{N}}\|(\mathbb{I}_{\theta_0}^*\mathbb{I}_{\theta_0})^{-1/2}(\psi)\|_{L^2(\mathscr{X})},$$

and either

(i) $\theta_0 = 0$*; or,*

(ii) $\theta_0 \in \widetilde{H}^\beta$ *and* τ_N *satisfies*

$$\tau_N \gg N^{-(\frac{2\beta-\alpha}{4\beta}\wedge\frac{1}{4})}; \tag{2.15}$$

then,

$$d_{weak}(\mathscr{L}(\sqrt{N}Z_N|\mathscr{D}^N), \text{Normal}(0, \|\varphi\|_{L^2(\mathscr{X})}^2)) \overset{P_{\theta_0}^N}{\to} 0.$$

The sufficient conditions of the previous two Theorems also suffice to yield a theorem characterizing the frequentist coverage of a credible interval for linear functionals (2.7).

Theorem 2.2.3. *The coverage defined as*

$$\liminf_{N\to\infty} P_{\theta_0}^N(\theta_0 \in I_N),$$

for I_N *in* (2.7)*, is*

(i) $c \in (1-\gamma, 1]$ *if* $\psi \notin R(\mathbb{I}^*_{\theta_0})$ *and Theorem 2.2.1 holds;*

(ii) $1-\gamma$ *if* $\psi \in R(\mathbb{I}^*_{\theta_0})$ *and Theorem 2.2.2 holds.*

For interpretation, let us first consider Theorem 2.2.2, which improves the assumptions needed for when Bernstein-von Mises theorem *does* hold compared with the previous result, namely Theorem 3.7, Monard, Nickl, and Paternain, 2021. The condition of this result, in our notation, may be written as

$$\psi \in R(\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0}\Lambda^{-\alpha/2}).$$

In contrast, we impose the weaker condition that $\psi \in R(\mathbb{I}^*_{\theta_0})$. As noted by A. Van Der Vaart, 1991, the demand $\psi \in R(\mathbb{I}^*_{\theta_0})$ is genuinely weaker than the demand $\psi \in R(\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0})$ when $\mathbb{I}^*_{\theta_0}$ does not have a closed range. The condition (2.15) is needed to guarantee the accuracy of a Gaussian approximation, for if $\theta_0 \notin \widetilde{H}^\alpha$, the "bias" of the credible set can have adverse effect on the frequentist coverage and τ_N may need to be adjusted judiciously. It can be easily shown that in the case when $\alpha = \beta$, the specific scaling choice of τ_N considered by Monard, Nickl, and Paternain, 2021 does imply our condition (2.15); a very similar proof is given for the diffusion coefficient estimation problem in the Appendix. We may also observe the weakened condition $\psi \in R(\mathbb{I}^*_{\theta_0})$ is *necessary* for the existence of a $\sqrt{N}$-consistent regular estimator of functional Ψ (A. Van Der Vaart, 1991; A. W. Van Der Vaart and Wellner, 1996), so that in general one cannot attain the same convergence result stated in Theorem 2.2.2 for any broader class of linear functionals. In this sense, Corollary 2.2.2 strictly improves the previous result.

When $\psi \notin R(\mathbb{I}^*_{\theta_0})$, Theorem 2.2.1 shows that the coverage of a credible ball based on L^2 distance then cannot meet its nominal level. The difficulty here is that a credible interval can have asymptotically non-negligible bias, unless θ_0 satisfies a strong smoothness condition. When $\alpha \geq \beta$, in particular, (2.14) is satisfied *only if* $\tau_N \gg 1$. In general, using such "diluted" priors leads to very poor estimates for the inverse problem. We discuss this matter further in the next Section.

2.3 Application to Divergence Form PDEs

This Section summarizes the application of the general theorems in the preceding Section to the special case of an inversion in an elliptic problem with a divergence form operator. All proofs for this Section are included in the Supplement.

We mainly focus on the special case of Darcy flow and study the frequentist coverage of Bayes credible intervals for linear functionals. Despite its simplicity, this problem is illuminating because it presents geometric obstructions to the compatibility between the family of efficiently estimable functionals and classical smoothness scales. For instance, as Nickl and Paternain, 2022 showed, $\psi \in C_0^\infty(\mathscr{Z})$ need not belong to $R(\mathbb{I}^*_{\theta_0})$ even for the simplest problems, so that *a fortiori* Bernstein-von Mises theorem must fail. This is disappointing, as one is left without knowledge about the frequentist coverage of credible intervals even in the large N-limit. We now show that under some smoothness conditions on θ_0, one can expect the intervals to be at least conservative. The fundamental trade-offs, between fast convergence rate in estimation and guarantees for such conservative coverage, will become clear.

Using the notation used in the previous Sections, let us consider the divergence form equation with inhomogeneous boundary condition:

$$\begin{cases} L_a u := \nabla \cdot (a \nabla u) = f & \text{on } \mathscr{Z}, \\ u = g & \text{on } \partial \mathscr{Z}. \end{cases} \tag{2.16}$$

The functions f, g are assumed to be given and belong to $C^\infty(\mathscr{Z})$. For a fixed, unknown conductivity field a_0, we observe the noisy "pointwise" measurements $Y_i = u(X_i) + \varepsilon_i$. We denote the corresponding PDE solution $u \equiv u_{f_0}$. If

$$\mathscr{C} := \{a \in H^s(\mathscr{Z}) : a(x) \geq K_{min} \text{ for every } x \in \mathscr{Z}\},$$

where $s > 1 + d/2$, the differential operator $u \mapsto \nabla \cdot (f \nabla u)$ is strongly elliptic. Furthermore, by Hölder embedding of H^s, we have $a \in C^r(\mathscr{Z})$ for some $r > 0$, so the equation (2.16) has a unique solution $u_{a_0} \in C(\overline{\mathscr{Z}}) \cap C^{2+r}(\mathscr{Z})$ by Theorem 6.13 (Gilbarg et al., 1977). The

parameter space $\mathscr{C}$, however, is not a linear space. A convenient practice here, as in Section 3.3 (Andrew M Stuart, 2010), is to parameterize

$$a_\theta = e^\theta + K_{min},\ \theta \in \Theta,$$

so that for a linear space Θ, $a_\theta \in \mathscr{C}$. The analysis can then exploit results in Section 2 (Nickl, 2023) that are tailored to the exponential map. An alternative that rules out the use of exponential map is to consider a more restrictive class of regular link functions due to Nickl, Geer, and S. Wang, 2020. While somewhat ad-hoc, we choose the above link function mainly for convenience in implementation.

We now define the forward map G as the mapping $\theta_0 \mapsto u_{a_0}$ that solves the equation (2.16) for $a_0 \equiv a_{\theta_0}$. The parameter space Θ can be now chosen to be H^s, $s > 1 + d/2$. For this class of problems, we consider the prior construction

$$\theta \sim \text{Normal}(0, -\Delta^{-\alpha}),\ \alpha > 1 + d. \tag{2.17}$$

Here Δ is the Dirichlet Laplacian operator. The condition on α, together with the well-known Weyl asymptotics (Weyl, 1911), implies that this is a Gaussian prior on $L^2(\mathscr{Z})$, satisfying Assumption 2. The Hilbert scales generated by Δ enforces increasingly strong boundary conditions on θ_0 and its derivatives. For instance, it is not difficult to derive the facts that $\widetilde{H}^1 = H_0^1$ and $\widetilde{H}^2 = H_0^1 \cap H^2$ (Taylor, 2013). These spaces are continuously embedded in H^s for each $s \geq 0$, thereby also meeting Assumption 1. Repeated application of Poincaré inequality shows that for every element $\theta \in \widetilde{H}^s$, $s \geq 0$, the norms $\|\cdot\|_{H^s}$ and $\|\cdot\|_{\widetilde{H}^s}$ are equivalent.

2.3.1 Convergence Rate and Uncertainty Quantification

The following assumption was first made by Richter, 1981 to derive continuous dependence of the inverse solution a_0 on the solution of the elliptic problem. As mentioned by the author, this encapsulates problems where the source function f, possessing sufficient regularity on the boundary, is strictly positive, or where the boundary data is zero. In this case, it is possible to deduce a quantitative estimate of $\|G(\theta) - G(\theta_0)\|_{L^\infty}$ based on $\|\theta - \theta_0\|_{C^2}$. We choose to operate in this setting, as the sufficient condition of our results will demand that

both the Gaussian prior samples and θ_0 are sufficiently smooth. We refer readers interested in an analytic exposition when θ_0 has low regularity to the results by Bonito et al., 2017.

Assumption 9. *There exist sets $\mathscr{X}_1$ and $\mathscr{X}_2$, possibly disconnected, that form a partition of $\mathscr{X}$, where we have*

$$\inf_{x\in\mathscr{X}_1} \|\nabla G(\theta_0)\|_{\mathbb{R}^d} \geq k_1 > 0, \ \inf_{x\in\mathscr{X}_2} \Delta G(\theta_0) \geq k_2 > 0.$$

While our proof of the following theorem mostly follows the already published proof of Giordano and Nickl, 2020, we consider slightly more general scaling sequences (τ_N) when θ_0 need not necessarily belong to the Cameron-Martin space $\widetilde{H}^\alpha$ of the prior.

Theorem 2.3.1. *Let $\Theta = H^p$, $p > 1 + \frac{d}{2}$ and suppose $\theta_0 \in \widetilde{H}^\beta$ where $\beta \geq p$. Suppose Π_0 is a centered Gaussian measure with covariance operator $(-\Delta)^\alpha$, with Δ a Dirichlet Laplacian on $\mathscr{Z}$ and $\alpha > 1 + d$. Define $\Pi_N = \tau_N^* \Pi_0$ with*

$$\tau_N^* := \begin{cases} N^{\frac{2\alpha-2\beta-d}{4\alpha+4+2d}} \wedge 1 & \text{if } \alpha \geq \beta, \\ N^{-\frac{d}{4\alpha+4+2d}} & \text{if } \alpha < \beta. \end{cases}. \tag{2.18}$$

Under Assumption 9, the posterior contracts at a rate ε_N^r:

$$P_{\theta_0}^N \left(\Pi\left(\theta : \|\theta - \theta_0\|_{L^2} \geq \varepsilon_N^r, \ \|\theta\|_{H^s} \leq M | \mathscr{D}^N\right) \geq e^{-LN\varepsilon_N^2} \right) \to 0, \tag{2.19}$$

where $s \in (\frac{d}{2}, \alpha - \frac{d}{2})$ and the rate sequence can be chosen to be

$$\varepsilon_N := \begin{cases} N^{-\frac{\alpha\wedge\beta+1}{2(\alpha\wedge\beta)+2+d}} & \text{if } \alpha \leq \beta + \frac{d}{2} \\ \gg N^{-\frac{\beta+1}{2\beta+2+d}} & \text{if } \alpha > \beta. \end{cases} \tag{2.20}$$

The exponent can be chosen to be $r = \frac{\beta'-1}{\beta'+1}$ with $\beta' = s \wedge \beta$. Furthermore, we can replace $L^2(\mathscr{Z})$-norm in (2.19) *with $L^\infty(\mathscr{Z})$-norm and the exponent r to be $\frac{\beta'-1-d/2}{\beta'+1}$ whenever $1 + \frac{d}{2} < \beta' < (\alpha - \frac{d}{2}) \wedge \beta$.*

Remark. When $\beta \leq \alpha \leq \beta + \frac{d}{2}$, the posterior of $G(\theta)$ (and not θ) contracts towards neighborhoods of $G(\theta_0)$ at a rate that is known to be minimax optimal in nonparametric regression with white

noise when the underlying function belongs to $B_{H^\beta}(M)$ (Bissantz, Hohage, and Munk, 2004). Even for this "forward problem," such a rate is not achievable if α is not close enough to β. We have explicitly restricted ourselves to considering only the scaling sequences $\tau_N \lesssim 1$. Proceeding as B. Knapik, A. v. d. Vaart, and J. v. Zanten, 2011, one may want to consider *dilating* sequences $\tau_N \to \infty$ when $\alpha > \beta + d/2$. Inspection of the proof with such a formal substitution shows that the posterior of $G(\theta)$ does contract towards $G(\theta_0)$ at the optimal rate in the above sense, even when $\alpha > \beta + \frac{d}{2}$. However, the argument breaks down when we want to transfer the claim to posterior contraction towards θ_0. This is because the method of proof, following Giordano and Nickl, 2020, relies on an "inverse stability" estimate of the form

$$\|\theta - \theta_0\|_{L^2} \lesssim \|G(\theta) - G(\theta_0)\|_{H^2}.$$

When $\tau_N \to \infty$, the multiplicative constant on the right hand side can no longer be chosen to be uniformly bounded in N, because Π_N does not concentrate on a subset of Θ that is uniformly bounded in norm. See also the discussion in Section 2.1.2 (Nickl, 2023) for a discussion of the difficulties unique to the nonlinear setting.

Theorem 2.3.2. *Consider* τ_N (2.18) *and* ε_N (2.20) *as in the previous Theorem and assume* $\alpha \leq \beta + \frac{d}{2}$. *The coverage* (2.2.3) *of* I_N *belongs to* $(1-\gamma, 1]$ *if* $\psi \notin R(\mathbb{I}^*_{\theta_0})$ *and*

$$\frac{2\alpha-1-d}{2\alpha+2+d}\frac{\alpha-1-d}{\alpha+1-d/2} > \begin{cases} \frac{1}{3} & \text{if } d=1, \\ F(\alpha) := \frac{\alpha-d/2}{3\alpha-2d+1} \vee \frac{(\alpha-d/2)^2}{4(\alpha-d)(\alpha+1-d)} & \text{otherwise}; \end{cases} \tag{2.21}$$

$$\alpha > \frac{d}{2} \times \begin{cases} \frac{6(\alpha-1)(\alpha+1)}{6(\alpha-1)(\alpha+1)-(\alpha+3)(4\alpha+3)} & \text{if } d=1, \\ \left(1 + \frac{d}{2}\frac{\alpha+3}{(\alpha-1)(\alpha+1)} - \frac{(\alpha+3)(2\alpha+2+d)}{(\alpha-1)(\alpha+1)}F(\alpha)\right)^{-1} & \text{otherwise}; \end{cases} \tag{2.22}$$

$$\beta > \alpha \times \frac{2(\alpha-1)(\alpha+1)}{2(\alpha-1)(\alpha+1)-d(\alpha+3)}. \tag{2.23}$$

On the other hand, the coverage of I_N *for the same choice of* τ_N *is* $1-\gamma$ *if* $\psi \in R(\mathbb{I}^*_{\theta_0})$ *and only the first display above is satisfied.*

While the conditions look formidable, they are easily checked whenever d is fixed. In fact, for any fixed d, it is apparent that the inequalities (2.21) and (2.22) must be satisfied

for all large enough α, because the right hand sides of these inequalities are strictly decreasing outside some neighborhood of zero. However, they can diverge to $-\infty$ or ∞ in that neighborhood, so that there are "spurious" regions of small αs that can still satisfy either conditions (2.21) or (2.22). This is why, strictly speaking, we cannot reduce the inequalities any further. When $d = 2$, it can be checked that (2.21) and (2.22) are met for integer orders $\alpha > 10$. In this region of α, condition (2.23) is met if $\beta \geq \alpha + 2$, so that the prior "slightly underestimates" the smoothness of θ_0.

3. Generalized Bayes Approach to Inverse Problems with Model Misspecification

In this Chapter, we hearken back to the general formulation of the problem, but now with *no assumption* about the existence of a true $\theta_0 \in \Theta$ indexing the likelihood that agrees with the data generating procedure giving rise to the observed data. We refer to this situation as "model misspecification"; vice versa, the "correctly specified" likelihood is the one which accurately describes the data generating procedure for *some* parameter value. We review the Gibbs posterior framework, proposed by various authors (P G Bissiri, C C Holmes, and S G Walker, 2016; Jiang and Tanner, 2008; Dunlop and Yang, 2021; Zou et al., 2019), that solves a variational problem on the space of probability measures to construct probabilistic inverse solutions similar to Bayes posteriors. Our novel contribution here is twofold, both relating to model selection. First, we present a generalized version of posterior predictive distributions that can be used to compare predictive performance of two distinct losses used to construct the variational problem. Second, we develop a cross-validation strategy for choosing the regularization parameter, denoted W, that balances the relative weight between the data-informed loss and the penalty functional. Some theoretical properties of Gibbs posteriors and their predictive analogues, along with numerical illustrations, close this Chapter.

A major caveat is that most of the analysis presented concerns *finite-dimensional* parameter spaces. The restriction of our analysis stands in contrast to the general results in Chapter 2 and existing works such as A M Stuart, 2010 which focus on infinite-dimensional Hilbert or Banach spaces. It also differs from the "high-dimensional" analysis of works such as Castillo and Rousseau, 2015, as the fixed number of dimension is not allowed to grow together with the number of observations. The practical justification for this restriction is the awareness on the statistician's part of a potentially severe model misspecification, a situation which reasonably describes the ultrasound vibrometry application covered in this Chapter. The asymptotic analysis can then serve the purpose of whether the finite-dimensional model still targets a "reasonable solution," without suggesting the model uncertainties can be resolved by even an infinite amount of observations. This notion of "reasonable solution" is

defined in the Chapter based on basic ideas from M-estimation theory.

Most of the materials in this Chapter have been published as Baek, Aquino, and Mukherjee, 2023.

3.1 Gibbs Posterior with Model Selection

We review the foundations for the Gibbs posterior framework and describe properties of the Gibbs posterior proposed by P G Bissiri, C C Holmes, and S G Walker, 2016. Section 3.1.4 describes the problem of model comparison and our original contribution of predictive model selection theory for Gibbs posteriors. While the range of problems we have in mind is the same as that mentioned in the previous Chapter, some notations will be revised to handle the problem from a different perspective.

3.1.1 Notations

Throughout this Chapter, we denote as $||\cdot||$ the norm in an Euclidean space $\mathbb{R}^m$. We write as $\Delta(\mathscr{X})$ the space of all probability distributions on $\mathscr{X} \subset \mathbb{R}^m$, assuming standard Borel σ-algebras. For two probability measures $\mu, \nu \in \Delta(\mathscr{X})$, $d_{TV}(\mu,\nu), d_H(\mu,\nu)$, and $D_{KL}(\mu,\nu)$ denote the total variation metric, Hellinger metric, and Kullback-Leibler (KL) divergence (Gibbs and Su, 2002), respectively. For two probability measures μ, ν (possibly on different spaces $\mathscr{X}_1$ and $\mathscr{X}_2$), we denote by $\mu \otimes \nu$ their product measure. For a probability measure $\mu \in \Delta(\mathscr{X})$, $L^q(\mathscr{X};\mu)$ is the space of all functions $f : \mathscr{X} \to \mathbb{R}$ that are L^q-integrable with respect to μ, where $q \in [1,\infty]$.

3.1.2 Parametric Inverse Problems with Model Uncertainty

Throughout this paper, we assume observing n i.i.d. variables that take values in $\mathscr{Y} \subset \mathbb{R}^d$ with an unknown probability distribution $\mathbb{P}$:

$$y_i \overset{iid}{\sim} \mathbb{P} \equiv \mathbb{P}_{\mathscr{F}(\theta_0)}. \tag{3.1}$$

Here, the parameter θ_0 is a physically meaningful parameter that characterizes the observed system. The parameter-to-observation map $\mathscr{F}(\theta)$ is often defined in relation to the parameterized PDE model as in the previous Chapter. Since our numerical studies consider more

involved examples in engineering applications, we explicitly parameterize a general PDE thus:

$$\mathscr{M}(u(\theta);\theta) = 0,\ u(\theta) \in \mathscr{U},\ \mathscr{M} : \mathscr{U} \to \mathscr{V}^*, \tag{3.2}$$

where $\mathscr{U}, \mathscr{V}$ are Hilbert spaces with $\mathscr{V}^*$ being the dual space of $\mathscr{V}$. We assume for every θ there exists a unique $u(\theta)$ that satisfies (3.2). The parameter-to-observation map is defined as $\mathscr{F}(\theta) := \mathscr{D}u(\theta)$, where $\mathscr{D}$ is the observation operator. Note that $\mathscr{F}$ thus essentially plays a role of the "forward operator" G in the previous Chapter. However, we no longer posit restrictions on the exact manner in which the observation is discretized. Nor do we posit an explicit relationship between the stochastic "noise" and the evaluation of the forward map $\mathscr{F}(\theta_0)$; in particular, the observations need not arise from an additive noise model over a grid of points.

In the classical Bayesian framework, the parameterization of the sampling distribution (3.1) by the forward model (3.2) is *known*. Examples include the additive white noise model (B T Knapik, A W van der Vaart, and J H van Zanten, 2011) and Poisson likelihood (Barmherzig and Sun, 2022). However, in practice, this need not be true because either the hypothesized parameterization is *incorrect*, or the model uncertainties are so large that such a parameterization is difficult. Both errors in the forward model and errors in the noise distribution contribute to an incorrect parameterization of the likelihood. There are many ways in which such a mismatch can arise. A concrete example in ultrasound vibrometry application is demonstrated in Section 3.4. Here, we only mention that both philosophical and asymptotic justification of Bayesian inference is more tenuous with model misspecification. While the modeler may use a "surrogate likelihood" to define a misspecified Bayes posterior and argue that one can obtain good approximations when the surrogate is "close" to $\mathbb{P}_\theta$, defining this "closeness" in nonlinear inverse problems is not trivial.

In the next Section, we review a variational formulation that bypasses these difficulties. Instead of trying to define the correctly parameterized $\boldsymbol{Q}_\theta = \mathbb{P}_\theta$, the variational perspective *defines* a discrepancy between the posited forward model and the observed data. The relative weights given to possible parameters θ are higher if they yield smaller discrepancy or loss.

3.1.3 Variational Framework for Gibbs Posteriors

Let $L : \Theta \times \mathbb{R}^d$ be a loss function. Let $\rho_0 \in \Delta(\Theta)$. We propose to solve the problem proposed by P G Bissiri, C C Holmes, and S G Walker, 2016:

$$\hat{\rho}_n^W(d\theta) := \underset{\rho \in \Delta(\Theta)}{\arg\min} \left[\mathscr{R}_W(\rho) = \int_\Theta \frac{1}{n} \sum_{i=1}^n L(\theta, y_i)\ \rho(d\theta) + \frac{1}{nW} D_{KL}(\rho || \rho_0) \right]. \tag{3.3}$$

Here, ρ_0 is the distribution quantifying our prior and $W > 0$ is a regularization parameter that we assume is given for now. If ρ is not absolutely continuous with respect to ρ_0, the divergence is defined to be $+\infty$. Often we will abbreviate the average loss over all data by $R_n(\theta) := \frac{1}{n}\sum_{i=1}^n L(\theta, y_i)$.

To ensure the existence of a solution, we will make assumptions on the structure of the problem, motivated by the assumptions of Cotter, Dashti, et al., 2009 and A M Stuart, 2010.

Assumption 10. *Let the loss $L(\theta, y)$ have the form $l(\mathscr{F}(\theta), y)$ and satisfy the following.*

1. *$L(\theta, y)$ is uniformly bounded from below:*

$$\inf_{\theta, y} L(\theta, y) \geq B > -\infty.$$

   We assume $B = 0$ without loss of generality.

2. *For every θ, y there exists $K \equiv K(||\theta||, ||y||) \in L^1(\Theta \times \mathscr{Y}; \rho_0 \otimes \mathbb{P})$ such that $L(\theta, y) \leq K(||\theta||, ||y||)$.*

3. *For every $r > 0$ there exists $C_1(r, y) > 0$ such that whenever $||\theta_1||, ||\theta_2|| < r$,*

$$|L(\theta_1, y) - L(\theta_2, y)| \leq C_1(r, y)||\theta_1 - \theta_2||$$

   with $C_1(y) \equiv C_1(r, y) \in L^1(\mathscr{Y}; \mathbb{P})$.

4. *For every $r > 0$ there exists $C_2(r, \theta) > 0$ such that whenever $||y_1||, ||y_2|| < r$,*

$$|L(\theta, y_1) - L(\theta, y_2)| \leq C_2(r, \theta)||y_1 - y_2||$$

   with $\exp(C_2(\theta)) \equiv C_2(r, \theta) \in L^1(\Theta; \rho_0)$.

Remark. Note that because L is defined to be a mapping on $\Theta \times \mathscr{Y}$, the regularity assumptions implicitly place restrictions on the forward model $\mathscr{F}$. We will use the squared ℓ^2 loss as an example to understand the regularity conditions

$$L(\theta, y) \equiv l(\mathscr{F}(\theta), y) = ||y - \mathscr{F}(\theta)||^2.$$

The properties of $\mathscr{F}$ dictate whether the loss L satisfies the assumptions. Various PDE-based models used in the literature, combined with the popular Gaussian prior distribution, satisfy assumptions 1 and 4 for squared ℓ^2 loss; see, e.g., Section 3, A M Stuart, 2010. On the other hand, the integrability conditions 2 and 3 depend on the unknown $\mathbb{P}$. One can check how mild or severe these conditions turn are for a specific loss function, by hypothesizing models like (3.20), without specifying a likelihood.

We will also make a mild smoothness assumption on the density of the prior distribution ρ_0. The Gaussian prior satisfies the assumption.

Assumption 11. *ρ_0 has positive density everywhere. Furthermore, for every $r > 0$ there exists $C_3(r) > 0$ such that whenever $||\theta_1||, ||\theta_2|| < r$,*

$$|\log \rho_0(\theta_1) - \log \rho_0(\theta_2)| \leq C_3(r)||\theta_1 - \theta_2||.$$

There exists a unique solution to (3.3) in $\Delta(\Theta)$, which has the following density for fixed $W > 0$:

$$\hat{\rho}_n^W(d\theta) := \frac{\exp\{-nWR_n(\theta)\}\rho_0(d\theta)}{Z_n^W}, \tag{3.4}$$

where the normalizing constant, or the "partition function" Z_n^W, is defined as

$$Z_n^W \equiv \int e^{-nWR(\theta)}\ \rho_0(d\theta). \tag{3.5}$$

To derive this formula, the objective functional can be rewritten as

$$\mathscr{R}_W(\rho) = \frac{1}{nW}\left\{D_{KL}(\rho||\hat{\rho}_n^W) - \log Z_n^W\right\}; \tag{3.6}$$

The first term is non-negative and uniquely attains zero at $\rho \equiv \hat{\rho}$. The second term does not depend on ρ, so the minimum of the functional is achieved at $\mathscr{R}_W(\hat{\rho}_n^W) = -\log Z_n^W$. The possible technical issues are measurability of $\exp(-nWR_n(\theta))$ and finiteness of $\log Z_n^W$, which follow from Assumption 10.

Remark. When we fix $W = 1$ and choose the loss to be the negative log-likelihood: $L(\theta, y) = -\log p(y|\mathscr{F}(\theta))$, the Gibbs posterior coincides with a Bayes posterior update, using $p(y|\mathscr{F}(\theta))$ as its likelihood component. Thus, our Gibbs posterior solution strictly generalizes the Bayes posterior.

We close the Section with some intuition of the role of W in the Gibbs posterior. In the limit $W \to 0$, $\hat{\rho}_n^W \equiv \rho_0$, so there is no update of information from the prior. Smaller W thus more heavily weighs prior information. In the limit $W \to \infty$, $\hat{\rho}_n^W$ concentrates on a set of θs minimizing the loss over the observed data. Larger W thus more heavily weighs information from the data and leads to increased sensitivity to perturbation under noise. Intuition suggests that a prior ρ_0 that is strictly positive on Θ should reflect large uncertainty and that W must be carefully chosen based on the amount of information in the data relative to the prior. Inspection of (3.3) suggests that we are implicitly implementing a discrepancy principle (Nair, 2009) since the divergence penalty has less influence when the sample size is large.

3.1.4 Extension to Model Selection

3.1.4.1 Predictive Model Selection

Solving the variational problem (3.3) still requires a pre-specified choice of loss L. It may appear this requirement is as restrictive as positing the generating process as a better choice of loss hinges on knowledge or assumptions on $\mathbb{P}$. Our first proposal is to *define* a valid way to compare two different losses without requiring knowledge of the data-generating mechanism. The key idea is to compare them based on the ability to make accurate *predictions*, measuring their discrepancy on a future observation. As mentioned, this principle is not new and has been used to improve the robustness of Bayesian model prediction and model checking. The novelty lies in the definition of the predictive density without assuming the likelihood, which

will serve as a natural discrepancy measure between the new observation and prediction.

Consider a common prior distribution ρ_0 and multiple competing losses, $L_1, \ldots, L_k$, defined on subsets $\Theta_1, \ldots, \Theta_k$ of Θ. Given the corresponding set of Gibbs posteriors $\hat{\rho}_{n,1}^{W_1}, \ldots, \hat{\rho}_{n,k}^{W_k}$, we propose the following predictive model comparison principle: map each distribution $\hat{\rho}_{n,m}^{W_m}$ to a prediction taking values in $\mathscr{Y}$ and measure its discrepancy from another measurement on the observation space, y^{new}. The solution minimizing this prediction discrepancy, i.e.

$$\underset{m \in \{1,\ldots,k\}}{\arg\min}\ D_{pred}(y^{new}; \hat{\rho}_{n,m}^{W_m}),$$

for a common discrepancy metric D_{pred}, is chosen to be optimal.

3.1.4.2 Gibbs Predictives

To implement the predictive model selection strategy, we need to choose the predictive discrepancy measure D_{pred}. In Bayesian statistics, similar problems arise when considering k different competing likelihood models describing the data-generating mechanism, none of them required to be correct. The default choice is the log-score of the *posterior predictive distribution* (Vehtari, Gelman, and Gabry, 2017):

$$-\log p_{n,m}(y^{new}) := -\log \int p_m(y^{new}|\theta)\ p_{n,m}(\theta) d\theta, \tag{3.7}$$

where $p_{n,m}(\theta)$ is the Bayes posterior for the m-th model after observing n data points, and $p_m(y|\theta)$ defines the model likelihood. The formed posterior predictive makes a distributional prediction, similar to how the posterior makes a probabilistic estimation by giving different probabilities to possible parameter values. The likelihood model that places the highest posterior predictive density on the next measurement made, y^{new}, is chosen to be an optimal description.

The second proposal in this Section is to generalize the notion of model selection based on posterior predictive score (3.7), similar to how we extended the notion of Bayes posterior in Section 3.1.3. This is a non-trivial problem because evaluating predictive power based on (3.7) already assumes a hypothesized likelihood. Our general Gibbs posterior framework is

more flexible, as it allows a general form of losses, but then we also cannot make sense of a natural distributional prediction (3.7). Nevertheless, we define below the notion of a Gibbs predictive distribution that subsumes the case of Bayes predictive distributions.

Definition 1. Let λ be a Lebesgue density of a probability measure on $\mathscr{Y}$. We define the Gibbs predictive distribution of y^{new} given $\boldsymbol{y} \equiv (y_1, \ldots, y_n)$ by its Lebesgue density, with respect to λ:

$$\hat{p}_n^W(y^{new}) := \frac{\int \exp\{-L(\theta, y^{new})\}\ \hat{\rho}_n^W(d\theta)\lambda(y^{new})}{\iint \exp\{-L(\theta, y^{new})\}\ d\lambda(y^{new})\ \hat{\rho}_n^W(d\theta)}. \tag{3.8}$$

Remark. Definition 1 subsumes the Bayes posterior predictive distributions as a special case. We choose the loss to be $-\log p(y|\theta)$ for some posited likelihood, restrict $W = 1$, and set $\lambda(y') \equiv 1$, to obtain:

$$\hat{p}_n^1(y^{new}) = \frac{\int p(y^{new}|\theta)\ p(\theta|\boldsymbol{y}) d\theta}{\iint p(y'|\theta)\ p(\theta|\boldsymbol{y}) d\theta\ dy'} \equiv p(y^{new}|\boldsymbol{y}),$$

since the denominator is automatically 1. The formal similarity suggests that we are proposing to define a predictive distribution based on the "likelihood" $\exp(-L) \times \lambda$. This generalization is only formal when $\mathscr{Y}$ is non-compact because $\lambda \equiv 1$ is no longer a probability measure. The denominator of (3.8) is often still finite, as is the case for all loss functions used in the numerical experiments in Section 3.4 below, but that does not immediately follow from the regularity assumptions 10.

Definition 1 completes the specification of our novel model selection principle in inverse problems under model uncertainty. Given different Gibbs posterior solutions $\hat{\rho}_{n,1}^W, \ldots, \hat{\rho}_{n,k}^W$ derived under a common prior ρ_0 and different losses $L_1, \ldots, L_k$, we propose to choose the optimal solution by the criterion

$$\underset{m \in \{1,\ldots,k\}}{\arg\min}\ -\log \hat{p}_{n,m}^{W_m}(y^{new}).$$

Given a set of n measurements $y_1, \ldots, y_n$ we implement an approximation of the criterion using LOOCV:

$$\underset{m \in \{1,\ldots,k\}}{\arg\min} \left\{ P_{CV}(m) := -\frac{1}{n} \sum_{i=1}^{n} \log \hat{p}_{(-i),m}^{W_m}(y_i) \right\}, \tag{3.9}$$

where $\hat{p}^{W_m}_{(-i),m}$ indicates the m-th Gibbs posterior solution derived from the dataset minus a datum y_i. Computational aspects will be discussed in the next Section, together with calibrating W_m, for which we again resort to LOOCV. The asymptotics for predictions using our construction as $n \to \infty$ are discussed in Section 3.3.

We conclude this discussion with a remark on computing pointwise evaluations of (1) involving high-dimensional integral $\tilde{Z}(\theta)$ nested inside an integral over Θ. Losses used in practical inverse problems often have enough local structure to yield an integral that can be decomposed into low-dimensional subparts. Furthermore, some popular but simple losses like squared ℓ^2-norm loss lead to analytically tractable integrals. A recipe for evaluating $\tilde{Z}(\theta)$ is only possible by utilizing a more detailed structure of the problem. We will show some examples in Section 3.4.

3.2 Model Calibration and Computation

In almost all practical scenarios, probability distributions of the form (3.4) do not admit a closed-form density, so we focus on sampling algorithms for Gibbs posteriors. MCMC algorithms allow us to draw samples that can target posteriors (Gelman et al., 2013). As long as the regularization parameter W is given, the computational problem in our framework is equivalent to Bayesian methods. The added complexity is to find W given the data.

We mention two important precursors that can be used to learn W. The first is the SafeBayes algorithm (P. D. Grünwald and Ommen, 2017). In view of theoretical results by P. D. Grünwald and Mehta, 2020, mentioned in the previous Section, it attempts to construct a Gibbs posterior that mimics the performance of that for an oracle choice of W. The other is the generalized posterior calibration algorithm (Syring and Martin, 2019), which chooses W so that the Gibbs posterior intervals satisfy valid frequentist coverage.

In inverse problems, one often deals with small sample size n and a forward model $\mathscr{F}$ is expensive to repeatedly evaluate. Calibrating frequentist coverage can be problematic conceptually (due to small n) and computationally (reliance on bootstrap methods). To address these concerns, when n is moderately large, we discount the value of W based on the

expected loss estimated through cross-validation. This step is quite similar to the SafeBayes algorithm, but less computationally demanding. Our method can be seamlessly integrated into the particle filter sampling algorithm for a sequence of Gibbs posteriors. The model selection procedure presented in Section 3.1.4 also relies on a cross-validation objective (3.9) and can be implemented using a particle filter.

3.2.1 Cross-Validation with Multiple Samples

Our choice of sampling method is motivated by the need to calibrate the regularization parameter W based on the data. When $n > 1$ sample observations of the underlying function are observed, frequentist statistics can be invoked to make a conscientious choice of W. The optimal choice of W is defined to minimize the error our inverse solution makes outside the observed data:

$$R(W) = \mathbb{E}\left[\int R_n(\theta)\; d\hat{\rho}_n^W(\theta)\right], \tag{3.10}$$

where the expectation is with respect to the observed data $y_1, \ldots, y_n$. (3.10) is a natural risk to minimize from a frequentist perspective, but we cannot compute it. Instead, we can estimate it through cross-validation. In the special case when each test set consists of a single y_i, we obtain the leave-one-out cross-validation (LOOCV) estimate of $R(\alpha)$ (Hastie, Tibshirani, and Friedman, 2009):

$$R_{CV}(W) = \frac{1}{n}\sum_{i=1}^{n} \int L(\theta, y_i)\; d\hat{\rho}_{(-i)}^W(\theta). \tag{3.11}$$

Here, $\hat{\rho}_{(-i)}^W$ denotes a Gibbs posterior derived as in (3.4), except that we hold out y_i from the dataset $\boldsymbol{y}$.

LOOCV is a popular choice for choosing regularization parameters in linear inverse problems since Golub, Heath, and Wahba, 1979, and there are many variants of the GCV algorithm. However, they are not applicable to our setting because they do not address the fundamental difficulty of computing expectations under different probability distributions with changing W. Instead, we turn to sampling. We set a grid of $\{W_0, W_1, \ldots, W_T\}$, fixing $W_0 \equiv 0$ and $W_1 \equiv \bar{W}$ for an upper bound $\bar{W}$. The goal is to approximate all expectations

involved in (3.11) by Monte Carlo samples drawn from the corresponding distributions. A naive Monte Carlo approach is expensive as it requires sampling from a sequence of in total nT different probability distributions. Thus, we need a sampling method that uses a sequence of distributions that can approximate all expectations involved in (3.11). In this section, we will show how importance sampling based on a carefully chosen sequence of T distributions allows us to estimate (3.11).

3.2.2 Importance Sampling Cross-Validation

Importance sampling yields a useful approximation when considering mild changes in the posterior distribution (Gelman et al., 2013). We first determine a "proposal distribution" $\tilde{\rho} \equiv \tilde{\rho}^W$ for each value of W. Then, the formula for estimating R_{CV} using S Monte Carlo draws from $\tilde{\rho}$ is

$$\hat{R}_{CV}(W) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{s=1}^{S} r_i^W(\theta^{(s)}) L(\theta^{(s)}, y_i)}{\sum_{s=1}^{S} r_i^W(\theta^{(s)})}, \tag{3.12}$$

where the importance weights have the formula

$$r_i^W(\theta) = \frac{\exp\{-W \sum_{j:j\neq i} L(\theta, y_j)\}}{\tilde{\rho}(\theta)}. \tag{3.13}$$

The choice of $\hat{\rho}$ is crucial to numerical stability of (3.12). For example, a seemingly natural choice of $\tilde{\rho}$ is the "full" posterior, $\hat{\rho}_n^W$, but it has a known problem of importance weights having possibly high or infinite variance (Vehtari, Gelman, and Gabry, 2017; Chatterjee and Diaconis, 2018). Recently, Silva and Zanella, 2022 have proposed a different choice of ρ^t that produces importance weights with finite asymptotic variance. The new distribution for each t has a density representation

$$\rho_{mix}^W(\theta) := \frac{\sum_{i=1}^{n} \exp\{-W \sum_{j:j\neq i} L(\theta, y_j)\}}{\sum_{i=1}^{n} \int \exp\{-W \sum_{j:j\neq i} L(\tilde{\theta}, y_j)\}\, d\rho_0(\tilde{\theta})}. \tag{3.14}$$

We set our choice of $\tilde{\rho}$ for each W to be precisely (3.14). Computation of importance weights (3.13) requires tractability of the numerator of density (3.14), as estimator (3.12) renormalizes the weights. As long as we can draw samples from each ρ_{mix}^W for $W \in \{W_1, \ldots, W_T\}$, we can

estimate (3.12). This reduces the burden of sampling from targeting nT distributions to just T distributions.

We can write explicit importance sampling estimators for both calibration of W and model selection (3.9). Let $\theta^{(1)},\ldots,\theta^{(S)}$ be (approximately) drawn from distribution (3.14). The calibration objective (3.11) is approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{s=1}^{S} r_i^W(\theta^{(s)})L(\theta^{(s)},y_i)}{\sum_{s=1}^{S} r_i^W(\theta^{(s)})} \tag{3.15}$$

The model selection objective (3.9) is approximated by

$$-\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\sum_{s=1}^{S} r_i^W(\theta^{(s)})e^{-L(\theta^{(s)},y_i)}}{\sum_{s=1}^{S} r_i^W(\theta^{(s)})\tilde{Z}(\theta^{(s)})}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\log\sum_{s=1}^{S} r_i^W(\theta^{(s)})\tilde{Z}(\theta^{(s)})-\log\sum_{s=1}^{S} e^{(W-1)L(\theta^{(s)},y_i)}\right). \tag{3.16}$$

3.2.3 Particle Filter Approximation

Usually, it is not possible to draw i.i.d. samples from any one distribution of the sequence $\rho_{mix}^{W_t}$ for increasing t. Particle filtering, or sequential Monte Carlo (SMC) algorithms, comprise a suite of techniques to obtain a sequence of S-sized particle approximations for each distribution. While filtering methods are often discussed in the context of estimating time-varying parameters, they are easily adapted to our setting by equating "time" with t.

We start with i.i.d. draws $\{\theta_0^{(1)},\ldots,\theta_0^{(S)}\}$ drawn from the prior, $\hat{\rho}_0 \equiv \rho_0$, since $W_0 = 0$. At each iteration t, an SMC sampler implements three moves:

(a) (Weighting) Each sample is weighted according to the importance sampling formula

$$w_t(\theta_{t-1}^{(s)}) \propto \frac{\rho_{mix}^{t}(\theta_{t-1}^{(s)})}{\rho_{mix}^{t-1}(\theta_{t-1}^{(s)})}$$

where $\rho_{mix}^{t} \equiv \rho_{mix}^{W_t}$. Since the normalizing constants are unknown, each weight is renormalized to

$$\bar{w}_t(\theta_{t-1}^{(s)}) := \frac{w_t(\theta_{t-1}^{(s)})}{\sum_l w_t(\theta_{t-1}^{(l)})}.$$

(b) (Resampling) S new particles $\{\tilde{\theta}_{t-1}^{(s)}\}$ are drawn with probabilities

$$Pr(\tilde{\theta}_{t-1}^{(s)} = \theta_{t-1}^{(s')}) = \tilde{w}(\theta_{t-1}^{(s')}),$$

for possible pairings (s, s').

(c) (Mutation) S Markov chains are run in parallel for $K > 0$ steps, each having $\bar{\theta}_{t-1}^{(s)}$ as initial states. The transition kernel of each chain is constructed to have an invariant measure ρ_{mix}^{t}. $\theta_t^{(s)}$ is assigned the value of the draw from a Markov chain run for K iterations.

The mutation step can be implemented using standard MCMC methods, such as Metropolis-Hastings with a Gaussian proposal. All steps in this algorithm are parallelizable across S particles, an attractive feature for computational efficiency.

3.2.4 Practical Considerations

The sampling method described has several hyperparameters that are potentially critical to reliable calibration and model selection.

First is the choice of grid size. For an upper bound, we fix $\bar{W} \equiv 1$ by normalizing the loss L with an appropriate scale, either physically meaningful or learned from the data. A number of standard estimators like the variance of the data can be used. We can even learn W with uncertainty, extending the loss function to include W as an unknown parameter (P G Bissiri, C C Holmes, and S G Walker, 2016). The grid spacing is also important. P. D. Grünwald and Ommen, 2017 suggest a grid of $(0, 1)$ that are spaced with exponentially growing width. This is at odds with the computational stability of filtering methods, and indeed Chopin and Papaspiliopoulos, 2020 suggest that a grid must be spaced with at least a geometrically decreasing width. Their theoretical results and suggestion do not exactly apply to the properties of the particular sequence of distributions we consider. In numerical implementations, we used the growing-width sequence of P. D. Grünwald and Ommen, 2017. Our observation is that (3.17) can quickly drop at the initial move from $W_0 = 0$ to W_1, but remain relatively stable throughout the rest of the sequence. If importance

weights degenerate quickly, further adaptive tempering steps by Jasra et al., 2011 may be incorporated to interpolate between two steps W_t and W_{t+1}.

Second, we must choose the number of samples S and the number of Markov chain transitions K in step (c). Ideally, these should not be tuning parameters but chosen based on theoretical bounds on the numerical approximation errors induced by approximating $\hat{\rho}_n^t$ with a discrete measure based on the particles. However, the current state-of-the-art bounds appear not sharp enough for many practical purposes. In general, increasing S and K trades better approximation with higher computational costs. Practically, one must tune the parameters based on computational budget and some experience.

Third, we must observe steps (a) and (c) require computing the loss at a new value of the forward map $\mathscr{F}(\boldsymbol{\theta})$, defined as a solution of a PDE. The computational cost for solving a PDE can ramp up quickly, as the mutation step requires K repeated evaluations of the PDE solver for S particles. We follow Chopin and Papaspiliopoulos, 2020 and trigger steps (b) and (c) only when importance weights show signs of degeneracy by diagnosing it with the effective sample size statistic (ESS):

$$\text{ESS}^t = \frac{(\sum_{s=1}^{S} w_t(\boldsymbol{\theta}_{t-1}^{(s)}))^2}{\sum_{s=1}^{S} w_t(\boldsymbol{\theta}_{t-1}^{(s)})^2}. \tag{3.17}$$

ESS is bounded between 0 and 1, higher when the distribution of the importance weights is wider; a necessary but not sufficient condition for good approximation guarantees of importance sampling-based estimate. The algorithm will only trigger steps (b) and (c) if ESS^t falls below a prescribed threshold.

Finally, we note that for applications in which few measurements are assumed independent (small n), greedy minimization of (3.11) is unwise, given the statistical noise inherent in small samples. We follow the suggestion of "one standard error rule" (Hastie, Tibshirani, and Friedman, 2009), which chooses the smallest W_t whose associated value of R_{CV} is one standard error within the value of R_{CV} at the minimizer in the grid. This approach favors smaller values of W than exactly minimizing (3.11), assuming it is better to be conservative

about uncertainties when the model error is potentially large.

3.3 Theoretical Analysis

In this Section, we present various desirable theoretical properties of (3.4). First, extant stability results for Bayesian inverse problem solutions (Cotter, Dashti, et al., 2009; A M Stuart, 2010) are improved to capture the correct order of growth for the metric used to measure perturbation in the posteriors. Second, we review statistical results that ensure consistency of Gibbs posteriors as $n \to \infty$. The explicit restriction on the dimension of the parameter space effectively converts the estimation problem from an "ill-posed" one (as in Chapter 2) to a "well-posed" one, so the consistency analysis can be carried out through rather standard methods in asymptotic statistics.

For the proof of the results in this Section, we refer the reader to the Appendix of Baek, Aquino, and Mukherjee, 2023.

3.3.1 Continuity in Data

The first stability result is on the continuity of Gibbs posterior in the underlying data. There exist many metrics for probability distributions, so one can obtain stronger theoretical guarantees depending on the choice. Cotter, Dashti, et al., 2009 and A M Stuart, 2010 have first presented stability results for *Bayes posterior* that bound the *Hellinger distance* from above by the distance between the underlying data. Since Hellinger distance is bounded above by 1, their bound is vacuous when the perturbation in the data is large. Below, we present a tighter, non-vacuous upper bound that is always at most 1. Our result first bounds the KL divergence between posteriors and exploit inequalities between different metrics of probability measures due to Bretagnolle and Huber, 1979. Although KL divergence is asymmetric, our bound is valid for "both directions."

Theorem 3.3.1. *Let $r > 0$, $W > 0$, and $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathscr{Y}^n$ satisfying $\max_{i=1}^n \{||y_{1,i}||, ||y_{2,i}||\} < r$. For two corresponding posteriors $\hat{\rho}_{n,1}^W$ and $\hat{\rho}_{n,2}^W$, there constants a, b that only depend on the prior ρ_0 and*

$r > 0$, such that

$$D_{KL}(\hat{\rho}_{n,1}^W || \hat{\rho}_{n,2}^W) \vee D_{KL}(\hat{\rho}_{n,2}^W || \hat{\rho}_{n,1}^W) \leq aW(1+e^{bnW})^2 \sum_{i=1}^{n} ||y_{1,i} - y_{2,i}||^2.$$

As a consequence,

$$d_H^2(\hat{\rho}_{n,1}^W, \hat{\rho}_{n,2}^W) \leq \sqrt{1 - \exp\left\{-aW(1+e^{bnW})^2 \sum_{i=1}^{n} ||y_{1,i} - y_{2,i}||\right\}}.$$

The Gibbs posterior is also continuous in the regularization parameter W. The continuity argument justifies learning W from the data as in Section 3.3. The result is an easy consequence of the proof of Theorem 3.3.1 and thus stated without proof.

Theorem 3.3.2. *Suppose $y \in \mathscr{Y}$. Let $W, W' > 0$. Then the map $W \mapsto \hat{\rho}_n^W$ is continuous in the Hellinger distance.*

3.3.2 Finite Approximation

In practice, the exact PDE model (3.2) is approximated using discretization schemes like finite elements methods. In this case, (3.2) is replaced with a discretized counterpart:

$$\mathscr{M}^h(u^h(\theta); \theta) = 0,\ u^h \in \mathscr{U}^h,$$

where $\mathscr{U}^h$ is a finite dimensional projection of $\mathscr{U}$. One then considers a sequence of discretization level indices h that decays to zero. Assuming each discrete forward model is well-posed, we can evaluate an approximate forward map $\mathscr{F}^h : \Theta \to \mathbb{R}^d$. When $u^h(\theta)$ converges to $u(\theta)$ for each θ, we desire convergence of the Gibbs posterior estimated using $\mathscr{F}^h(\theta)$ to that estimated using $\mathscr{F}(\theta)$. This is our next result. Again, a weaker result has been derived by A M Stuart, 2010 for Bayesian problems, and the bounds are improved similarly as in Theorem 3.3.1.

Theorem 3.3.3. *Define a surrogate loss $L^h(\theta, y) = l(\mathscr{F}^h(\theta), y)$, and suppose the new loss satisfies Assumption 10. Furthermore, suppose for every $r > 0$ and for every h, there exist $C(\theta) \equiv C(r, \theta)$*

such that whenever $||y|| < r$,

$$|L(\theta, y) - L^h(\theta, y)| = C(\theta)\psi(h),\ C(\theta) \in L^1(\Theta; \rho_0),$$

where $\psi(h) \to 0$ *as* $h \to 0$. *Then, for two corresponding Gibbs posteriors* $\hat{\rho}_n^W$ *and* $\hat{\rho}_{n,h}^W$, *there exist constant* a, b *that only depend on the prior* ρ_0 *and* $r > 0$ *such that*

$$D_{KL}(\hat{\rho}_n^W || \hat{\rho}_{n,h}^W) \vee D_{KL}(\hat{\rho}_{n,h}^W || \hat{\rho}_n^W) \leq aW(1 + e^{bnW})^2 \psi(h).$$

As a consequence,

$$d_H^2(\hat{\rho}_n^W, \hat{\rho}_{n,h}^W) \leq \sqrt{1 - \exp\{-aW(1 + e^{bnW})^2 \psi(h)\}}.$$

3.3.3 Statistical Consistency

The preceding stability results have been derived conditional on the data $\boldsymbol{y}$. Frequentist analysis of a sequence of Gibbs posteriors $(\hat{\rho}_n^W)_{n=1}^{\infty}$, for a sequence of growing length, reveals the asymptotic convergence properties of our solution. To rigorously define consistency, we need to appropriately define convergence of a sequence of posteriors $(\hat{\rho}_W^n)_{n=1}^{\infty}$. In Bayesian statistics, a sequence of posterior distributions ρ_n is said to be *consistent* if

$$\rho_n \to \delta_{\theta_0} \quad \text{with probability 1;} \tag{3.18}$$

i.e., the distribution converges to a Dirac measure at a true, unknown model parameter θ_0, in the weak topology of probability measures. In typical consistency proofs, we use another definition that is equivalent if Θ is a metric space (Ghosal and A W van der Vaart, 2017):

$$\rho_n(\Theta \setminus U) \to 0 \quad \text{with probability 1,} \tag{3.19}$$

for every open neighborhood U of θ_0. Recall now the general Gibbs posterior framework accounts for the possibility of model misspecification. While it is possible there exists a physically meaningful true parameter θ_0, the true likelihood parameterization $\mathbb{P}_{\theta_0}$ is unknown. Therefore, the same definition of "convergence to truth" is inappropriate.

It is reasonable to instead consider convergence to a set of parameters that minimize the following expected loss (Kleijn and A. v. d. Vaart, 2012):

$$\Theta^* := \left\{ \theta^* : R(\theta^*) = \min_{\theta} R(\theta),\ R(\theta) = \int L(\theta, y)\ \mathbb{P}(y) dy \right\},$$

It is easily checked that by Assumption 10, R is continuous in θ (c.f. proof in appendix). Therefore, if Θ is compact, Θ^* is guaranteed to be non-empty. We also note that in general it is not guaranteed that $\theta_0 \in \Theta^*$.

Consistency in this context is defined as

$$\hat{\rho}_n^W(\Theta \setminus U) \to 0 \quad \text{with probability 1,}$$

for every open set $U \supset \Theta^*$, generalizing (3.19). The following theorem, essentially a consequence of theorems due to Kleijn and A. v. d. Vaart, 2012, provides sufficient conditions under which (3.19) describes the asymptotic behavior of Gibbs posteriors.

Theorem 3.3.4. *In addition the assumptions in Section 3.1, suppose furthermore*

1. *Θ is compact;*
2. *There exists $\bar{W}$ such that $C_1(y)e^{WL(\theta^*, y)} \in L^1(\mathscr{Y}; \mathbb{P})$ for $C_1(y)$ in Assumption 10.3.*

Then, Θ^* is non-empty. Furthermore, for every $W \in (0, \bar{W})$ and every open set $U \supset \Theta^*$,

$$\hat{\rho}_n^W(\Theta \setminus U) \to 0 \quad \textit{with probability 1.}$$

Remark. Condition 2 does not follow from previously stated regularity assumptions. That W must be small enough is intuitively reasonable, as the posterior becomes less stable with large W at any finite n. For the technical subtleties of this condition, see Nicholas Syring and Ryan Martin, 2023. We point out that mere continuity of L and compactness of Θ are in general insufficient to guarantee consistency results under model misspecification as studied by Kleijn and A. v. d. Vaart, 2012.

In Section 3.1.4 and 3.2, we proposed that an optimal loss must be chosen based on its predictive performance. Asymptotic analysis of the Gibbs predictive, (1), can provide

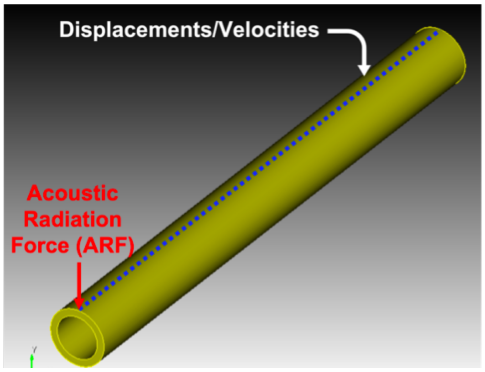


FIGURE 3.1: Waveguide with excitation and measurement locations.

one justification of why our procedure is reasonable. The following theorem shows that if given 3.3.4, the corresponding sequence of Gibbs predictives weakly converges to a certain "likelihood" involving the recovered $\theta*$. An interpretation is that even though we did not explicitly specify a generative model, our predictive criterion (3.9) asymptotically learns a likelihood model that uses the best possible parameter estimate.

Theorem 3.3.5. *Suppose* $\Theta^* = \{\theta^*\}$. *Assume further conditions given by Theorm 3.3.4, so that* $\hat{\rho}_n^W$ *is consistent. Then, for a sequence of Gibbs predictives* $\hat{p}_n^W(y^{new})$,

$$D_{KL}(\mathbb{P}||\hat{p}_n^W) - D_{KL}(\mathbb{P}||\tilde{p}_{\theta^*}) \to 0 \quad \textit{with probability 1,}$$

where $\tilde{p}_{\theta^*}$ *has the density*

$$\tilde{p}_{\theta^*}(y^{new}) = \frac{\exp\{-L(\theta^*; y^{new})\}\lambda(y^{new})}{\int \exp\{-L(\theta^*; y^{new})\} d\lambda(y^{new})}.$$

3.4 Numerical Illustrations

To numerically illustrate the entire workflow in practice based on our proposed method, we consider the inverse problem of estimating the material properties and geometry of a cylindrical waveguide using observations of the wave speed of propagating modes. This type of problem arises in numerous applications, such as non-destructive evaluation of pipes and ultrasound elastography of arteries (Roy, Urban, et al., 2021; Roy and Guddati, 2021).

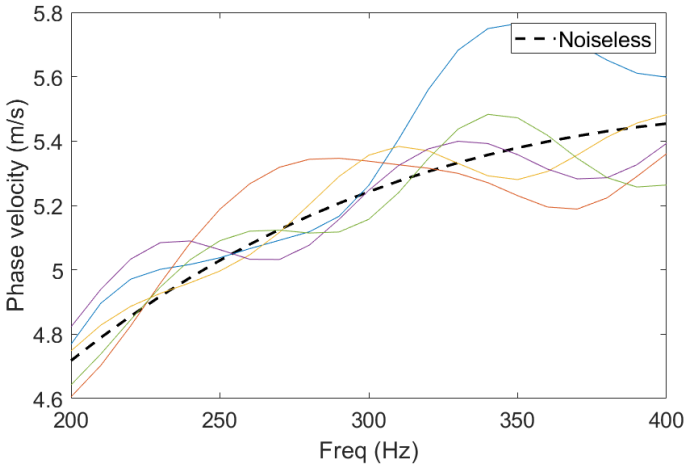


FIGURE 3.2: Simulated noisy dispersion curves for the experiment ($n = 5$).

Figure 3.1 shows a waveguide configuration typically found in ultrasound-based dispersion vibrometry (Bernal et al., 2011; Hugenberg et al., 2021; Capriotti et al., 2022).

The description of the governing equations for elastic and viscoelastic waveguides can be found in Roy, Urban, et al., 2021 and Roy and Guddati, 2021. We focus on wave propagation in isotropic, linear elastic media with an unknown shear modulus. We treat the density and bulk modulus as known, a realistic assumption for nearly incompressible materials like soft tissues. Since we deal with cylindrical waveguides, the geometry is completely defined by their radius and thickness. In summary, our problem is to estimate with uncertainty the shear modulus, thickness, and radius of the waveguide, given noisy waveguide observations.

We now describe the data acquisition and noise modeling for the problem at hand. The structure is excited with a localized force, and the resulting displacements or velocities are measured on a line along the length of the waveguide, as shown in Figure 3.1. Let us assume that these observations (i.e., velocities) can be described by the following additive noise model (3.20):

$$y_i = x(\boldsymbol{\theta}) + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} P_e,\ \mathbb{E}\varepsilon_1 = 0. \tag{3.20}$$

Here, $\boldsymbol{\theta}$ is the unknown material parameter plus geometry, and $x(\boldsymbol{\theta})$ is the particle velocity field. To use waveguide models, we convert the spatiotemporal observations to dispersion relations (phase velocities as functions of frequency). This step entails both a 2D Fourier

transform of the data and highly nonlinear peak-finding operations. Readers interested in the details of this step are referred to Bernal et al., 2011.

The nonlinear transformation described maps the spatiotemporal variable $x(\theta)$ to discrete phase velocity curve y. We can formalize this by introducing a new processing operator $\mathscr{G}$ to the additive noise model (3.20):

$$y_i = \mathscr{G}(x(\theta) + \varepsilon_i) =: \mathscr{F}(\theta) + \delta_i(\theta), \tag{3.21}$$

where $\mathscr{F}$ is the waveguide model and i.i.d. errors δ_i are potentially dependent on θ. Even when space-time domain errors ε_i's are Gaussian, specifying a closed-form likelihood for (3.21) remains a challenge, as a formal Taylor expansion shows that $\delta_i(\theta)$ depends nonlinearly on θ and random variables ε_is. Hence, the type of problem considered in this Section provides a prime example for model misspecification and the utility of Gibbs posterior framework.

To emphasize the methodological aspects of the problem, we will use simulated data, instead of *in vivo* experimental data, to study the performance of Gibbs posteriors. To this end, we generated data samples by simulating the physical response through a waveguide model, using fixed ground truth parameter $\theta_0 \in (\mathbb{R}^+)^3$, and evaluated the output of our forward waveguide model $\mathscr{F}$. The first dominant mode is then perturbed by random noise. We used an independent multiplicative noise process for each of the simulated $n = 5$ sample curves, following a log-Gaussian process in the frequency domain (see Figure 3.2).

We hypothesize that prior information for the possible range of θ is available based on physical knowledge, but different notions can exist for measuring the misfit between the inverse solution and the data. We consider two different losses already considered by Roy and Guddati, 2021:

- Squared ℓ^2-norm loss: $L_{ls}(\theta, y) := ||\mathscr{F}(\theta) - y||^2$;
- ℓ^1 error loss: $L_{l1}(\theta, y) := ||\mathscr{F}(\theta) - y||_{\ell^1}$.

Model selection rule proposed in Section 3.1.4 can be used to compare the associated predictive accuracy of the two losses. Furthermore, for a meaningful comparison, we have scaled the losses based on a scale estimate from the data, $\hat{W}_0$, based on the geometric mean of

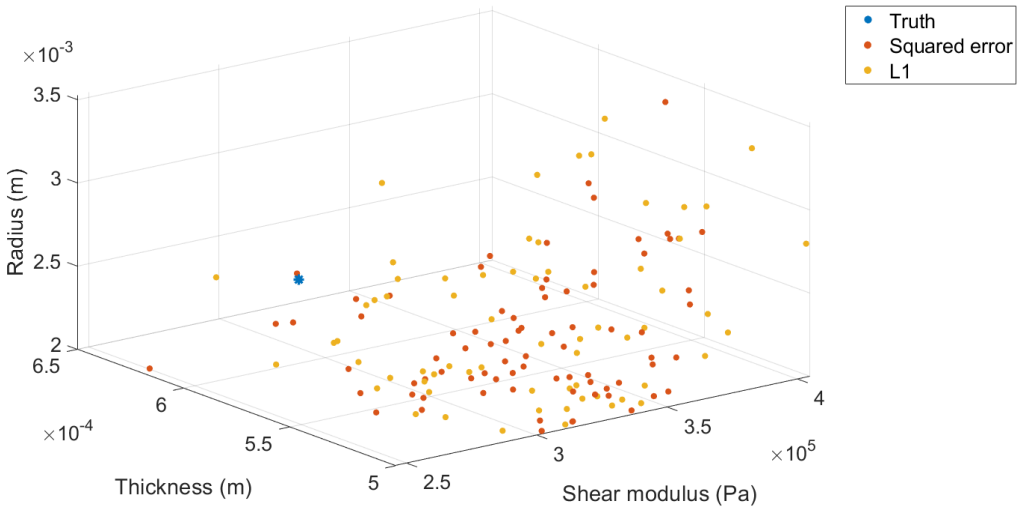


FIGURE 3.3: Joint comparison of Gibbs posterior sample draws for θ using losses L_{ls} and L_{l1}.

Table 3.1: Table of posterior mean squared errors and standard deviations. Each quantity is scaled by the magnitude of parameter $|\theta_{0,i}|$ used for simulation.

	L_{ls}			L_{l1}		
	Modulus	Thickness	Radius	Modulus	Thickness	Radius
$\lvert\theta_{0,i} - \mathbb{E}\theta_i\rvert$	0.13	0.11	0.11	0.14	0.11	0.14
$\sqrt{Var(\theta_i)}$	0.08	0.04	0.13	0.09	0.04	0.15

Table 3.2: Table of calibrated parameters for each model among a grid $[2^{-10}, 2^{-9}, \ldots, 2^{-1}, 1]$. Lower P_{CV} (3.9) implies relative predictive optimality. SE is the approximate standard error of P_{CV}.

	L_{ls}	L_{l1}
W	2^{-5}	2^{-6}
P_{CV}	-23	3
SE	-2	18

sample variances of y_i's across frequencies. This turns out to be ≈ 41.

An independent Beta$(1,3)$ prior was placed on each component of Θ. The prior was translated and re-scaled to have support on the interval dictated by physical knowledge for each parameter. The intervals were between 5 and 95 kPa for shear modulus, between 5 mm and 6.5 mm for arterial wall thickness, and between 2 mm and 4 mm for artery radius, respectively. We used the SMC algorithm of Section 3.2.3 to draw $S = 400$ samples from all four Gibbs posteriors. In successive iterations, particles were mutated by running

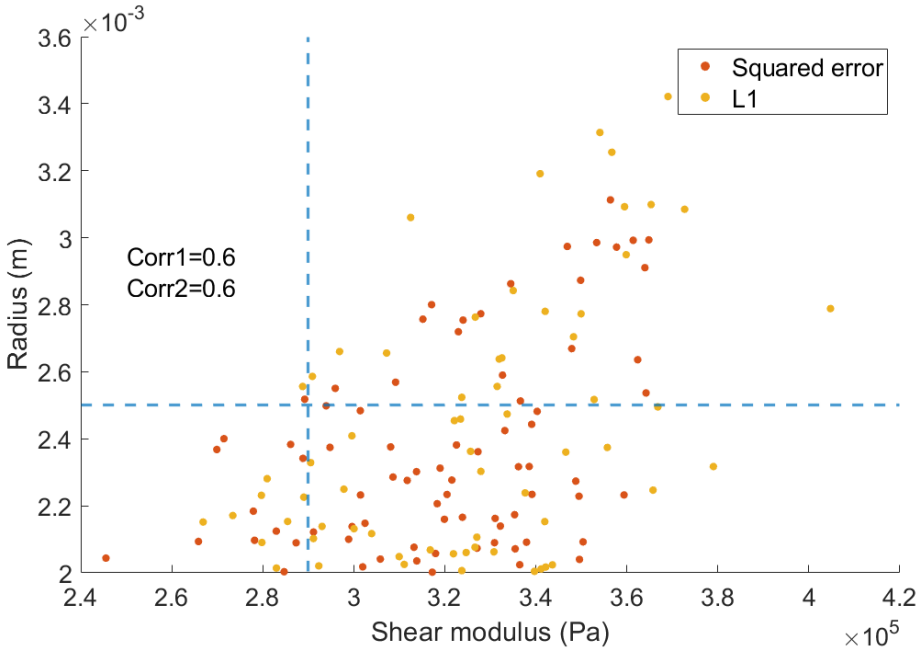


FIGURE 3.4: 2D projection of Figure 3.3 onto radius on shear modulus axes. Large positive correlations are induced *a posteriori* (see legends).

a Metropolis-Hastings algorithm for $K = 5$ steps.

Figure 3.3 plots two distinct sets of posterior sample draws obtained after terminating the algorithm. Apparently, both L_{ls} and L_{l1} lead to posteriors characterized by similar high probability regions. Table 3.1 vindicates this based on mean squared errors of the posterior mean estimators (accuracy) and posterior standard deviations (nominal uncertainty) for each parameter. For both posteriors, W chosen through LOOCV was less than 1. Figure 3.4 shows that both distributions also exhibit a positive linear correlation between shear modulus and radius parameters, a structure learned as opposed to an independent prior specified on the parameters.

Figure 3.5 demonstrates the difficulty of inferring simultaneously material and geometric parameters in waveguides. Here, histograms of posterior sample draws are compared against the prior densities. For the thickness and radius parameters, posterior density estimates closely follow the prior density function. This implies little learning has taken place for these parameters marginally. On the other hand, the marginal density estimate for the

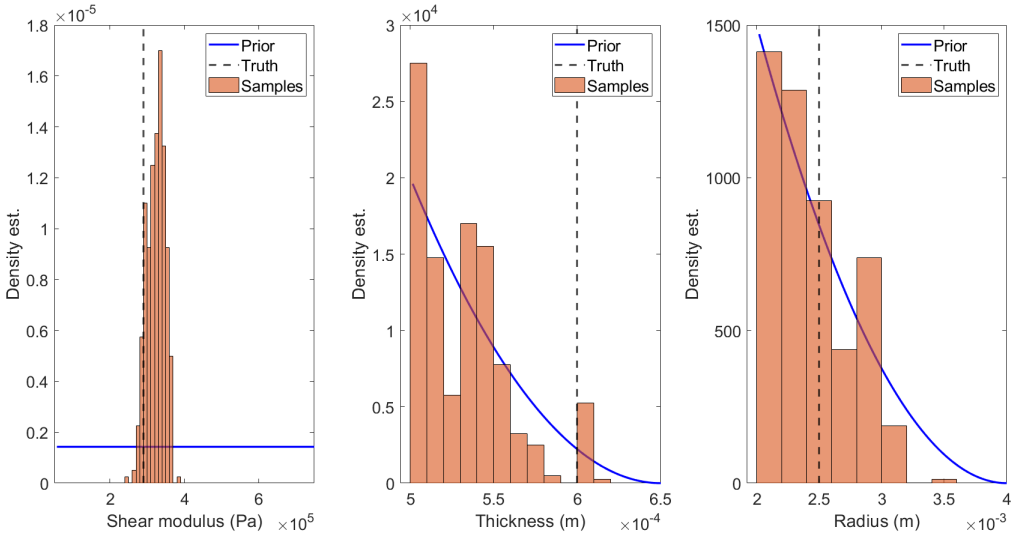


FIGURE 3.5: Marginal comparison of prior density against posterior sample draws for the three model parameters, using loss L_{ls}.

shear modulus parameter is highly concentrated on a neighborhood of the true value used in simulation, in contrast with a uniform prior on the support. While the Figure shows only the posterior for L_{ls}, a similar lack of learning took place for L_{l1}, indicating a lack of information in the data and the need for stronger prior knowledge to distinguish the effect of using different loss functions. Practically, one can resolve the problem by collecting richer data with more propagating modes and different frequency ranges.

Table 3.2 lists the statistics P_{CV} for the two losses being compared. Roy, Urban, et al., 2021 have remarked that empirically, there seems to be little difference in choosing either loss. From the point of view of predictive criterion we suggest, on the other hand, using L_{ls} yields a much smaller statistic P_{CV} on average.

4. Asymptotics of Bayesian Uncertainty Estimation in Random Features Regression

In this Chapter, we tackle a different topic from that of the previous Chapters. As described in the Introduction, the so-called "random features (RF) models" have been studied as toy models that combine theoretical tractability with interesting behaviors that mimic those of deep neural networks (DNNs) trained in practice. We briefly review this model and how common least squares objective training can be viewed as a maximum a posteriori (MAP) estimation of a corresponding probabilistic model. The penalized training loss function then also yields a quantity that can be viewed as the width of the ℓ^2-norm ball quantifying Bayes prediction risk. Comparison of this quantity against the generalization risk on unseen data, studied by Mei and Montanari, 2022, illustrates some unexpected consequences of a "naive" Bayes-like interpretation of RF models for uncertainty quantification. We conjecture some similarities between RF asymptotics and the classical results of Freedman, 1999 for white noise models that pose interesting theoretical challenges for the application of random matrix theory to deep neural networks.

Most of the materials in this Chapter have been published as Baek, Berchuck, and Mukherjee, 2024.

4.1 Background on Random Features Model

Let inputs $\boldsymbol{x}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, n$ be drawn i.i.d. from a uniform measure (denoted τ) on a d-dimensional sphere with respect to the conventional Euclidean norm:

$$\mathbb{S}^{d-1}(\sqrt{d}) := \{\boldsymbol{x} \in \mathbb{R}^d : ||\boldsymbol{x}|| = \sqrt{d}\}. \tag{4.1}$$

Let outputs y_i be generated by the following model:

$$y_i = f_d(\boldsymbol{x}_i) + \varepsilon_i,\ f_d(\boldsymbol{x}) = \beta_{d,0} + \langle \boldsymbol{x}, \boldsymbol{\beta}_d \rangle + f_d^{NL}(\boldsymbol{x}). \tag{4.2}$$

The model is decomposed into a linear component and a nonlinear component, f_d^{NL}. The random error ε_i's are assumed to be i.i.d. with mean zero, variance τ^2, and finite fourth moment. We allow $\tau^2 = 0$, in which case the generating model is *noiseless*. In the analysis,

these quantities will be assumed to obey further conditions, including an appropriate scaling so that all d-dependent quantities are $O_d(1)$.

4.1.1 Training with Ridge Regularization

We focus on learning the optimal *random features model* that best fits the training data. This is a class of functions

$$\mathscr{F} := \left\{ f : f(\boldsymbol{x}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_j \rangle / \sqrt{d}) \right\}, \tag{4.3}$$

which is dependent on N random features, $(\boldsymbol{\theta}_j)_{j=1}^N$, which are drawn i.i.d. from $\mathbb{S}^{d-1}$, and a nonlinear activation function, σ. The training objective solves a regularized least squares problem for the linear coefficients $\boldsymbol{a} \equiv (a_j)_{j=1}^N$:

$$\min_{\boldsymbol{a} \in \mathbb{R}^N} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{x}_i, \theta_j \rangle / \sqrt{d}) \right)^2 + d\psi_{1,d}\psi_{2,d}\lambda ||\boldsymbol{a}||^2 \right\}, \tag{4.4}$$

where $\psi_{1,d} = N/d$, $\psi_{2,d} = n/d$, and $\lambda > 0$. The optimal weights, $\widehat{\boldsymbol{a}} \equiv \widehat{\boldsymbol{a}}(\lambda)$, determine an optimal ridge predictor denoted by $\widehat{f} \equiv f(\cdot; \widehat{\boldsymbol{a}}(\lambda))$. The dependence of the trained predictor on the dataset $(\mathscr{X}, \boldsymbol{y})$ and features $\boldsymbol{\Theta}$ are suppressed in notation.

There exist both practical and theoretical motivations for studying RF regression. On the practical side, RF regression has been suggested by Rahimi and Recht, 2007 as a randomized approximation scheme for training kernel ridge regression (KRR) models for large datasets. On the theoretical side, (4.4) describes a "lazy training" algorithm for a 2-layer neural network with activation function σ. Previous studies have focused on the approximation power of such a function class (Jacot, Gabriel, and Hongler, 2018; Chizat, Oyallon, and Bach, 2019) and the optimization landscape involved in training problems of type (4.4) (Mei, Montanari, and Nguyen, 2018; Mei, Misiakiewicz, and Montanari, 2019).

Given a new input feature $\boldsymbol{x}$, the ridge regularized predictor for the unknown output y has the form

$$\widehat{f}(\boldsymbol{x}) \equiv f(\boldsymbol{x}; \widehat{\boldsymbol{a}}) = \sigma(\boldsymbol{x}^T \boldsymbol{\Theta} / \sqrt{d}) \widehat{\boldsymbol{a}}, \tag{4.5}$$

with the optimal ridge weights $\widehat{\boldsymbol{a}}$ and a resolvent matrix that defines the joint posterior of the weights:

$$\widehat{\boldsymbol{a}} \equiv \widehat{\boldsymbol{a}}(\lambda) := \widehat{\boldsymbol{\Sigma}}(\lambda)\mathscr{Z}^T\boldsymbol{y}/\sqrt{d}, \tag{4.6}$$

$$\widehat{\boldsymbol{\Sigma}}(\lambda) := \left(\mathscr{Z}^T\mathscr{Z} + \psi_{1d}\psi_{2d}\lambda\mathbf{I}_N\right)^{-1}. \tag{4.7}$$

Here, $\mathscr{Z} := \boldsymbol{\sigma}(\mathscr{X}\boldsymbol{\Theta}/\sqrt{d})/\sqrt{d}$ for input design matrix $\mathscr{X} \in \mathbb{R}^{n\times d}$ and output vector $\boldsymbol{y} \equiv (y_i)_{i=1}^n$. Similarly we write $\boldsymbol{\sigma}(\boldsymbol{x}) := \boldsymbol{\sigma}(\boldsymbol{x}^T\boldsymbol{\Theta}/\sqrt{d})$. The $L^2 \equiv L^2(\mathbb{S}^{d-1}(\sqrt{d}); \tau)$ generalization error of predictor $\widehat{f}$ is defined by

$$R_{RF}(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}, \lambda) := \mathbb{E}_{\boldsymbol{x}}||f_d(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x})||^2 \equiv ||f_d - \widehat{f}||^2_{L^2}, \tag{4.8}$$

We emphasize that (4.8) is a random quantity, as it depends on $(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta})$.

4.1.2 RF as Bayesian Model

The objective function (4.4) can be interpreted as a MAP estimation problem for an equivalent Bayesian model. Formally, we adopt a d-dependent weight prior distribution, denoted $p(\boldsymbol{a})$, and also a normal likelihood, denoted $p(\boldsymbol{y}|\mathscr{X}, \boldsymbol{\Theta}, \boldsymbol{a})$, centered around a function in the class (4.3) with variance ϕ^{-1}.

$$\boldsymbol{a} \sim \text{Normal}\left(0, \phi^{-1}\frac{\psi_{1,d}\psi_{2,d}\lambda}{d}\mathbf{I}_N\right)$$

$$\boldsymbol{y} \mid \mathscr{X}, \boldsymbol{\Theta}, \boldsymbol{a} \sim \text{Normal}\left(\boldsymbol{\sigma}(\mathscr{X}\boldsymbol{\Theta}/\sqrt{d})\boldsymbol{a}, \phi^{-1}\mathbf{I}_n\right)$$

The normal likelihood model need not agree with the generating process (4.2), as is often the case for Bayesian deep learning. Technically and inferentially, it can be justified as a coherent update of belief given a squared error loss (Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker, 2016). The joint likelihood of $(\boldsymbol{y}, \boldsymbol{a})$ is defined conditional on both the random features $\boldsymbol{\Theta}$ and an "inverse temperature" ϕ. The choice to condition on $\boldsymbol{\Theta}$ instead of learning them can be unnatural in certain settings and is mainly for the convenience of analysis. However, we do note that the RF model has been used by Crawford et al., 2018 as an approximation of fully Bayesian kernel learning.

The posterior predictive distribution, or Bayesian model average over the posterior of $\boldsymbol{a}$, is defined as

$$p(y \mid \boldsymbol{x}, \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}) := \int_{\mathbb{R}^N} p(y \mid \boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{a})\ p(\boldsymbol{a} \mid \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta})\ d\boldsymbol{a}, \tag{4.9}$$

where the posterior of $\boldsymbol{a}$ is a probability distribution placing higher mass near settings of $\boldsymbol{a}$ minimizing ridge objective (4.4):

$$p(\boldsymbol{a} \mid \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}) \propto \exp\left[-\frac{\phi}{2}\left\{\sum_{i=1}^{n}(y_i - f(\boldsymbol{x};\boldsymbol{a}))^2 + \frac{\psi_{1,d}\psi_{2,d}\lambda}{d}||\boldsymbol{a}||^2\right\}\right]. \tag{4.10}$$

This is a Gaussian measure centered around $\widehat{\boldsymbol{a}}$ (4.6) and covariance matrix $\widehat{\boldsymbol{\Sigma}}(\lambda)/d$. Thus the posterior predictive at new input $\boldsymbol{x}$ is also a normal distribution centered around $\widehat{f}$, with variance

$$s^2(\boldsymbol{x}) \equiv s^2(\boldsymbol{x};\lambda) := \phi^{-1}(1 + \boldsymbol{\sigma}(\boldsymbol{x})^T\widehat{\boldsymbol{\Sigma}}(\lambda)\boldsymbol{\sigma}(\boldsymbol{x})/d). \tag{4.11}$$

We refer to (4.11) as the PPV. This quantity dictates the width of the uncertainty interval centered around (4.4) evaluated at $\boldsymbol{x}$. The expected PPV over $\boldsymbol{x}$ is defined by

$$S^2_{RF}(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}, \lambda) := \mathbb{E}_{\boldsymbol{x}}[s^2(\boldsymbol{x})] \equiv \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{V}[y \mid \boldsymbol{x}, \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}]\right]. \tag{4.12}$$

The expectation over $\boldsymbol{x}$, similarly as in (4.8), yields the radius of the posterior credible ball for a posterior Gaussian process f centered around the optimal predictor $\widehat{f}$. Furthermore, the Gaussian likelihood model simplifies the expected PPV into a decomposition of this radius and the inverse temperature of the posterior:

$$S^2_{RF} = \int ||f - \widehat{f}||^2_{L^2}\ dp(f \mid \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}) + \phi^{-1}, \tag{4.13}$$

where $p(f \mid \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta})$ is the law of Gaussian process f induced by the weight posterior $p(\boldsymbol{a} \mid \boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta})$. Thus, both summands in the display depend on the training features and are random. It is worth contrasting this definition with (4.12); in the next section, we explain in detail the motivation for comparing the two quantities.

The extra quantity introduced in the Bayesian formulation, ϕ, governs how much the resulting posterior should concentrate around the ridge predictor (4.6). Since we are interested in the regime where $N, d \to \infty$, it is reasonable to assume the scale of the likelihood, ϕ, must appropriately decrease with d, similar to how the prior on $\boldsymbol{a}$ is rescaled by $1/d$. Practitioners may adopt some prior distribution on this parameter and perform hierarchical inference. We adopt a simpler, empirical Bayes approach and choose it to maximize the marginal likelihood:

$$\widehat{\phi}^{-1} \equiv \widehat{\phi}^{-1}(\lambda) := \frac{\langle \boldsymbol{y}, \boldsymbol{y} - \widehat{f}(\mathscr{X}) \rangle}{n}. \tag{4.14}$$

This choice coincides with the training set error attained by predictor (4.4), so it will be decreasing as $N \to \infty$. If $N > n$ and $\lambda \to 0^+$, the training error vanishes as the model can perfectly interpolate the training set. We note the precise asymptotics of the training error has been already characterized by Mei and Montanari, 2022 (Section 6).

A fundamental question here is whether R_{RF} and S^2_{RF} have similar asymptotics, as both quantities summarize uncertainty about our prediction in different ways. R_{RF} is the "frequentist's true risk," which requires assumptions about the unknown generative process. S^2_{RF}, on the other hand, is the "Bayesian's risk" that can be actually computed without any model assumptions. Its asymptotics *does* depend on model assumptions, as both the prior and likelihood need not capture the generative process (4.2). In particular, it is desired that it agrees with R_{RF} in some limiting sense. Throughout the rest of this work, we probe the question: Do R_{RF} and S^2_{RF} converge to the same value as $d, n, N \to \infty$?

4.1.3 Previous Works

We have introduced our problem of comparing two different quantities, R_{RF} and S^2_{RF}, and provided their respective interpretations. Such comparison, between the frequentist risk and the variance of the posterior, was done by Freedman, 1999 in a white noise model, where one observes an infinite square-summable sequence with Gaussian noise. The key finding is that the distributions of two corresponding statistics, re-normalized, have different variances. They are in fact radically different distributions in that they are essentially orthogonal. B.

Knapik, A. v. d. Vaart, and J. v. Zanten, 2011 clarified the situation by showing frequentist coverage of credible ball depends heavily on the smoothness of the prior covariance operator. Johnstone, 2010 have extended the results to a sequence of finite but growing length, which is a setup more similar to ours. Our study for the RF model addresses a much simpler question, as the theoretical results in Section 4.2 only address whether two statistics converge to the same limits. Unlike in the work by previous authors, the identity of limits cannot be taken for granted. A key feature driving the different asymptotics of the two quantities is that the Bayesian's prior and the likelihood are "mismatched" with respect to the data generating process; i.e., f_d need not belong to the RF class (4.3). In Section 4.2.3, we demonstrate some distributional properties of R_{RF} and S_{RF}^2 empirically observed in numerical simulations.

Uncertainty quantification in Bayesian learning has recently garnered much attention in the theoretical deep learning literature. We highlight, among many works: Clarté et al., 2023b, characterizing exact asymptotic calibration of a Bayes optimal classifier and a classifier obtained from the generalized approximate message passing algorithm, and Clarté et al., 2023a, demonstrating double descent-like behavior of the Bayes calibration curve in RF models. In particular, quantities R_{RF} and S_{RF}^2 studied by our work can be related, respectively, to the "Bayes-optimal classifier" and "empirical risk classifier" of the latter work (Section 2). The recent work of Guionnet et al., 2023, on the other hand, explicitly addresses the model mismatch in Bayesian inference by deriving the exact, universal asymptotic formula for generalization of a misspecified Bayes estimator in a rank-1 signal detection problem.

4.2 Results

4.2.1 Asymptotic Characterization

We present the main results and illustrate them through numerical simulations. We operate under assumptions identical to those of Mei and Montanari, 2022, stated below. For proofs, we refer the reader to the Appendix of Baek, Berchuck, and Mukherjee, 2024.

Assumption 12. *Define* $\psi_{1,d} = N/d$ *and* $\psi_{2,d} = n/d$. *We assume* $\psi_{1,d} \to \psi_1 < +\infty$ *and* $\psi_{2,d} \to \psi_2 < +\infty$ *as* $d \to \infty$.

Assumption 13. *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be weakly differentiable and* $|\sigma(x)|, |\sigma'(x)| < c_0 e^{c_1|x|}$ *for some* $c_0, c_1 < +\infty$. *For* $G \sim \text{Normal}(0,1)$, *the coefficients:*

$$\mu_0 = \mathbb{E}[\sigma(G)],\ \mu_1 = \mathbb{E}[G\sigma(G)],\ \mu_*^2 = E[\sigma^2(G)] - (\mu_0^2 + \mu_1^2) \tag{4.15}$$

are assumed to be greater than 0 (ruling out a linear σ*) and finite.*

Assumption 14. *The generating model* (4.2) *satisfies*

$$\beta_{0,d}^2 \to F_0^2 < +\infty,\ ||\boldsymbol{\beta}_d||^2 \to F_1^2 < +\infty; \tag{4.16}$$

furthermore, $f_{NL,d}(\cdot)$ *is a centered Gaussian process on* $\mathbb{S}^{d-1}(\sqrt{d})$, *whose covariance function has the form:*

$$\mathbb{E}[f_{NL,d}(\boldsymbol{x}_1) f_{NL,d}(\boldsymbol{x}_2)] = \Sigma_d(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d). \tag{4.17}$$

The kernel Σ_d *satisfies*

$$\mathbb{E}[\Sigma_d(x_{(1)}/\sqrt{d})] = 0,\ \mathbb{E}[x_{(1)}\Sigma_d(x_{(1)}/\sqrt{d})] = 0 \quad \textit{and} \quad \Sigma_d(1) \to F_*^2 < +\infty, \tag{4.18}$$

where $x_{(1)}$ *is the first entry of* $\boldsymbol{x} \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$. *The signal-to-noise ratio (SNR) of the model is defined by*

$$\rho = \frac{F_1^2}{F_*^2 + \tau^2}. \tag{4.19}$$

In this "linear proportional" asymptotic regime, we derive an asymptotic formula for the expected PPV akin to that of Mei and Montanari, 2022 for the generalization error. Analysis shows Definition 1 of Mei and Montanari, 2022, characterizing the asymptotics of R_{RF}, can be straightforwardly applied to also characterize the asymptotics of S_{RF}^2.

Proposition 1. *Denote by* $\mathbb{C}_+$ *the upper half complex plane:* $\{a + b\boldsymbol{i} : \boldsymbol{i} = \sqrt{-1},\ b > 0\}$. *Let functions* $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ *be uniquely defined by the conditions: on* $\mathbb{C}_+$, $\nu_1(\xi), \nu_2(\xi)$ *are analytic and uniquely satisfy the equations*

$$\nu_1 = \psi_1\left(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2}\right)^{-1}, \tag{4.20}$$

$$\nu_2 = \psi_2\left(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2}\right)^{-1}, \tag{4.21}$$

when $|\nu_1(\xi)| \leq \psi_1/\mathrm{Im}(\xi)$ *and* $|\nu_2(\xi)| \leq \psi_2/\mathrm{Im}(\xi)$*, with* $\mathrm{Im}(\xi) > C$ *for sufficiently large constant* C*. Define*

$$\chi = \nu_1(i\sqrt{\psi_1\psi_2\lambda}/\mu_*)\nu_2(i\sqrt{\psi_1\psi_2\lambda}/\mu_*),\ \zeta = \frac{\mu_1}{\mu_*}. \tag{4.22}$$

Under Assumptions 12-14 and for $\lambda > 0$*,*

$$S^2_{RF}(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}, \lambda) \xrightarrow{P} \mathscr{S}^2 = \frac{F_1^2}{1-\chi\zeta^2} + F_*^2 + \tau^2 \text{ as } d,n,N \to \infty.$$

Note that $\mathscr{S}^2$ depends on $(\psi_1, \psi_2, \lambda)$ only through function χ. The following facts follow from this fact and the asymptotic characterization of function χ when $\lambda \to 0^+$.

Proposition 2. *The following holds for the map* $(\psi_1, \psi_2, \lambda) \mapsto \mathscr{S}^2$ *defined in Proposition 1:*

1. *It is non-decreasing in* λ*.*
2. $\lim_{\lambda\to 0^+} \mathscr{S}^2 < +\infty$ *when* $\psi_1 = \psi_2$*.*

Item 2, Proposition 2 deserves special attention, as it seems to suggest that there is no "double descent" in the asymptotics of S^2 when using an improper prior. This fact does *not* possess an operational meaning at any finite $N = n$, due to the ordering of the limits $N, n \to \infty$ and $\lambda \to 0^+$. In fact, one may expect that the distribution of the least singular value of the relevant matrix $\mathscr{Z}$ causes numerical instability for small λ when $N = n$. In Section 4.2.3, this hypothesis is empirically validated through simulations. Subtly, the theoretical prediction does accurately describe the numerical simulations for small choices of λ. On the other hand, the asymptotic prediction for the frequentist risk does diverge when $\lambda \to 0^+$ and $\psi_1 = \psi_2$.

4.2.2 Comparison with Generalization Error

The asymptotics of (4.8), the "frequentist" risk for our comparison, was characterized by Mei and Montanari, 2022 through a theorem similar to Proposition 1. Our main interest lies in comparing the risk against the Bayesian variance in two limiting cases:

1. Highly overparameterized models where $\psi_1 \to \infty$. The number of parameters grows faster than any constant multiple of the number of samples.

2. Large sample models where $\psi_2 \to \infty$. The number of samples grows faster than any constant multiple of the number of parameters.

The first regime has been studied as the relevant regime in modern deep learning. The second regime is closer to the classical regime of asymptotic statistics where only the number of samples n diverges. Below, we re-state the asymptotic formulae for R_{RF} of Mei and Montanari, 2022 in these special cases, which admits simplifications relative to when both ψ_1, ψ_2 are finite.

Proposition 3. (Theorem 4-5, Mei and Montanari, 2022) *Under the notation of Proposition 1, define a function*

$$\omega \equiv \omega(\zeta, \psi, \bar{\lambda}) = -\frac{(\psi\zeta^2 - \zeta^2 - 1) + \sqrt{\psi\zeta^2 - \zeta^2 - 1}}{2(\bar{\lambda}\psi + 1)}.$$

In the highly overparameterized regime, where $\psi_1 \to \infty$, asymptotic risk is given as

$$\mathscr{R}_{wide}(\rho, \zeta, \psi_2, \bar{\lambda}) = \lim_{\psi_1 \to \infty} \lim_{d \to \infty} \mathbb{E} R_{RF}(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}, \lambda) = \frac{(F_1^2 + F_*^2 + \tau^2)(\psi_2\rho + \omega_2^2)}{(1+\rho)(\psi_2 - 2\omega_2\psi_2 + \omega_2^2\psi_2 - \omega_2^2)} + F_*^2$$

with $\omega_2 = \omega(\zeta, \psi_2, \lambda/\mu_*^2)$.

In the large sample regime, where $\psi_2 \to \infty$, asymptotic risk is given as

$$\mathscr{R}_{lsamp}(\zeta, \psi_1, \bar{\lambda}) = \lim_{\psi_2 \to \infty} \lim_{d \to \infty} \mathbb{E} R_{RF}(\boldsymbol{y}, \mathscr{X}, \boldsymbol{\Theta}, \lambda) = \frac{F_1^2(\psi_1\zeta^2 + \omega_1^2)}{\zeta^2(\psi_1 - 2\omega_1\psi_1 + \omega_1^2\psi_1 - \omega_1^2)} + F_*^2$$

with $\omega_1 = \omega(\zeta, \psi_1, \lambda/\mu_*^2)$.

A simple question, mentioned in Section 4.1.2, was whether the two quantities agree. It turns out that in the second regime, at least, the two formulae converge to the same limit, which is the main content of the next Proposition. In the first regime, whether the limits agree is determined by the signal-to-noise ratio ρ. Mei and Montanari, 2022 showed that if ρ is larger than a certain critical threshold ρ_*, which depends on (ψ_2, ζ), the optimal regularization is achieved by $\lambda \to 0^+$, whereas if ρ is smaller than ρ_*, there exists an optimal regularization λ_* bounded away from 0. This phase transition also determines the agreement of the risk and the PPV in the limit.

Proposition 4. (Proposition 5.2,Mei and Montanari, 2022+α) *Define quantities*

$$\rho_*(\zeta,\psi_2) = \frac{\omega_{0,2}^2 - \omega_{0,2}}{(1-\psi_2)\omega_{0,2} + \psi_2},$$

$$\omega_{0,2} = \omega(\zeta,\psi_2,0)$$

under the notation of Proposition 3.

1. *If* $\rho < \rho_*$*, then* $\min_{\bar{\lambda}\geq 0}\mathscr{R}_{wide}(\rho,\zeta,\psi_2,\bar{\lambda})$*, is attained at some* $\lambda^{opt} > 0$*:*

$$\lambda^{opt} := \arg\min_{\bar{\lambda}\geq 0}\mathscr{R}_{wide}(\rho,\zeta,\psi_2,\bar{\lambda}) = \frac{\zeta^2\psi_2 - \zeta^2\omega_*\psi_2 + \zeta^2\omega_* + \omega_* - \omega_*^2}{(\omega_*^2 - \omega_*)\psi_2},$$

$$\omega_* := \omega(\sqrt{\rho},\psi_2,0).$$

   Furthermore, $\mathscr{R}_{wide}(\rho,\zeta,\psi_2,\lambda^{opt}) = \lim_{\psi_1\to\infty}\mathscr{S}^2(\psi_1,\psi_2,\lambda^{opt}) - \tau^2$.

   If, on the other hand, $\rho > \rho_*$*, then* $\arg\min_{\bar{\lambda}\geq 0}\mathscr{R}_{wide}(\rho,\zeta,\psi_2,\bar{\lambda}) = 0$*. Furthermore,* $\mathscr{R}_{wide}(\rho,\zeta,\psi_2,0) < \lim_{\psi_1\to\infty}\mathscr{S}^2(\psi_1,\psi_2,0) - \tau^2$.

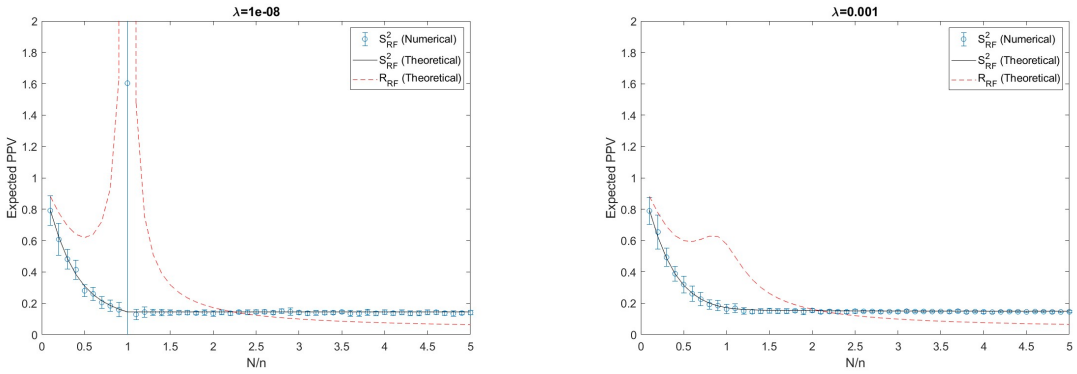
2. $\arg\min_{\bar{\lambda}\geq 0}\mathscr{R}_{lsamp}(\zeta,\psi_1,\bar{\lambda}) = 0$ *and* $\mathscr{R}_{lsamp}(\zeta,\psi_1,0) = \lim_{\psi_2\to\infty}\mathscr{S}^2(\psi_1,\psi_2,0) - \tau^2$.

Note the subtraction of the noise level τ^2 from $\mathscr{S}^2$. This is because S_{RF}^2 is computed based on the training error in Proposition 1, which includes both the approximation error and the variance of data, whereas R_{RF} is computed based only on the approximation error in the test set.

4.2.3 Numerical Simulations

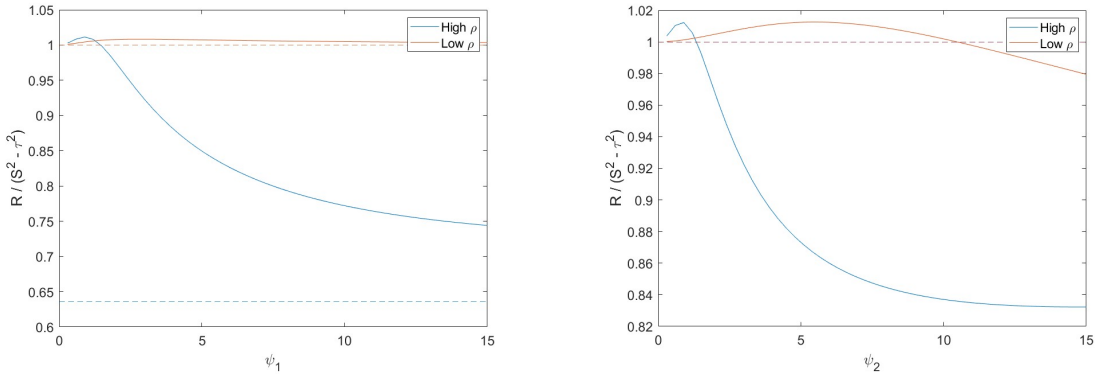
In this section, we first compare evaluations of the asymptotic formulae for R_{RF} and S_{RF}^2 for varying (ψ_1,ψ_2,λ). We highlight their difference for finite (ψ_1,ψ_2) at the optimal choice of λ for the MAP risk. We also present numerical simulations, which both validate the formulae (they concentrate fast) and empirically exhibit further interesting distributional properties (suggesting the need for second-order asymptotics).

Figure 4.1 shows the dramatic mismatch between R_{RF} and S_{RF}^2 in the low-noise regime. It turns out the conservativeness of the credible ball persists for the choice of λ that depends on (ψ_1,ψ_2). The situation becomes more delicate in the high-noise regime because there exists a phase transition in the optimal choice of λ that depends on (4.19), which decreases



(a) $\lambda =$ 1e-8 (b) $\lambda =$ 1e-3

FIGURE 4.1: Comparison of asymptotic formula and 20 instances of S^2_{RF} (4.12). Data are generated via noiseless linear model $y = \langle x, \beta \rangle$ ($||\beta|| = 1$, $\rho = \infty$). Activation is ReLU: $\sigma(x) = \max\{0, x\}$. d and n are fixed to 100 and 300, respectively. The asymptotic formula for R_{RF} (4.8) is plotted for comparison (red, dashed).



(a) $\psi_2 = 3$ (b) $\psi_1 = 3$

FIGURE 4.2: Ratio of $\mathscr{R}(\lambda^{opt})$ to $\mathscr{S}^2(\lambda^{opt}) - \tau^2$ as a function of ψ_1 (4.2a) and of ψ_2 (4.2b). In each plot, ψ_2 and ψ_1 are respectively fixed to 3, while $F_1 = 1$, $F_* = 0$, and $\rho = 1/\tau^2$ for noise variance $\tau^2 \in \{.2, 5\}$.

with the noise variance τ^2. Figures 4.2.4.2a and 4.2b compare the two curves, $\mathscr{R}$ and $\mathscr{S}$, for the "optimally tuned" λ at which the best possible frequentist risk is attained, for a fixed pair of (ψ_1, ψ_2). In a low-noise task, with $\rho = 5$, the ratio of the frequentist risk of the posterior mean predictor to the width of the posterior predictive credible ball is less than 1 for a wide range of ψ_1. The situation is more nuanced and possibly more favorable for the Bayesian case in the high-noise task with $\rho = 1/5$.

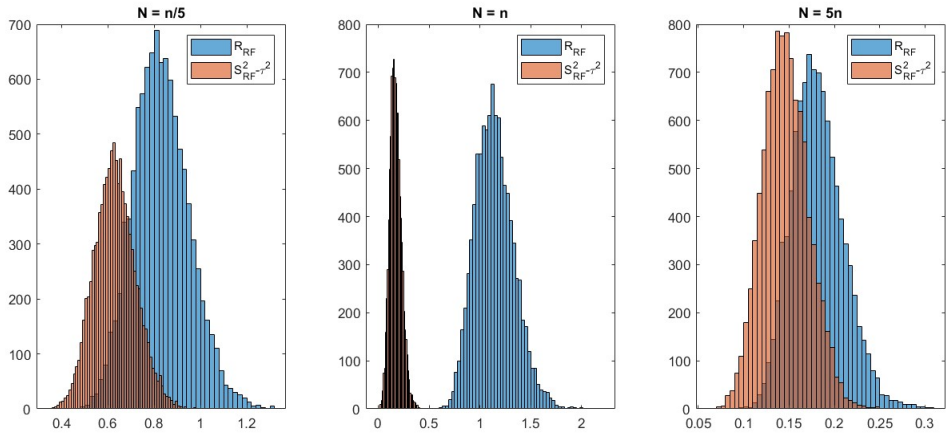


FIGURE 4.3: Histograms of 1e+4 draws of R_{RF} and $S^2_{RF} - \tau^2$ under low-noise linear model $y = \langle \boldsymbol{x}, \boldsymbol{\beta} \rangle + \tau^2$ with $\tau^2 = 1/5$.

While Figure 4.1 validates good concentration properties of R_{RF} and S^2_{RF} with respect to their asymptotic formulae, we may want to investigate the rate of fluctuations for these quantities. Figure 4.3 suggests an interesting phenomenon: both quantities appear Gaussian, but are nearly orthogonal precisely near the "interpolation boundary" where R_{RF} exhibits double descent (2nd subplot). Our empirical results should be compared with the results of Freedman, 1999 and Johnstone, 2010. The Gaussianity of R_{RF} and S^2_{RF}, if true, strengthens the agreement between the expected width S^2_{RF} and expected risk R_{RF} immediately transfers to the agreement between the frequentist coverage of $(1-\alpha)-\%$ Bayes credible ball around the mean and the nominal coverage. Again, such claims are generally true in finite-dimensional settings, but not true for high-dimensional settings. For instance, Freedman, 1999 showed that asymptotically, the posterior variance fluctuates like a Gaussian random variable with a variance strictly smaller than that of the frequentist risk. Extracting such information is possible only if we study *second-order information* of R_{RF} and S^2_{RF}; in particular, suggests a Gaussian central limit for these quantities. A rigorous proof of a central limit theorem for R_{RF} and S^2_{RF} goes beyond the scope of this work. Nevertheless, we conjecture the following for the fluctuations of R_{RF} and S^2_{RF}:

1. Both $d(R_{RF} - \mathbb{E}R_{RF})$ and $d(S^2_{RF} - \mathbb{E}S^2_{RF})$ weakly converge to Gaussian distribution with

appropriate variances. The faster convergence rate of d rather than $\sqrt{d}$ is known to be common in central limit theorems for linear spectral statistics (Lytova and Pastur, 2009).

2. When $\psi_1 \leq \psi_2$, asymptotic variance of S_{RF}^2 is smaller than R_{RF}. When $\psi_1 = \psi_2$, the two distributions are nearly orthogonal. For large enough ψ_1, the asymptotic variances are of the same order.

These conjectures are of independent interest and pose interesting theoretical challenges. We must note that second-order asymptotics has received less attention in the deep learning community. Only recently did Li, Xie, and Q. Wang, 2021 study the asymptotic normality of prediction risk achieved by a min-norm least squares interpolator. Their result relies on a central limit theorem for linear statistics of eigenvalues of large sample covariance matrices, established by Bai and Silverstein, 2004. Many central limit theorems in random matrix theory seem insufficient to handle kernel learning or learning with random features.

5. Conclusions

Below, we summarize interesting future directions suggested by the work in Chapters 2-3 and Chapter 4, respectively.

5.1 Bayesian Inverse Problems and Model Misspecification

The results in Chapter 2 greatly extends our understanding of frequentist coverage for Bayes posteriors when the prior is induced by a Gaussian process. There remain two limitations. First, one may want to extend the methods to cover non-Gaussian priors that are specifically tailored to the problem at hand, such as the high-dimensional prior considered by Nickl, 2020. While we believe the techniques of this work allude to the possibility of such an extension, this definitely goes beyond the scope of our current work. Second, our results are not *adaptive*, in that it is assumed the Gaussian prior covariance operator and the N-dependent scale sequence is adequately chosen with respect to the Sobolev smoothness of the estimated parameter θ_0. To offer a more reassuring guarantee for practitioners, one would want an adaptive Bayes prior which yields near-optimal rate and conservative or exact asymptotic coverage without correctly specifying *a priori* the smoothness parameter. Near-optimal contraction rate results, based on a hierarchical prior for the smoothness parameter, are available for nonparametric regression (Chapter 10, Subhashis Ghosal and Aad W van der Vaart, 2017) and linear inverse problems (Bartek T Knapik et al., 2016). The strategies used in either of these applications are not, however, readily amenable to nonlinear inverse problems. The problem of adaptive coverage is even more challenging, as it is known that it is possible to obtain such results only for certain classes of functions with "tame tails" even in linear inverse problems (Szabó, A. v. d. Vaart, and J. v. Zanten, 2015).

As observed by many authors (Owhadi, Scovel, and Sullivan, 2015; Owhadi and Scovel, 2017), the adoption of a strictly Bayesian approach to inverse problems can lead to various fundamental questions. The complexity of the forward model and the parameter space pose challenges like extreme sensitivity to prior selection, model perturbations, and numerical approximation errors. The methodological concerns of the materials presented in Chap-

ter 3 are a step in the direction of reconciling these challenges with the many benefits of Bayesian inference. Another important challenge is to rigorously formulate the adequacy of a given uncertainty quantification.Kleijn and A. v. d. Vaart, 2012, for example, studied the frequentist coverage and Bernstein-von Mises phenomena for misspecified Bayes posteriors, with mixed results. The variational formulation of this work does not address these issues, but it is an interesting question whether a correct choice of W can yield exact asymptotic coverage and can be algorithmically approximated. A promising approach has been laid out by Syring and Martin, 2019, who proposed the GPC algorithm to choose W that explicitly meets frequentist coverage demands. A technically interesting question is whether rigorous, finite-sample guarantees can be given for various strategies of choosing W, including that of Syring and Martin, 2019 and the cross-validation sampler proposed in Chapter 3. To our understanding, such theoretical guarantees are available only for the SafeBayes algorithm (P. Grünwald, 2012) under restrictive assumptions.

5.2 RF Models, Neural Networks and Probabilistic Inference

While our technical calculations are only applicable to the RF model, we believe the general perspective of comparing frequentist versus Bayesian estimation approaches to the deep learning model can offer useful heuristics explaining empirically observed phenomena in training Bayesian deep learning models.

One suggestion is that the different regimes of asymptotics may explain away the "cold posterior effect", first empirically observed by Wenzel et al., 2020. The authors forcefully suggested that raising the posterior of a deep learning model to an even larger power (i.e., more concentrated) can improve generalization. That the posteriors can be too wide in highly overparameterized, near-noiseless regimes may explain this issue. Even though in this simple RF model setup, the posterior predictive mean is left unchanged from that of the posterior, it is plausible that for more complicated models this need not be so. When that posterior is much wider than the expected risk of the mean, we will mix over too many bad predicting weights, so it will be actually worse to average over the posterior than not.

The technically most interesting but quite challenging direction, suggested by our numerical simulations in Section 3.2.3, is to study second-order fluctuations of random matrix quantities often encountered in deep learning. While at this point a vast amount of literature exists that systematically overcomes the non-standard definitions of quantities like generalization error from the viewpoint of random matrix theory, we are aware of no study as of yet that overcomes the same issues to show a central limit-type result. On the other hand, central limit theorems for linear spectral statistics remain an active area of research in random matrix theory and are generally acknowledged to require more advanced, subtler arguments that are fundamentally different from first-order results (i.e., convergence of empirical spectral distribution). For a start, a separate work on the central limit theorem for linear spectral statistics of kernel matrices is in preparation by the authors.

Another interesting direction is to extend the comparison between frequentist and Bayesian modeling approaches to more complex models of training dynamics. For instance, Adlam and Pennington, 2020 have derived precise asymptotics of generalization error in the neural tangent kernel (NTK) regime, in which the features $\boldsymbol{\Theta}$ are "learned" through a linearized approximation of the gradient flow in training. Mei, Montanari, and Nguyen, 2018; Mei, Misiakiewicz, and Montanari, 2019 have suggested a different asymptotic model, in which $\boldsymbol{\Theta}$ evolve through nonlinear dynamics. An interesting challenge, both technically and conceptually, is to adapt their analysis to the Bayesian setting, where the gradient flow on the space of posterior measures of $\boldsymbol{\Theta}$ is approximated and specific projections are tracked.

Appendix A. Proofs for Chapter 2

A.1 Proof of Theorem 2.2.1

For this proof, we closely follow the proof of Theorem 3.7 (Monard, Nickl, and Paternain, 2021), making some important modifications to account for the fact that ψ need not belong to $R(\mathbb{I}^*_{\theta_0})$. The main idea is to show that the sequence of random variables (2.11) has a Laplace transform converging pointwise to a Laplace transform of a standard normal distribution. First, write by $\Pi_N^{\Theta_\infty(M,\bar{\varepsilon}_N)}(\cdot|\mathscr{D}^N)$ the posterior distribution truncated and renormalized to the set $\Theta_\infty(M,\bar{\varepsilon}_N)$. For any Borel set $A \subset \Theta$, we have

$$\begin{aligned}&\Pi_N(A|\mathscr{D}^N) - \Pi^{\Theta_\infty(M,\bar{\varepsilon}_N)}(A|\mathscr{D}^N)\\ &= \Pi_N(\Theta_\infty(M,\bar{\varepsilon}_N)^c \cap A|\mathscr{D}^N) - \Pi_N(\Theta_\infty(M,\bar{\varepsilon}_N)^c|\mathscr{D}^N)\Pi_N^{\Theta_\infty(M,\bar{\varepsilon}_N)}(A|\mathscr{D}^N),\end{aligned}$$

so the total variation distance between the posterior measure and its restriction to $\Theta_\infty(M,\bar{\varepsilon}_N)$ is bounded above as

$$\sup_{A \text{ Borel}} |\Pi_N(A|\mathscr{D}^N) - \Pi_N^{\Theta_\infty(M,\bar{\varepsilon}_N)}(A|\mathscr{D}^N)| \leq 2\Pi_N(\Theta_\infty(M,\bar{\varepsilon}_N)^c|\mathscr{D}^N).$$

By Assumption 7, the right hand side of the above display is $o^N_{P_{\theta_0}}(1)$. Since the total variation topology is finer than the weak topology of measures, we also have

$$d_{weak}(\Pi_N(\cdot|\mathscr{D}^N), \Pi_N^{\Theta_\infty(M,\bar{\varepsilon}_N)}(\cdot|\mathscr{D}^N)) = o^N_{P_{\theta_0}}(1).$$

Therefore, under the maintained assumptions, it suffices to show the assertion of Theorem 2.2.1 holds for $\Pi_N^{\Theta_\infty(M,\bar{\varepsilon}_N)}(A|\mathscr{D}^N)$.

We now show that the Laplace transform corresponding to the probability law of random variable $s_N^{-1}Z_N$ (2.11), induced by the truncated posterior measure $\Pi_N^{\Theta_\infty(M,\bar{\varepsilon}_N)}(\cdot|\mathscr{D}^N)$:

$$\frac{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)} e^{ts_N^{-1}Z_N - l_N(\theta)}\, d\Pi_N(\theta)}{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)} e^{-l_N(\theta)}\, d\Pi_N(\theta)}, \tag{A.1}$$

converges for each $t \in \mathbb{R}$ to $e^{t^2/2}$ in probability. If this claim is true, the asserted convergence of the Theorem follows by Lemmas 1-2 of the Supplement (Castillo and Rousseau, 2015).

We now consider the following expansion of the numerator of (A.1):

$$\int_{\Theta_\infty(M,\bar{\varepsilon}_N)} e^{ts_N^{-1}Z_N - l_N(\theta) + l_N(\theta^{(t)}) - l_N(\theta^{(t)})}\, d\Pi_N(\theta), \tag{A.2}$$

where $\theta^{(t)} := \theta - ts_N^{-1}\overline{\psi}_N$ for $\overline{\psi}_N$ to be defined. The perturbative analysis now needs a small but important modification from that of the previous authors. For a given $\psi \in L^2(\mathscr{Z})$, we define an N-dependent "approximate posterior scale"

$$s_N \equiv s_{N,\theta_0,\psi} = \sqrt{\langle \psi, (N\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0} + \tau_N^{-2}\Lambda^\alpha)^{-1}(\psi)\rangle_{L^2(\mathscr{Z})}} \tag{A.3}$$

and the N-dependent perturbative vector

$$\overline{\psi}_N \equiv \bar{\psi}_{N,\theta_0} = (N\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0} + \tau_N^{-2}\Lambda^\alpha)^{-1}(\psi). \tag{A.4}$$

Because $\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0}$ is a bounded operator, it is relatively compact with respect to Λ^α (Kato, 2013), so the operator $\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0} + \frac{1}{N}\Lambda^\alpha$ is a self-adjoint positive operator on $D(\Lambda^\alpha)$ for any fixed N. By (2.2), this operator has a bounded inverse on $D(\Lambda^\alpha)$, which we continuously extend to the whole of $L^2(\mathscr{Z})$. Thus, displays (A.3) and (A.4) make sense for any finite N. In particular, $\Lambda^\alpha\overline{\psi}_N$ (A.4) then belongs to $L^2(\mathscr{Z})$. The rationale behind introducing s_N is that for large enough N, it approximates the variance of the underlying posterior which is approximately Gaussian. The key to this proof will be an estimate for the rate of decay of s_N, along with some other quantities that will later arise in the analysis.

Now we proceed with the expansion :

$$
\begin{aligned}
l_N(\theta) - l_N(\theta^{(t)}) &= \sum_{i=1}^{N} \frac{1}{2}(|Y_i - G(\theta)(X_i)|^2 - |Y_i - G(\theta^{(t)})(X_i)|^2) \\
&= \sum_{i=1}^{N} \frac{1}{2}(|G(\theta)(X_i) - G(\theta_0)(X_i) - \varepsilon_i|^2 - |G(\theta^{(t)})(X_i) - G(\theta_0)(X_i) - \varepsilon_i|^2) \\
&= \underbrace{\sum_{i=1}^{N} \left(\frac{|G(\theta) - G(\theta_0)|^2(X_i)}{2} - \frac{|G(\theta^{(t)}) - G(\theta_0)|^2(X_i)}{2} \right)}_{(I)} \\
&\quad - \underbrace{\sum_{i=1}^{N} \varepsilon_i (G(\theta) - G(\theta_0))(X_i)}_{(II)}.
\end{aligned}
$$

We first expand the term (II) as

$$
ts_N^{-1} \sum_{i=1}^{N} \varepsilon_i \mathbb{I}_{\theta_0}(\overline{\psi}_N)(X_i) + \sum_{i=1}^{N} \varepsilon_i R_{\theta_0}(\theta - \theta_0 - ts_N^{-1}\overline{\psi}_N)(X_i) \tag{A.5}
$$

By Lemma 1, the absolute supremum of the second summand in the above display over $\theta \in \Theta_\infty(M, \overline{\varepsilon}_N)$ is $o_{P_{\theta_0}}^N(1)$. Next, we expand the term (I) as

$$
\sum_{i=1}^{N} \left(\frac{|G(\theta) - G(\theta_0)|^2(X_i)}{2} - \frac{\|G(\theta) - G(\theta_0)\|_{L^2(\mathscr{X})}}{2} \right) \tag{A.6}
$$

$$
+ \sum_{i=1}^{N} \left(\frac{|G(\theta^{(t)}) - G(\theta_0)|^2(X_i)}{2} - \frac{\|G(\theta^{(t)}) - G(\theta_0)\|_{L^2(\mathscr{X})}}{2} \right) \tag{A.7}
$$

$$
- t^2 s_N^{-1} \frac{N\|\mathbb{I}_{\theta_0}(\overline{\psi}_N)\|_{L^2(\mathscr{X})}^2}{2} + ts_N^{-1} \langle \theta - \theta_0, N\mathbb{I}_{\theta_0}^* \mathbb{I}_{\theta_0}(\overline{\psi}_N) \rangle_{L^2(\mathscr{X})} \tag{A.8}
$$

$$
+ N\langle \mathbb{I}_{\theta_0}(\theta - \theta_0), R(\theta - \theta_0) \rangle_{L^2(\mathscr{X})} - N\langle \mathbb{I}_{\theta_0}(\theta^{(t)} - \theta_0), R(\theta^{(t)} - \theta_0) \rangle_{L^2(\mathscr{X})} \tag{A.9}
$$

$$
+ \frac{N\|R(\theta - \theta_0)\|_{L^2(\mathscr{X})}^2}{2} - \frac{N\|R(\theta^{(t)} - \theta_0)\|_{L^2(\mathscr{X})}^2}{2} \tag{A.10}
$$

By Lemma 1, the absolute supremum of (A.6) and (A.7) over $\theta \in \Theta_\infty(M, \overline{\varepsilon}_N)$ is $o_{P_{\theta_0}}^N(1)$. The absolute value of (A.9) is bounded above by a $2N\sigma_N\overline{\varepsilon}_N$ by Cauchy-Schwarz inequality and thus is $o_{P_{\theta_0}}^N(1)$ by Assumption 8. Similarly, the absolute value of (A.10) is bounded above

by $N\sigma_N^2$ which is $o^N_{P_{\theta_0}}(1)$. Putting now the above with (II) (A.5), we now conclude that the numerator (A.2) of the Laplace transform can be re-written as

$$\exp\left(t^2 s_N^{-2}\frac{\langle\overline{\psi}_N, N\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0}(\overline{\psi}_N)\rangle_{L^2(\mathscr{X})}}{2}\right)\times$$

$$\int_{\Theta_\infty(M,\bar{\varepsilon}_N)} e^{ts_N^{-1}\langle\theta-\theta_0,\tau_N^{-2}\Lambda^\alpha(\overline{\psi}_N)\rangle_{L^2(\mathscr{X})}-l(\theta^{(t)})}\, d\Pi_N(\theta)\times e^{o^N_{P_{\theta_0}}(1)}.$$

Define a measure Π'_N as the pushforward measure under a shift map $T_{\overline{\psi}_N}:\theta\mapsto\theta+ts_N^{-1}\overline{\psi}_N$. Since $\overline{\psi}_N\in D(\Lambda^\alpha)$ by construction, the Cameron-Martin theorem (2.4) implies

$$e^{ts_N^{-1}\langle\theta,\tau_N^{-2}\Lambda^\alpha(\overline{\psi}_N)\rangle_{L^2(\mathscr{X})}}=\frac{d\Pi'_N}{d\Pi_N}(\theta)\times\exp\left(t^2s_N^{-2}\frac{\|\tau_N^{-1}\Lambda^{\alpha/2}(\overline{\psi}_N)\|^2_{L^2(\mathscr{X})}}{2}\right)$$

Therefore, the Laplace transform (A.1) can now be rewritten as

$$\exp\left(-ts_N^{-1}\langle\theta_0,\Lambda^\alpha(\overline{\psi}_N)\rangle_{L^2(\mathscr{X})}+\frac{t^2}{2}\right)\times\frac{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)}e^{-l_N(\theta^{(t)})}\,d\Pi'_N(\theta)}{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)}e^{-l_N(\theta)}\,d\Pi_N(\theta)}\times e^{o^N_{P_{\theta_0}}(1)}. \tag{A.11}$$

We may define the "bias" comprising the exponent of the first multiplier as

$$b_N\equiv b_N(\psi,\theta_0)=\langle\theta_0,\Lambda^\alpha(N\tau_N^2\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0}+\Lambda^\alpha)^{-1}(\psi)\rangle_{L^2(\mathscr{X})}. \tag{A.12}$$

Lemma 5 implies that under our condition (2.14), we have $b_N/s_N\to 0$. Finally, we now observe that the quotient comprising the second multiplier is equal to

$$\frac{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)}e^{-l_N(\theta^{(t)})}\,d\Pi'_N(\theta)}{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)}e^{-l_N(\theta)}\,d\Pi_N(\theta)}=\frac{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)^{(t)}}e^{-l_N(\theta)}\,d\Pi_N(\theta)}{\int_{\Theta_\infty(M,\bar{\varepsilon}_N)}e^{-l_N(\theta)}\,d\Pi_N(\theta)}, \tag{A.13}$$

where $\Theta_\infty(M,\bar{\varepsilon}_N)^{(t)}:=\Theta_\infty(M,\bar{\varepsilon}_N)-ts_N^{-1}\overline{\psi}_N$. The above display is shown to converge in probability to 1, again by Lemma 1. Putting all the preceding facts together, we conclude that the simplified expression (A.11) converges in probability to $e^{t^2/2}$ for each $t\in\mathbb{R}$. The Laplace transform (A.1) therefore converges pointwise in probability to the Laplace transform of a standard normal distribution, which proves Theorem 2.2.1.

Lemma 1. *Define*

$$\mathscr{F}_1(t) := \sup_{\theta\in\Theta_\infty(M,\overline{\varepsilon}_N)} \left| \sum_{i=1}^N \varepsilon_i R_{\theta_0}(\theta-\theta_0-ts_N^{-1}\overline{\psi}_N)(X_i) \right|;$$

$$\mathscr{F}_2(t) := \sup_{\theta\in\Theta_\infty(M,\overline{\varepsilon}_N)} \left| \sum_{i=1}^N \left(\frac{|G(\theta)-G(\theta_0)|^2(X_i)}{2} - \frac{\|G(\theta)-G(\theta_0)\|^2_{L^2(\mathscr{X})}}{2} \right) \right|.$$

Then $\mathscr{F}_1(t) = o_{P^N_{\theta_0}}(1)$, $\mathscr{F}_2(t) = o_{P^N_{\theta_0}}(1)$ *and the quotient in* (A.13) *is* $e^{o_{P^N_{\theta_0}}(1)}$ *for each* $t\in\mathbb{R}$.

Proof. The proof of the first two assertions follows, almost line by line, the proof of Lemmas 4.3-4 (Monard, Nickl, and Paternain, 2021). We therefore do not reproduce the full proof here, but highlight the slight differences in the estimates that are needed to apply the previous authors' argument. The main idea behind showing $\mathscr{F}_1(t) = o_{P^N_{\theta_0}}(1)$ and $\mathscr{F}_2(t) = o_{P^N_{\theta_0}}(1)$ according to the previous authors' argument is to exploit Assumption 8 to verify the sufficient conditions of Theorem 3.5.4 (Giné and Nickl, 2021). Inspecting the proof of Lemmas 4.3-4 (Monard, Nickl, and Paternain, 2021), we observe that all the arguments of this proof apply identically and uniformly in $\theta\in\Theta_\infty(M,\overline{\varepsilon}_N)$, provided we can show that

$$s_N^{-1}\|\overline{\psi}_N\|_{L^\infty(\mathscr{X})} \ll \overline{\varepsilon}_N. \tag{A.14}$$

The reasoning is as follows. First, if the above statement is true, then for every fixed $t\in\mathbb{R}$ and uniformly in $\theta\in\Theta_\infty(M,\overline{\varepsilon}_N)$, we have

$$\|\theta-\theta_0-ts_N^{-1}\overline{\psi}_N\|_{L^\infty(\mathscr{X})} \le \|\theta-\theta_0\| + ts_N^{-1}\|\overline{\psi}_N\|_{L^\infty(\mathscr{X})} \lesssim \overline{\varepsilon}_N,$$

where the multiplicative constant is uniform in $\theta\in\Theta_\infty(M,\overline{\varepsilon}_N)$ since M is a constant. By Assumption 5, this then also implies that, uniformly in $\theta_0\in\Theta_\infty(M,\overline{\varepsilon}_N)$,

$$R_{\theta_0}(\theta-\theta_0-ts_N^{-1}\overline{\psi}_N) \lesssim \overline{\varepsilon}_N^\rho.$$

Therefore, our choice of the sequence σ_N from Assumption 8 is an upper bound also for $R_{\theta_0}(\theta-\theta_0-ts_N^{-1}\overline{\psi}_N)$. Now turning to the proof of Lemmas 4.3-4 (Monard, Nickl, and Paternain, 2021), one can check line by line that these two consequences, along with Assumption 8, guarantee that

each argument is also applicable to our case. But now the statement (A.14) is indeed true, as can be seen by combining Lemma 6 and the assumed condition (2.13).

Second, we claim that (A.13) converges in probability to 1. We recall the definition of the events $\Theta_\infty(M,\overline{\varepsilon}_N)$, as in (2.9), and $\Theta_\infty(M,\overline{\varepsilon}_N)^{(t)} = \Theta_\infty(M,\overline{\varepsilon}_N) - ts_N^{-1}\overline{\psi}_N$. Combination of Lemma 6 and the assumed condition (2.13) implies that $ts_N^{-1}\|\overline{\psi}_N\|_{L^\infty(\mathscr{Z})} \ll \overline{\varepsilon}_N$ and $t\|\overline{\varepsilon}_N\|_{H^s} \ll \overline{\varepsilon}_N \ll M$, by Assumption 1 and since s can be arbitrarily chosen as long as it belongs to $(\frac{d}{2}, \alpha - \frac{d}{2})$. The assertion is therefore true. □

A.2 Proof of Theorem 2.2.2

By Lemma 7, $s_N = O(1/\sqrt{N})$ and the limit of Ns_N^2 is the same as $\|\varphi\|^2_{L^2(\mathscr{Z})}$ for $\mathbb{I}^*_{\theta_0}\varphi = \psi$. Therefore, in the proof of Theorem 2.2.1, we can replace all appearances of s_N with $\frac{\|\varphi\|_{L^2(\mathscr{Z})}}{\sqrt{N}}$. One can then check the same argument of the proof of Theorem 2.2.1 is applicable line by line, given that one can still check:

(i) b_N (A.12) satisfies $s_N^{-1}b_N = o(1)$;

(ii) the second summand in (A.5) and terms (A.6)-(A.7) are all also $o^N_{P_{\theta_0}}(1)$ uniformly in $\theta \in \Theta_\infty(M,\overline{\varepsilon}_N)$;

(iii) the quotient (A.13) converges to 1 in probability.

All the above facts can be proven as follows. First, by Lemma 7 and the maintained condition (2.15), $s_N^{-1}b_N$ converges to zero. Second, by Lemma 9, we have $s_N^{-1}\|\psi_\lambda\|_{L^\infty(\mathscr{Z})}$ and $s_N^{-1}\|\psi_\lambda\|_{\tilde{H}^s}$ bounded above by a constant multiple of $1/\sqrt{N}$ for any $s \in (\frac{d}{2}, \alpha - \frac{d}{2})$. By Assumptions 3 and 7, the sequence $\overline{\varepsilon}_N \gg 1/\sqrt{N}$. Therefore, as stated in the proof of Lemma 1, the arguments used in the proof of Lemmas 4.3-4 of Monard, Nickl, and Paternain, 2021 are again applicable, and the second item can be proven. Finally, the very same arguments for the proof of the second item also shows the quotient (A.13) converges in probability to 1, as outlined in the proof of Lemma 1.

A.3 Proof of Theorem 2.2.3

For the proof of this Theorem for either of the sufficient conditions (Theorem 2.2.1 or 2.2.2), we first claim that

$$d_{weak}(\mathscr{L}(\widehat{s_N}^{-1}\Psi(\overline{\theta}_N - \theta_0)),\ \text{Normal}(0, t_N/s_N)) \to 0, \tag{A.15}$$

where t_N defines the "approximate standard deviation" of the centering term $\widehat{\Psi}_N$ (2.12):

$$t_N := \sqrt{N}\|\mathbb{I}_{\theta_0}(\overline{\psi}_N)\|_{L^2(\mathscr{X})}. \tag{A.16}$$

Application of central limit theorem shows that $t_N^{-1}\widehat{\Psi}_N$ converges weakly to a standard normal distribution. We observe that t_N/s_N either converges to a finite nonzero constant ≤ 1 or to zero, since $t_N < s_N$ for all finite N by Lemma 5. In the latter case, the convergence statement can be replaced by

$$d_{weak}(\mathscr{L}(\widehat{s_N}^{-1}\Psi(\overline{\theta}_N - \theta_0),\ \delta_{\{0\}}) = o(1),$$

where $\delta_{\{0\}}$ is a Dirac mass at zero. In either case, the limiting measure, which we write as Q, is a tight probability measure on $\mathbb{R}$ and places zero probability mass on the boundary of the interval $[-z_{1-\gamma/2},\ z_{\gamma/2}]$. Now, by the Portmanteau theorem (Resnick, 2019), we deduce the convergence of coverage probability:

$$\begin{aligned} P_{\theta_0}^N(\theta_0 \in I_N) &= P_{\theta_0}^N(\widehat{s_N}^{-1}\Psi(\overline{\theta}_N - \theta_0) \in [-z_{1-\gamma/2,\, z_{1-\gamma/2}}]) \\ &\to Q([-z_{1-\gamma/2},\ z_{1-\gamma/2}]). \end{aligned}$$

The limiting probability belongs to the interval $(1-\gamma, 1]$ whenever $t_N/s_N \to c \in [0,1)$ and is exactly $1-\gamma$ whenever $t_N/s_N \to 1$. By Lemma 8, the latter case holds if and only if $\psi \in R(\mathbb{I}_{\theta_0}^*)$, which proves the assertion of the Theorem for both cases.

The proof of the claim (A.15) follows, almost line by line, the proof of Theorem 3.8 (Monard, Nickl, and Paternain, 2021) as outlined in Section 4.5 therein. We therefore do not reproduce the full proof here, but highlight the slight differences in the estimates that are needed to apply the previous authors' argument. We first claim that all p-th moments

of $s_N^{-1} Z_N$ are stochastically bounded thus:

$$s_N^{-2p} \mathbb{E}[Z_N^{2p} | \mathscr{D}^N] = O^N_{P_{\theta_0}}(1). \tag{A.17}$$

We can then easily recognize that the assertion of Lemma 4.6 (Monard, Nickl, and Paternain, 2021) is a special case of (A.17) when $s_N \asymp 1/\sqrt{N}$ and $p = 1$. Inspection of the proof therein shows that the very same arguments are applicable even for general p and s_N, provided we can show, for each $p \geq 1$,

$$N^2 e^{-N\varepsilon_N^2} \int_\Theta \|\theta - \theta_0\|^4 d\Pi_N(\theta) = O(N^2 \tau_N^4 e^{-N\varepsilon_N^2}) = o(1);$$

$$2^p e^{p\widehat{\Psi}_N} = O_{P^N_{\theta_0}}(1).$$

The first claim is indeed true under the maintained condition on the sequences τ_N and ε_N in Assumptions 2 and 3, respectively. The second assertion is also true due to the following reasoning. Application of central limit theorem to $\widehat{\Psi}_N$ along with the asymptotics of s_N/t_N discussed above, shows that it weakly converges to a centered normal distribution with variance c^2, where $c \in [0,1)$. The case $c = 0$ is equivalent to this measure being a Dirac mass at zero. The second assertion now follows by the continuous mapping theorem. Therefore, we can show that the assertion (A.17) is also true. The remaining proof of the claim (A.15) now follows by a contradiction argument identical to the ensuing proof in Section 4.5 (Monard, Nickl, and Paternain, 2021).

A.4 Asymptotics of Key Quantities

The proof of our main results relies on characterizing the asymptotics of three quantities: s_N (A.3), t_N (A.16) and b_N (A.12), along with the $L^\infty(\mathscr{Z})$-norm of $\overline{\psi}_N$ (A.4). Note that all of these quantities do not depend on θ, so that all inequalities in this Section do not specify the exact multiplicative constants.

For the results to follow, we first define a new linear bounded operator from $L^2(\mathscr{Z})$ to $L^2(\mathscr{X})$

$$\mathbb{K}_{\theta_0,\alpha} := \mathbb{I}_{\theta_0} \Lambda^{-\alpha/2}. \tag{A.18}$$

We then observe that $\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha}$ is a positive, self-adjoint, injective (due to Assumption 6) linear operator on $L^2(\mathscr{X})$ and is trace-class, its trace being bounded above by the operator norm of $\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0}$ times the trace of $\Lambda^{-\alpha}$. Hence it possesses an eigendecomposition consisting of positive eigenvalues and eigenvectors that form orthonormal bases of $L^2(\mathscr{X})$.

Lemma 2. *Let $\psi \in L^2(\mathscr{Z})$. Under Assumption 6, there exists $b > 0$ and $C = C(\psi, \Lambda, \alpha, \beta) > 0$ such that*

$$\langle \psi_\lambda, \psi_\lambda - u\rangle_{L^2(\mathscr{Z})} \leq \frac{1-b}{2}\|\psi_\lambda - u\|^2_{L^2(\mathscr{Z})} + C\|\mathbb{K}_{\theta_0,\alpha}(\psi_\lambda - u)\|^{\eta^{\text{lin}}}_{L^2(\mathscr{X})}, \tag{A.19}$$

where $\psi_\lambda := \Lambda^{-\alpha/2}\psi$.

Proof. Write $\psi_\lambda := \Lambda^{-\alpha/2}(\psi) \in \widetilde{H}^\alpha$. Observe that $\psi_\lambda \notin R((\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})^{1/2})$ since we assumed $\psi \notin R((\mathbb{I}^*_{\theta_0}\mathbb{I}_{\theta_0})^{1/2})$. We show the following variational inequality is true for some $b > 0$ and every $u \in L^2(\mathscr{Z})$: Fix some $b < 1$. On the one hand, whenever

$$\|u\|_{L^2(\mathscr{Z})} \geq \frac{1+b}{1-b}\|\psi_\lambda\|_{L^2(\mathscr{Z})},$$

by triangle and Cauchy-Schwarz inequality, the left hand side of the above display is bounded above by

$$\|\psi_\lambda\|_{L^2(\mathscr{Z})}\|\psi_\lambda - u\|_{L^2(\mathscr{Z})} \leq \frac{1-b}{2}\|\psi_\lambda - u\|^2_{L^2(\mathscr{Z})}.$$

On the other hand, whenever $u \in B_{L^2(\mathscr{Z})}(\frac{1+b}{1-b})$, again using Cauchy-Schwarz inequality and the fact that $\psi_\lambda \in \widetilde{H}^\alpha$, we instead obtain an upper bound

$$\langle \psi_\lambda, \psi_\lambda - u\rangle_{L^2(\mathscr{Z})} \leq \|\Lambda^{\alpha/2}\psi_\lambda\|_{L^2(\mathscr{Z})}\|\Lambda^{-\alpha/2}(\psi_\lambda - u)\|_{L^2(\mathscr{Z})}.$$

By Assumption 6, there exists a constant C, depending only on $\psi_\lambda, \Lambda, \alpha$ and β, such that the right hand side of the above display is bounded above by $C\|\mathbb{K}_{\theta_0,\alpha}(\psi_\lambda - u)\|^{\eta^{\text{lin}}}_{L^2(\mathscr{Z})}$. □

Next, for a fixed vector u, we introduce the distance function

$$d_{1/2}(R) := \inf_{v \in B_{L^2(\mathscr{X})}(R)} \|u - \mathbb{K}^*_{\theta_0,\alpha}v\|. \tag{A.20}$$

It is clear from the definition that there exists a finite R and a vector v such that $d_{1/2}(R) = 0$ whenever $u \in R(\mathbb{K}^*_{\theta_0,\alpha})$.

The following result is proven by Hofmann, 2006 (Lemma 2.5).

Lemma 3. *When $u \notin R(\mathbb{K}^*_{\theta_0,\alpha})$, the function* (A.20) *is positive, continuous, strictly decreasing, and convex. Furthermore, for each R, there exists a unique vector $v^\dagger \in B_{L^2(\mathscr{Z})}(R)$ which is orthogonal to the nullspace of $\mathbb{K}^*_{\theta_0,\alpha}$ and satisfies*

$$d_{1/2}(R) = \|u - (\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})^{1/2} v^\dagger\|_{L^2(\mathscr{Z})}.$$

Next, we derive the asymptotic order of the distance function for a vector ψ_λ as $R \to \infty$, which plays an important role in our analysis.

Lemma 4. *Let $\psi \notin R(\mathbb{I}^*_{\theta_0})$ and let the corresponding distance function* (A.20) *be defined with respect to $u \equiv \psi_\lambda = \Lambda^{-\alpha/2}\psi$. Under Assumption 6, $d_{1/2}(R) \asymp R^{-\frac{\omega}{1-\omega}}$ as $R \to \infty$, with $\omega = \frac{\eta^{\text{lin}}}{2-\eta^{\text{lin}}}$.*

Proof. Since $\psi \notin R(\mathbb{I}^*_{\theta_0})$, $\psi_\lambda \notin R(\mathbb{K}^*_{\theta_0,\alpha})$, due to the injectivity of $\Lambda^{-\alpha/2}$. The assertion now follows from combining Lemma 2 with Theorem 4.5, or rather, Remark 4.6 of Flemming, 2012. □

We are now ready to state the main lemmas needed for the proof of Theorem 2.2.1.

Lemma 5. *Let $\psi \notin R(\mathbb{I}^*_{\theta_0})$. Under Assumption 6, s_N* (A.3)*, t_N* (A.16) *and b_N* (A.12) *satisfy the following.*

(i) We have $t_N < s_N$ for every finite N and

$$s_N \asymp \tau_N \times (\sqrt{N}\tau_N)^{-\omega},\ \omega = \frac{\eta^{\text{lin}}}{2-\eta^{\text{lin}}}.$$

(ii) We have

$$s_N^{-1} b_N \lesssim \tau_N^{-1} \times \begin{cases} (\sqrt{N}\tau_N)^{\omega(1-\frac{\beta}{\alpha})} & \text{if } \alpha \geq \beta, \\ (\sqrt{N}\tau_N)^{-\omega(1-\frac{\alpha}{\beta})} & \text{if } \alpha < \beta. \end{cases}$$

The exponents in the two cases above are non-negative and strictly negative, respectively.

Proof. We freely use notions from the spectral projection theory and functional calculus for self-adjoint operators; see, e.g., Section 2.3 (Heinz Werner Engl, Hanke, and Neubauer, 1996). Define, in particular, the spectral distribution function $F_{\psi_\lambda}(\varepsilon)$ and the spectral projection operator E_t through

$$F^2_{\psi_\lambda}(t) := \|E_t \psi_\lambda\|^2 = \|\mathbf{1}_{(0,t]}(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|^2_{L^2(\mathscr{Z})},$$

where $\mathbf{1}_{(0,t]}$ is the indicator function of the interval $(0,t]$. We will first show that the spectral distance function yields the order of the lower and upper bound for s_N and the upper bound for t_N, then use interpolation to deduce the second assertion about the bias term b_N.

(i) By Lemma 4 and the equivalence relations of Theorem 2 (Flemming, Hofmann, and Mathé, 2011), we obtain the fact

$$F_{\psi_\lambda}(\varepsilon) \asymp \varepsilon^{\frac{\varpi}{2}} \text{ as } \varepsilon \to 0. \tag{A.21}$$

The two quantities s_N and t_N can be rewritten as follows:

$$s_N = \tau_N \|f(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{L^2(\mathscr{Z})},$$

$$t_N = \tau_N \|g(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{L^2(\mathscr{Z})};$$

where $f, g : (0,\infty) \to \mathbb{R}$ are defined as $f(x) = \frac{1}{\sqrt{1+N\tau_N^2 x}}$ and $g(x) = \frac{\sqrt{N\tau_N^2 x}}{(1+N\tau_N^2 x)}$. This representation clearly shows that $t_N < s_N$ for every finite N, proving the first part of the claim. We now observe that f, f^2, g are continuous and bounded above by 1. For f and f^2, by strict monotonicity, there exists some $c > 0$ such that whenever $x \in (0, \frac{c}{N\tau_N^2}]$, we have $f(x) \wedge f^2(x) \geq 1/2$. Hence, we have

$$s_N^2 \geq \tau_N^2 \int_0^{\frac{c}{N\tau_N^2}} f^2(t)\, dF^2_{\psi_\lambda}(t) \geq \frac{\tau_N^2}{4} \int_0^{\frac{c}{N\tau_N^2}} dF_{\psi_\lambda}(t)^2 = \frac{\tau_N^2}{4} F^2\left(\frac{c}{N\tau_N^2}\right). \tag{A.22}$$

By a similar reasoning, we can also show

$$h_N := \|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{L^2(\mathscr{Z})} \geq \frac{1}{2} F\left(\frac{c}{N\tau_N^2}\right), \tag{A.23}$$

which estimate will be needed for the proof later.

To derive the upper bound, following the proof of Theorem 5.5 (Hofmann and Mathé, 2007), we proceed as follows. First, consider the equation

$$\Phi_{1/2}(R) := \frac{d_{1/2}(R)}{R} = \sqrt{\varepsilon}. \tag{A.24}$$

As shown by Hofmann and Mathé, 2007 and Flemming, Hofmann, and Mathé, 2011, $\Phi_{\psi_\lambda} : (0,\infty) \to (0,\infty)$ is a continuous, strictly decreasing function for $\psi_\lambda \notin R(\mathbb{K}^*_{\theta_0,\alpha})$. Therefore, the equation in the above display has a unique solution $\widehat{R} = \widehat{R}(\varepsilon)$ for all ε. Now, observe that there exists some $c' > 0$ such that the mappings $x \mapsto \sqrt{x} f(x)$, $x \mapsto \sqrt{x} g(x)$, and $x \mapsto \sqrt{x} f^2(x)$ are all bounded above by $\frac{c'}{\sqrt{N}\tau_N}$. Choose $\varepsilon = \frac{1}{N\tau_N^2}$, $\widehat{R} = \widehat{R}(N)$ the solution of (A.24) for this choice, and $v^\dagger$ the minimizer of the distance function (A.20) for this choice of $\widehat{R}$. Then, for s_N:

$$\begin{aligned}
\tau_N^{-1} s_N &= \| f(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda \|_{L^2(\mathscr{Z})} \\
&\leq \| f(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})(\psi_\lambda - \mathbb{K}^*_{\theta_0,\alpha} v^\dagger) \|_{L^2(\mathscr{Z})} + \| f(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\mathbb{K}^*_{\theta_0,\alpha} v^\dagger \|_{L^2(\mathscr{Z})} \\
&\leq d_{1/2}(\widehat{R}) + \frac{c'}{\sqrt{N}\tau_N}\widehat{R} \\
&\leq (1+c') d_{1/2}\left(\Phi^{-1}_{1/2}\left(\frac{1}{\sqrt{N}\tau_N} \right) \right).
\end{aligned}$$

By Corollary 1 (Flemming, Hofmann, and Mathé, 2011), for every $\varepsilon \in (0,\infty)$,

$$d_{1/2}(2\Phi^{-1}_{1/2}(\sqrt{\varepsilon})) \leq F_{\psi_\lambda}(\varepsilon) \leq 2 d_{1/2}(\Phi^{-1}_{1/2}(\sqrt{\varepsilon})).$$

Since $(2\varepsilon)^{\frac{\omega}{1-\omega}} \leq 2^{\frac{\omega}{1-\omega}} \varepsilon^{\frac{\omega}{1-\omega}}$, for ε belonging to a sufficiently small neighborhood of zero, the upper bound of the above display can be bounded above by $2^{-\frac{\omega}{1-\omega}}$ times the lower bound due to Lemma 4. Combined with the lower bound (A.22) and the asymptotic estimate (A.21), this shows that $s_N \asymp \tau_N \times (N\tau_N^2)^{-\omega/2}$. The exact same reasoning can be applied to an upper bound of h_N from (A.23), so that $h_N \asymp (N\tau_N^2)^{-\omega/2}$.

(ii) We now turn to the estimate for b_N. For $\theta_0 \in \widetilde{H}^\beta$, we may rewrite, using Cauchy-Schwarz inequality,

$$|b_N| \leq \|\theta_0\|_{\widetilde{H}^\beta} \times \| f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda \|_{\widetilde{H}^{\alpha-\beta}}. \tag{A.25}$$

When $\beta < \alpha$, using the interpolation inequality of Theorem 2.1.1,

$$\|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{\widetilde{H}^{\alpha-\beta}} \leq \|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|^{\frac{\beta}{\alpha}}_{L^2(\mathscr{Z})}\|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|^{\frac{\alpha-\beta}{\alpha}}_{\widetilde{H}^{\alpha}}. \quad \text{(A.26)}$$

Since $f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})$ has an operator norm bounded by 1 for every N and $\psi_\lambda = \Lambda^{-\alpha/2}\psi$, the $\widetilde{H}^{\alpha}$-norm of $f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda$ is bounded above by the $L^2(\mathscr{Z})$-norm of ψ. The L^2-norm of the same vector is exactly h_N in (A.23), which was shown above to be of order $(N\tau_N^2)^{-\omega/2}$. Therefore, combined with the estimate for s_N, we conclude that

$$s_N^{-1}b_N \lesssim \tau_N^{-1} \times (N\tau_N^2)^{\frac{\omega}{2}(1-\frac{\beta}{\alpha})},$$

which proves the second assertion of the Lemma for this case. In the second situation, where $\beta \geq \alpha$, we can again use interpolation and the fact that $\|\cdot\|_{\widetilde{H}^{-\beta}} \leq \|\cdot\|_{\widetilde{H}^{-\alpha}}$ to deduce

$$\|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{\widetilde{H}^{\alpha-\beta}} \leq \|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|^{\frac{\beta-\alpha}{\beta}}_{\widetilde{H}^{-\alpha}}\|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|^{\frac{\alpha}{\beta}}_{L^2(\mathscr{Z})}. \quad \text{(A.27)}$$

Since $\|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{\widetilde{H}^{-\alpha}}$ is equal to $\|\Lambda^{-\alpha/2}f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{L^2(\mathscr{Z})}$, by Assumption 6, the $\widetilde{H}^{-\alpha}$-norm of $f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda$ appearing in the right hand side is bounded above by a constant multiple of the factor

$$\left(\frac{1}{\sqrt{N}\tau_N}\|g(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{L^2(\mathscr{Z})}\right)^{\eta^{\text{lin}}\frac{\beta-\alpha}{\beta}} \asymp \left\{(N\tau_N^2)^{-\frac{1}{2}-\frac{\omega}{2}}\right\}^{\eta^{\text{lin}}\frac{\beta-\alpha}{\beta}}.$$

Then, again using the estimates for h_N and doing some algebra with the exponent, we obtain:

$$s_N^{-1}b_N \lesssim \tau_N^{-1} \times (N\tau_N^2)^{-\frac{\omega}{2}(1-\frac{\alpha}{\beta})},$$

concluding the proof of the second assertion of the Lemma.

□

Lemma 6. *If $\psi \notin R(\mathbb{I}^*_{\theta_0})$, under Assumption 6, we have $s_N^{-1}\|\overline{\psi}_N\|_{L^\infty(\mathscr{Z})} \leq s_N^{-1}\|\overline{\psi}_N\|_{\widetilde{H}^{\alpha'}} \lesssim \tau_N \times (\sqrt{N}\tau_N)^{-\omega(1-\frac{\alpha'}{\alpha})}$ for any $\alpha' \in (\frac{d}{2}, \alpha)$.*

Proof. Let f be defined as in the proof of Lemma 5. For any choice of α' under the assumed condition, $\|\cdot\|_{\widetilde{H}^{\alpha'}}$ dominates $\|\cdot\|_{L^\infty(\mathscr{Z})}$ due to Assumption 1. By interpolating as in Theorem 2.1.1, we can write:

$$\|\overline{\psi}_N\|_{L^\infty(\mathscr{Z})} \leq \|\overline{\psi}_N\|_{\widetilde{H}^{\alpha'}} \leq \|\overline{\psi}_N\|_{L^2(\mathscr{Z})}^{\frac{\alpha-\alpha'}{\alpha}} \|\overline{\psi}_N\|_{\widetilde{H}^{\alpha}}^{\frac{\alpha'}{\alpha}}. \tag{A.28}$$

The $\widetilde{H}^{\alpha}$-norm and $L^2(\mathscr{Z})$-norm of $\overline{\psi}_N$ is equal to τ_N^2 times the term h_N appearing in (A.23) and τ_N^2 times the $\widetilde{H}^{-\alpha}$-norm of the vector $f^2(\mathbb{K}_{\theta_0,\alpha}^* \mathbb{K}_{\theta_0,\alpha})\psi_\lambda$ appearing in the previous proof, respectively. Since $\alpha > \alpha'$, by using the same reasoning as in the previous proof, we obtain the claimed upper bound on $L^\infty(\mathscr{Z})$-norm. □

When ψ does belong to the range of $\mathbb{I}_{\theta_0}^*$ as in Theorem 2.2.2, we obtain quite different asymptotics.

Lemma 7. *If $\psi \in R(\mathbb{I}_{\theta_0}^*)$, then $Ns_N^2 \to \|(\mathbb{I}_{\theta_0}^* \mathbb{I}_{\theta_0})^{-1/2}\psi\|_{L^2(\mathscr{Z})}^2$ and*

$$s_N^{-1} b_N \lesssim \begin{cases} N^{\frac{\alpha-2\beta}{2\alpha}} \tau_N^{-\frac{2\beta}{\alpha}} & \text{if } \alpha \geq \beta, \\ \frac{1}{\sqrt{N}\tau_N^2} & \text{if } \alpha < \beta. \end{cases}$$

Proof. First, we observe that if $\psi \in R(\mathbb{I}_{\theta_0}^*) = R((\mathbb{I}_{\theta_0}^* \mathbb{I}_{\theta_0})^{1/2})$, then $\psi_\lambda = \Lambda^{-\alpha/2}\psi \in R((\mathbb{K}_{\theta_0,\alpha}^* \mathbb{K}_{\theta_0,\alpha})^{1/2})$. For s_N (A.3), we have

$$Ns_N^2 = N\tau_N^2 \int_0^K \frac{1}{1+N\tau_N^2 t^2}\, dF_{\psi_\lambda}^2(t) = \int_0^K \left(\frac{N\tau_N^2 t^2}{1+N\tau_N^2 t^2}\right) \frac{dF_{\psi_\lambda}^2(t)}{t^2},$$

where K stands in for the operator norm of $\mathbb{K}_{\theta_0,\alpha}^* \mathbb{K}_{\theta_0,\alpha}$. We apply the dominated convergence theorem to the spectral distribution function $dF_{\psi_\lambda}^2$, $(1/t)^2$ being an integrable envelope, and conclude the above display converges to $\|(\mathbb{K}_{\theta_0,\alpha}^* \mathbb{K}_{\theta_0,\alpha})^{-1/2}\psi_\lambda\|_{L^2(\mathscr{Z})}^2 = \|(\mathbb{I}_{\theta_0}^* \mathbb{I}_{\theta_0})^{-1/2}\psi\|_{L^2(\mathscr{Z})}^2$ as $N \to \infty$.

For b_N, we return to the Cauchy-Schwarz inequality of (A.25) in the proof of Lemma 5, where it was shown it suffices to bound the term

$$\|f^2(\mathbb{K}_{\theta_0,\alpha}^* \mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{\widetilde{H}^{\alpha-\beta}}$$

for $f(x) = \frac{1}{\sqrt{1+N\tau_N^2 x}}$. When $\alpha \geq \beta$, again using the interpolation inequality (A.26), it suffices to bound the $L^2(\mathscr{Z})$-norm instead, since the $\widetilde{H}^\alpha$-norm is bounded above by the $L^2(\mathscr{Z})$-norm of ψ. Under the maintained assumption, there exists some $w \in B_{L^2(\mathscr{Z})}(R)$, $R > 0$, such that

$$\|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda\|_{L^2(\mathscr{Z})} = \|f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\mathbb{K}^*_{\theta_0,\alpha}w\|_{L^2(\mathscr{Z})} =: \|\tilde{f}(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})w\|_{L^2(\mathscr{Z})},$$

where $\tilde{f}(x) = \frac{\sqrt{x}}{1+N\tau_N^2 x}$. There exists $c' > 0$ such that $\tilde{f}(x) \leq \frac{c'}{N\tau_N^2}$, so the right hand side of the above display can be bounded from above by $\frac{c'R}{N\tau_N^2}$. The claimed bound on $s_N^{-1}b_N$ for $\alpha \geq \beta$ now follows because $s_N \asymp N^{-1/2}$. The bound for the case when $\alpha < \beta \iff \alpha - \beta < 0$ follows simply by bounding the $\widetilde{H}_{\alpha-\beta}$ norm by the $L^2(\mathscr{Z})$-norm. □

Lemma 8. *$s_N/t_N \to 1$ if and only if $\psi \in R(\mathbb{I}^*_{\theta_0})$.*

Proof. We first claim that $s_N/t_N \to 1$ if and only if, for every $c > 0$,

$$\int_0^{\frac{c}{\sqrt{N}\tau_N}} \frac{1}{1+N\tau_N^2 t^2}\, dF^2_{\psi_\lambda}(t) \ll \tau_N^{-2}s_N^2. \tag{A.29}$$

Given the claim, the order of the left hand side can be deduced as follows. Since the function $f(t) = \frac{1}{1+N\tau_N^2 t}$ on the range of integration is bounded below by $1/(1+c)$ and above by 1, we have

$$F^2_{\psi_\lambda}\left(\frac{1}{1+c}\frac{1}{N\tau_N^2}\right) \leq \int_0^{\frac{c}{\sqrt{N}\tau_N}} f(t)\, dF^2_{\psi_\lambda}(t) \leq F^2_{\psi_\lambda}\left(\frac{1}{N\tau_N^2}\right).$$

If $\psi_\lambda \notin R(\mathbb{K}^*_{\theta_0,\alpha})$, we also have $F^2_{\psi_\lambda}(\frac{1}{N\tau_N^2}) \asymp \tau_N^{-2}s_N^2$, as shown in the proof of Lemma 5. On the other hand, if $\psi_\lambda \in R(\mathbb{K}^*_{\theta_0,\alpha})$, then $F^2_{\psi_\lambda}(\frac{1}{N\tau_N^2}) \ll \frac{1}{N\tau_N^2}$ by Lemma 3 (Flemming, Hofmann, and Mathé, 2011), but Ns_N^2 converges to a finite nonzero limit by Lemma 7. Therefore, $s_N/t_N \to 1$ if and only if $\psi_\lambda \in R(\mathbb{K}^*_{\theta_0,\alpha})$, which is if and only if $\psi \in R(\mathbb{I}^*_{\theta_0})$ by the injectivity of $\Lambda^{-\alpha/2}$.

The proof of the equivalence between the statement $s_N/t_N \to 1$ and (A.29) simply follows the argument used by B. Knapik, A. v. d. Vaart, and J. v. Zanten, 2011 in the proof of Theorem 5.4 therein. We reproduce their proof here for completeness. Write $s_N^2 = A_N + B_N$ where

$$A_N := \int_{\frac{c}{\sqrt{N}\tau_N}}^K f(t)\, dF^2_{\psi_\lambda}(t),\ B_N := \int_0^{\frac{c}{\sqrt{N}\tau_N}} f(t)\, dF^2_{\psi_\lambda}(t),$$

where K stands in for the operator norm of $\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha}$. Similarly, write $t_N^2 = C_N + D_N$. Our goal then becomes equivalent to showing $t_N/s_N \to 1$ if and only if $B_N/A_N \to 0$.

We observe that t_N can be written as

$$t_N = \int_0^K \frac{N\tau_N^2 t^2}{1+N\tau_N^2 t^2} \times f(t)\, dF^2_{\psi_\lambda}(t),$$

so that

$$\frac{D_N}{B_N} \leq \frac{c}{1+c} \leq \frac{C_N}{A_N} \leq 1.$$

To show that $t_N/s_N \to 1$ implies $B_N/A_N \to 0$, we write

$$\frac{t_N^2}{s_N^2} = \frac{C_N/A_N + (D_N/B_N)(B_N/A_N)}{1+B_N/A_N} \leq \frac{1+(B_N/A_N)c/(1+c)}{1+B_N/A_N}.$$

Fix $r \in (0,1)$. The mapping $x \mapsto \frac{1+rx}{1+x}$ is strictly decreasing from 1 (at $x=0$) to $r<1$ (as $x \to \infty$). It follows that the right hand side of the above display converges to 1 if and only if $B_N/A_N \to 0$, with its limit infimum strictly smaller than 1 otherwise, which proves the implication $t_N/s_N \to 1 \implies B_N/A_N \to 0$.

To show the other implication: $B_N/A_N \to 0 \implies t_N/s_N \to 1$, we write

$$\frac{t_N^2}{s_N^2} \geq \frac{C_N}{A_N+B_N} = \frac{C_N/A_N}{1+B_N/A_N} \geq \frac{c/(1+c)}{1+B_N/A_N}.$$

If $B_N/A_N \to 0$, then the limit infimum of the left hand side is at least $c/(1+c)$. Since this is true for every $c>0$, we conclude that $t_N/s_N \to 1$. □

Lemma 9. *If $\psi \in R(\mathbb{I}^*_{\theta_0})$, then $s_N^{-1}\|\overline{\psi}_N\|_{L^\infty(\mathscr{X})} \leq s_N^{-1}\|\overline{\psi}_N\|_{\widetilde{H}^{\alpha'}} \lesssim N^{-1/2}$ for any $\alpha' \in (\frac{d}{2}, \alpha)$.*

Proof. The reasoning is very similar to the proof of Lemma 6: again, we use the interpolation as in (A.28), and observe that it suffices to bound τ_N^2 times the $L^2(\mathscr{X})$-norm of the vector $f^2(\mathbb{K}^*_{\theta_0,\alpha}\mathbb{K}_{\theta_0,\alpha})\psi_\lambda$, or h_N as appearing in (A.23). By using the result from the proof of Lemma 7, we obtain

$$s_N^{-1}\|\overline{\psi}_N\|_{\widetilde{H}^{\alpha'}} \lesssim \tau_N^2 \times \sqrt{N} \times \frac{1}{N\tau_N^2} = \frac{1}{\sqrt{N}}.$$

□

A.5 Proof of Theorem 2.3.1

The proof of the Theorem essentially follows the proof of Theorems 4-5 (Giordano and Nickl, 2020). We therefore will only highlight the major differences in choosing a different scaling rate τ_N. Prior to the main proof, we will first establish a few quantitative estimates useful for all following proofs and also verify Assumptions 3-4, provided $\theta_0 \in \widetilde{H}^\beta$, $\beta > 1 + \frac{d}{2}$ (only the first assertion of Assumption 4 is needed for this Theorem).

Analytic Estimates

By the mean value theorem, for every $\theta, \theta_0 \in B_{\widetilde{H}^{\beta'}}(M)$, there exists some $\tilde{\theta} : \mathscr{X} \to \mathbb{R}$ such that

$$a_\theta - a_{\theta_0} = e^\theta - e^{\theta_0} = e^{\tilde{\theta}}(\theta - \theta_0).$$

Recall that the Gaussian prior is supported on $\widetilde{H}^s$ whenever $s < \alpha - \frac{d}{2}$. If $\theta, \theta_0 \in B_{\widetilde{H}^s}(M)$, then they are uniformly bounded in $\|\cdot\|_{C^1}$ by Assumption 1 and Hölder embedding of Sobolev spaces. A collection of elements $\tilde{\theta}$ for which the above display is met is then also uniformly bounded in $\|\cdot\|_{C^1}$. This implies that

$$\|a_\theta - a_{\theta_0}\|_{C^1} \asymp \|\theta - \theta_0\|_{C^1}, \tag{A.30}$$

where the multiplicative constant depends only on M. Furthermore, the above holds when the norm in C^1 is replaced with L^p for any $p \in [1, \infty]$.

Equation 2.16 specifies a strongly elliptic divergence form operator L_a, $a \in \mathscr{C}$. Recall that $\beta' = s \wedge \beta > 1 + \frac{d}{2}$ by the maintained assumption on the choice of s and $\beta \geq s$. Whenever $a \in B_{H^{\beta'}}(M)$, Proposition A.5.2 (Nickl, 2023) then yields an explicit expression for the constant $C \equiv C(\beta', d, \mathscr{X}, K_{min}, f, g, M) > 0$ such that

$$\sup_{\|\theta\|_{\widetilde{H}^{\beta'}} \leq M} \|G(\theta)\|_{H^{\beta'+1}} \leq C < \infty. \tag{A.31}$$

The above display also provides a uniform bound on $\|\cdot\|_{C^\xi}$ for any $\xi < \beta' - \frac{d}{2}$ by Hölder embedding of Sobolev spaces.

Verifying the Assumptions

Assumption 3 is proven in Proposition 2.1.3 (Nickl, 2023) for a constant U that is uniform in M.

Assumption 4 : First, by Proposition 2.1.3 (Nickl, 2023) for $\kappa_1 = 1$, there exists some $L(M)$ such that

$$\|G(\theta) - G(\theta_0)\|_{L^2} \leq L(M)\|\theta - \theta_0\|_{(H^1)^*}$$

whenever $\theta, \theta_0 \in B_{\widetilde{H}^s}(M)$. where $(H^1)^*$ is the topological dual space of H^1. Since $\widetilde{H}^1 = H_0^1$, we have $\widetilde{H}^{-1} = H^{-1} = (H_0^1)^*$. Since the norms $\|\cdot\|_{(H^1)^*}$ and $\|\cdot\|_{H_0^1}$ are equivalent on H_0^1, we conclude the first inequality of Assumption 4 is also satisfied for $\kappa_1 = 1$.

We now show the second inequality is met for a certain exponent η that depends on d. By Hölder embedding of H^1, we have

$$\|G(\theta) - G(\theta_0)\|_{L^\infty} \leq \|G(\theta) - G(\theta_0)\|_{H^\xi}$$

for any $\frac{d}{2} < \xi$. If $d = 1$, then ξ can be exactly chosen to be 1. Furthermore, $G(\theta) = G(\theta_0)$ on $\partial\mathscr{X}$, so by Poincaré inequality the right hand side is equivalent to $\|\nabla G(\theta) - \nabla G(\theta_0)\|_{L^2}$. By Lemma 1 (Richter, 1981),

$$\|\nabla G(\theta) - \nabla G(\theta_0)\|_{L^2} \leq K_{min}^{-1}\|\nabla G(\theta_0)\|_{L^2}\|\theta - \theta_0\|_{L^\infty}.$$

We conclude that when $d = 1$, the second claim of Assumption 4 is met with exponent $\eta = 1$. Next, consider the case when $d > 1$. Choosing $\xi > \frac{d}{2}$ and using Sobolev interpolation inequality:

$$\|G(\theta) - G(\theta_0)\|_{L^\infty} \leq \|G(\theta) - G(\theta_0)\|_{H^\xi} \leq \|G(\theta) - G(\theta_0)\|_{H^1}^{1-\frac{\xi-1}{\beta'}}\|G(\theta) - G(\theta_0)\|_{H^{\beta'+1}}^{\frac{\xi-1}{\beta'}}.$$

We conclude that when $d > 1$, the claim is met with any exponent in $(0, \frac{\beta'-d-1}{\beta'})$.

Main Proof

Given the above Assumptions, we now prove the "pushforward posterior" of $G(\theta)$ contracts towards a neighborhood of $G(\theta_0)$:

$$P_{\theta_0}^N(\Pi_N(\theta : \|G(\theta) - G(\theta_0)\|_{L^2} > M\varepsilon_N,\ \|\theta\|_{\widetilde{H}^s} > M|\mathscr{D}^N) \geq e^{-LN\varepsilon_N^2}) \to 0. \tag{A.32}$$

Throughout this Section, we write $d_G(\theta,\theta_0) := \|G(\theta) - G(\theta_0)\|$. Given this claim, the results of Theorem 2.3.1 will be derived on a local stability estimate for the inverse of G. For the proof, it will suffice to show the following three statements hold for some sequence of measurable subsets $\Theta_N \subset \Theta$, for all sufficiently large N and every A, B, C:

$$\Pi_N(\theta : \|G(\theta) - G(\theta_0)\|_{L^2} \leq M\varepsilon_N) \geq e^{-AN\varepsilon_N^2}; \tag{A.33}$$

$$\Pi_N(\Theta_N^c) \leq e^{-BN\varepsilon_N^2}; \tag{A.34}$$

$$\log \mathscr{N}(\Theta_N, d_H, \varepsilon_N) \leq CN\varepsilon_N^2. \tag{A.35}$$

Note, in particular, that the proof of condition (A.33) implies Assumption 3 for every choice of sequence ε_N maintained in this proof. Combination of Theorem 14 and Lemma 22 (Giordano and Nickl, 2020) then implies the claimed contraction theorem (A.32).

We choose the sequence of subsets Θ_N for a given prior Π_N to be

$$\Theta_N := \{B_{\widetilde{H}^{-1}}(M\varepsilon_N;\theta_0) + B_{\widetilde{H}^{\alpha\wedge\beta}}(M\sqrt{N}\varepsilon_N\tau_N;\theta_0)\} \cap B_{\widetilde{H}^s}(M). \tag{A.36}$$

We verify the conditions one by one. First, by Assumption 4, we have

$$\Pi_N(\theta : d_G(\theta,\theta_0) \leq M\varepsilon_N) \geq \Pi_N\left(\theta : \|\theta - \theta_0\|_{\widetilde{H}^{-1}} \leq \frac{\varepsilon_N}{L(M')},\ \|\theta - \theta_0\|_{\widetilde{H}^s} \leq M\right) \tag{A.37}$$

with $M' = M + \|\theta_0\|$, through triangle inequality. The event in the above display is a symmetric set shifted by a fixed vector $\theta_0 \in \widetilde{H}^\beta$. If $\alpha \leq \beta$, θ_0 belongs to the Cameron-Martin space $\widetilde{H}^\alpha$, so by Corollary 2.6.18 (Giné and Nickl, 2021) and Gaussian correlation inequality, we obtain a lower bound

$$e^{-\tau_N^{-2}\|\theta_0\|_{\widetilde{H}^\alpha}}\Pi_N\left(\theta : \|\theta\|_{(H^1)^*} \leq \frac{\varepsilon_N}{L(M')}\right)\Pi_N(\theta : \|\theta\|_{\widetilde{H}^s} \leq M).$$

The two prior probabilities involved in the above display can then be bounded from below as in the proof of Lemma 16 (Giordano and Nickl, 2020) by a term $e^{-AN\varepsilon_N^2}$ in this case, for some $A > 0$. By the choice of $\tau_N = \frac{1}{\sqrt{N}\varepsilon_N}$ in the $\alpha \leq \beta$ case, condition (A.33) follows. On the other hand, suppose $\alpha > \beta$. Define a finite-dimensional projection of θ_0 by

$$\theta_0^{(N)} := E_D \theta_0,$$

where $D = N^{\frac{d}{2\beta+2+d}}$ and $E_D = \mathbf{1}_{(0,D]}(-\Delta^{\alpha/2})$ is the spectral projection operator. By the well-known Weyl asymptotics for the Laplacian (Weyl, 1911) and the choice of ε_N for this case, we obtain

$$\|\theta_0^{(N)} - \theta_0\|_{\widetilde{H}^{-1}} \lesssim (D)^{-\frac{2}{d}\frac{\beta+1}{2\beta+2+d}} \|\theta_0\|_{\widetilde{H}^\beta} \lesssim \varepsilon_N^2.$$

Similar reasoning also yields the estimates

$$\|\theta_0^{(N)} - \theta_0\|_{\widetilde{H}^s} \asymp \|\theta_0^{(N)} - \theta_0\|_{\widetilde{H}^s} \lesssim N^{-\frac{\beta-s}{2\beta+2+d}} \|\theta_0\|_{\widetilde{H}^\beta} \lesssim 1$$

and

$$\|\theta_0^{(N)}\|_{\widetilde{H}^\alpha} \lesssim N^{\frac{\alpha-\beta}{2\beta+2+d}}.$$

By triangle inequality and by the fact that $\widetilde{H}^s$-norm dominates the $\widetilde{H}^{-1}$-norm,

$$\Pi_N\left(B_{(H^\kappa)^*}\left(\frac{\varepsilon_N}{L(M')};\theta_0\right) \cap B_{\widetilde{H}^s}(M;\theta_0)\right) \geq \Pi_N\left(B_{(H^\kappa)^*}\left(\frac{\varepsilon_N}{2L(M')};\theta_0^{(N)}\right) \cap B_{\widetilde{H}^s}(M;\theta_0^{(N)})\right),$$

and since $\theta_0^{(N)} \in \widetilde{H}^\alpha$, we again apply Corollary 2.6.18 (Giné and Nickl, 2021) to obtain a lower bound

$$e^{-\tau_N^{-2}\|\theta_0^{(N)}\|_{\widetilde{H}^\alpha}} \Pi_N\left(\theta : \|\theta\|_{(H^1)^*} \leq \frac{\varepsilon_N}{2L(M')}\right) \Pi_N(\theta : \|\theta\|_{\widetilde{H}^s} \leq M).$$

The choice of τ_N in the $\alpha > \beta$ case implies the first multiplier is bounded below by $e^{-AN\varepsilon_N^2}$ in this case. The two other probabilities can be bounded from below using the same arguments as in the $\alpha \leq \beta$ case, but with a different choice of τ_N and ε_N. Proceeding again as in the proof of Lemma 16 (Giordano and Nickl, 2020), condition (A.33) again follows.

The other two conditions (A.34) and (A.35) can be shown exactly as in the proof of Lemmas 17 and 18 (Giordano and Nickl, 2020), only with a different choice of sequences τ_N and ε_N.

Conditional Stability Estimates

We can now derive Theorem 2.3.1 from (A.32) based on conditional stability estimates as first proposed by Vollmer, 2013. We essentially retrace the steps taken by Giordano and Nickl, 2020. First, under Assumption 9 and whenever $\theta = \theta_0$ on $\partial\mathscr{X}$, Proposition 2.1.7 (Nickl, 2023) shows the existence of a constant $C = C(M, K_{min}, k_1, k_2, \mathscr{X}, d, f, g)$ such that

$$\|a_\theta - a_{\theta_0}\|_{L^2} \leq C\|a_{\theta_0}\|_{C^1}\|G(\theta) - G(\theta_0)\|_{H^2};$$

note that $\|\cdot\|_{C^1}$ of a_{θ_0} is bounded above by a constant multiple of $\|\theta_0\|_{\widetilde{H}^\beta}$. Interpolating between Sobolev scales and using the uniform bound (A.31), we can deduce

$$\|a_\theta - a_{\theta_0}\|_{L^2(\mathscr{Z})} \lesssim \|G(\theta) - G(\theta_0)\|_{L^2}^{\frac{\beta'-1}{\beta'+1}}$$

where the multiplicative constant is uniform in $\theta \in B_{\widetilde{H}^s}(M)$. Using (A.30), the left hand side is bounded from below as a constant multiple of $\|\theta - \theta_0\|_{L^2}$. On the other hand, Corollary 1 (Richter, 1981) shows that whenever $\theta = \theta_0$ on $\partial\mathscr{X}$,

$$\|a_\theta - a_{\theta_0}\|_{L^\infty(\mathscr{Z})} \leq C(k_1, k_2, u_{\theta_0})\|a_{\theta_0}\|_{C^1(\mathscr{Z})}\|G(\theta) - G(\theta_0)\|_{C^2(\mathscr{X})}.$$

By Hölder embedding of H^ξ, $\xi > 2 + \frac{d}{2}$ and again by interpolating between Sobolev scales, we can deduce

$$\|a_\theta - a_{\theta_0}\|_{L^\infty(\mathscr{Z})} \lesssim \|G(\theta) - G(\theta_0)\|_{L^\infty(\mathscr{Z})}^{\frac{\beta'-1-d/2}{\beta'+1}}$$

up to a multiplicative constant uniform in $\theta \in B_{\widetilde{H}^s}(M)$, completing the proof of the Theorem.

A.6 Proof of Theorem 2.3.2

A.6.1 Case 1 : Range Condition Does Not Hold

We apply Theorem 2.2.3 in the case when $\psi \notin R(\mathbb{I}^*_{\theta_0})$. To do so, we verify the remaining Assumptions 5-8.

Assumption 5 is verified by Theorem 3.3.2 (Nickl, 2023). In particular, we obtain an explicit expression for $\mathbb{I}_{\theta_0}$, its adjoint and the order of the remainder term (2.8):

$$\mathbb{I}_{\theta_0}(h) = -L_{a_0}^{-1}[\nabla \cdot (e^{\theta_0} h \nabla u_{a_0})], h \in L^2(\mathscr{Z}); \tag{A.38}$$

$$\mathbb{I}_{\theta_0}^*(g) = e^{\theta_0} \nabla u_{a_0} \cdot \nabla L_{a_0}^{-1}[g], g \in L^2; \tag{A.39}$$

$$\|R_{\theta_0}(h)\|_{L^2} = O(\|h\|_{L^\infty(\mathscr{Z})}^2), h \in L^2(\mathscr{Z}). \tag{A.40}$$

Assumption 6 : Since $\alpha > 1+d$, every $h \in \widetilde{H}^\alpha$ also belongs to $\widetilde{H}^1 = H_0^1$, the domain of $(-\Delta)^{1/2}$. For every such h, by Theorem 5 (Nickl and Paternain, 2022), we have

$$C\|\mathbb{I}_{\theta_0}(h)\|_{H^2(\mathscr{X})} \geq \|h\|_{L^2},$$

with constant $C = C(k_1, k_2, a_0, \mathscr{Z})$, where μ, c_0 are constants in Assumption 9. Interpolating between Sobolev spaces,

$$\|\mathbb{I}_{\theta_0}(h)\|_{H^2(\mathscr{X})} \leq \|\mathbb{I}_{\theta_0}(h)\|_{L^2}^{\frac{\alpha\wedge\beta-1}{\alpha\wedge\beta+1}} \|\mathbb{I}_{\theta_0}(h)\|_{H^{\alpha\wedge\beta+1}(\mathscr{X})}^{\frac{2}{\alpha\wedge\beta+1}}.$$

Now $e^{\theta_0} h \in \widetilde{H}^{\alpha\wedge\beta}$, which is continuously embedded in $H^{\alpha\wedge\beta}$. Since this element belongs to a domain of higher powers of the Dirichlet Laplacian, repeated application of Poincaré inequality yields the estimate

$$\|\nabla \cdot (e^{\theta_0} h \nabla u_{f_0})\|_{H^{\alpha\wedge\beta+1}} \leq CM\|e^{\theta_0}\|_{H^{\alpha\wedge\beta}} \|u_{a_0}\|_{H^{\alpha\wedge\beta}}, C = C(d, \alpha, \beta, \mathscr{X}).$$

We conclude that the exponent η^{lin} can be chosen to be $\frac{\alpha\wedge\beta-1}{\alpha\wedge\beta+1}$.

Assumption 7 is the content of Theorem 2.3.1.

Assumption 8 : Display (A.40), combined with Lemma 6 and the maintained condition (2.13), implies the sequence σ_N can be chosen to be $\overline{\varepsilon}_N^2$, up to some multiplicative constant independent of θ and N. Next, the entropy number involved in the Dudley-type integral (2.10) is bounded above by the entropy number for the set $B_{\widetilde{H}^s}(M)$ due to the definition (2.9). In turn, this entropy number can be bounded from above as

$$\log \mathscr{N}(B_{\widetilde{H}^s}(M), \|\cdot\|_{L^\infty}, \delta) \leq \log \mathscr{N}(B_{H^s}(M), \|\cdot\|_{C(\bar{\Omega})}, \delta) \lesssim \left(\frac{1}{\delta}\right)^{\frac{d}{s}}, \beta > 0.$$

Here, $C(\bar{\Omega})$ is the space of functions in $C(\Omega)$ that can be continuously extended to $\bar{\Omega}$. The first inequality is a striaghtforward consequence of Assumption 1; the second inequality follows from display (8), Section 4.10.3 (Triebel, 1978). Since s is assumed to be greater than $\frac{d}{2}$, we can now deduce that all integrals involved in this Assumption are finite and characterize their orders. Verifying the Assumption now reduces to verifying the following statement, since $\eta \leq 1$:

$$\max\left\{N\overline{\varepsilon}_N^{2+\eta}, \sqrt{N}\overline{\varepsilon}_N^{(2-\frac{d}{s})\eta}, \sqrt{\log N}\overline{\varepsilon}_N^{(1-\frac{2d}{s})\eta}\right\} \to 0.$$

For this to hold, it is necessary that $s > 2d$, so that $\alpha > \frac{5}{2}d$. Now it is clear it suffices to show only the two terms in the above display converge to zero. (2.21). In the proof of Theorem 2.3.1, the exponent η was derived to be exactly 1 when $d = 1$ and any exponent in $(0, \frac{\beta' - d/2 - 1}{\beta'}$ when $d > 1$. Some algebra with the exponent shows that in the first case, it suffices to have

$$\frac{2\alpha - 2}{2\alpha + 3}\frac{\alpha - 2}{\alpha + 1/2} > \frac{1}{3}$$

while when $d > 1$, it suffices to have

$$\frac{2\alpha - 1 - d}{2\alpha + 2 + d}\frac{\alpha - 1 - d}{\alpha + 1 - d/2} > \frac{\alpha - d/2}{3\alpha - 2d + 1} \vee \frac{(\alpha - d/2)^2}{4(\alpha - d)(\alpha + 1 - d)}.$$

Now to apply Theorem 2.2.1, we need to verify two additional conditions (2.13) and (2.14). From the derivation of η^{lin} above, we know ω can be chosen to be $\frac{\alpha - 1}{\alpha + 3}$. Condition (2.23) then implies condition (2.14) whenever $\alpha \leq \beta$, as can be seen through some tedious algebra. Now to show condition (2.13), we use the fact that our choice of τ_N and ε_N for the case $\alpha \leq \beta$ satisfies the relation

$$\sqrt{N}\tau_N = \frac{1}{\varepsilon_N},\ \tau_N = \frac{1}{\sqrt{N}\varepsilon_N}.$$

Consider first the case $d = 1$. From the verification of Assumption 8, we can choose $\overline{\varepsilon}_N$ to be $N^{-\frac{1}{3}-\varepsilon}$ for an arbitrarily small ε that may also depend on α, β when $d = 1$. It can be

now shown with some tedious algebra that if

$$\alpha > \frac{3(\alpha-1)(\alpha+1)}{6(\alpha-1)(\alpha+1)-(\alpha+3)(4\alpha+3)},$$

then there exists $a \in (\frac{d}{2}, 1-\frac{d}{2\alpha})$ satisfying the condition (2.13).

In the case when $d > 1$, a similar reasoning with even more tedious algebra shows that it suffices to have the condition

$$\alpha > \frac{d}{2}\left\{1-\left(\frac{\alpha+3}{3\alpha-2d+1}\frac{2\alpha-3d/2}{\alpha-1}\right)\vee \left(\frac{(\alpha+3)(2\alpha^3+(2-3d)\alpha^2+d(7d/2-4)\alpha-d^2(7d/2-5)/2)}{4(\alpha-d)(\alpha+1-d)(\alpha-1)(\alpha+1)}\right)\right\}^{-1},$$

completing the proof.

A.6.2 Case 2 : Range Condition Holds

In the case when $\psi \in R(\mathbb{I}^*_{\theta_0})$, we note that (2.21) still is necessary to verify Assumption 8 and that Assumption 6 is not necessary. Provided this, (2.15) can be now verified (with some algebra) for the chosen τ_N (2.18) in the case $\alpha \leq \beta$. The case $\beta \leq \alpha < 2\beta$ is seen to be trivial because $\tau_N \lesssim 1$. We have assumed $\alpha \leq \beta + \frac{d}{2}$, which implies $\alpha < 2\beta$; hence, the assertion now follows. Hence, Theorem 2.2.2 applies.

Bibliography

Adlam, Ben and Jeffrey Pennington (2020). "The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization". In: *International Conference on Machine Learning*. PMLR, pp. 74–84.

Baek, Youngsoo, Wilkins Aquino, and Sayan Mukherjee (2023). "Generalized Bayes approach to inverse problems with model misspecification". In: *Inverse Problems* 39.10, p. 105011.

Baek, Youngsoo, Samuel Berchuck, and Sayan Mukherjee (2024). "Asymptotics of Bayesian Uncertainty Estimation in Random Features Regression". In: vol. 36.

Bai, ZD and Jack W Silverstein (2004). "CLT for linear spectral statistics of large-dimensional sample covariance matrices". In: *The Annals of Probability* 32.1A, pp. 553–605.

Barmherzig, David A and Ju Sun (2022). "Towards practical holographic coherent diffraction imaging via maximum likelihood estimation". In: *Optics Express* 30.5, pp. 6886–6906.

Belkin, Mikhail (2021). "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation". In: *Acta Numerica* 30, pp. 203–248.

Bernal, M et al. (2011). "Material property estimation for tubes and arteries using ultrasound radiation force and analysis of propagating modes". In: *J. Acoust. Soc. Am.* 129.3, pp. 1344–1354.

Bernardo, J M and A F M Smith (2009). *Bayesian theory*. West Sussex: John Wiley & Sons.

Beskos, A et al. (2015). "Sequential Monte Carlo methods for Bayesian elliptic inverse problems". In: *Statistics and Computing* 25.4, pp. 727–737.

Bhattacharya, A, D Pati, and Y Yang (2019). "Bayesian fractional posteriors". In: *Ann. Statist.* 47.1, pp. 39–66.

Bissantz, Nicolai, Thorsten Hohage, and Axel Munk (2004). "Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise". In: *Inverse Problems* 20.6, p. 1773.

Bissiri, P G, C C Holmes, and S G Walker (2016). "A general framework for updating belief distributions". In: *J. R. Statist. Soc. B* 78.5, pp. 1103–1130.

Bissiri, Pier Giovanni, Chris C Holmes, and Stephen G Walker (2016). "A general framework for updating belief distributions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 1103–1130.

Bogachev, Vladimir Igorevich (1998). *Gaussian measures*. 62. American Mathematical Soc.

Bonito, Andrea et al. (2017). "Diffusion coefficients estimation for elliptic partial differential equations". In: *SIAM Journal on Mathematical Analysis* 49.2, pp. 1570–1592.

Bretagnolle, Jean and Catherine Huber (1979). “Estimation des densités: risque minimax”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 47, pp. 119–137.

Capriotti, M et al. (2022). “The influence of acoustic radiation force beam shape and location on wave spectral content for arterial dispersion ultrasound vibrometry”. In: *Phys. Med. Biol.* 67.13, p. 135002.

Castillo, Ismaël and Richard Nickl (2013). “Nonparametric Bernstein–von Mises Theorems in Gaussian White Noise”. In: *The Annals of Statistics* 41.4, pp. 1999–2028.

Castillo, Ismaël and Judith Rousseau (2015). “A Bernstein–von Mises Theorem for Smooth Functionals in Semiparametric Models”. In: *The Annals of Statistics* 43.6, pp. 2353–2383.

Chatterjee, S and P Diaconis (2018). “The sample size required in importance sampling”. In: *Ann. Appl. Probab.* 28.2, pp. 1099–1135.

Chizat, Lenaic, Edouard Oyallon, and Francis Bach (2019). “On lazy training in differentiable programming”. In: *Advances in neural information processing systems* 32.

Chopin, N and O Papaspiliopoulos (2020). *An Introduction to Sequential Monte Carlo*. New York, NY: Springer.

Clarté, Lucas et al. (2023a). “On double-descent in uncertainty quantification in overparametrized models”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 7089–7125.

— (2023b). “Theoretical characterization of uncertainty in high-dimensional linear classification”. In: *Machine Learning: Science and Technology* 4.2, p. 025029.

Cotter, S L, M Dashti, et al. (2009). “Bayesian inverse problems for functions and applications”. In: *Inverse Problems* 25.11, p. 115008.

Cotter, S L, G O Roberts, et al. (2013). “MCMC methods for functions: modifying old algorithms to make them faster”. In: *Statist. Sci.* 28.3, pp. 424–446.

Cox, Dennis D (1993). “An analysis of Bayesian inference for nonparametric regression”. In: *The Annals of Statistics*, pp. 903–923.

Crawford, Lorin et al. (2018). “Bayesian Approximate Kernel Regression With Variable Selection”. In: *Journal of the American Statistical Association* 113.524, pp. 1710–1721.

Dunlop, M M and Y Yang (2021). “Stability of Gibbs posteriors from the Wasserstein loss for Bayesian full waveform inversion”. In: *SIAM/ASA Journal on Uncertainty Quantification* 9.4, pp. 1499–1526.

Engl, Heinz W, Karl Kunisch, and Andreas Neubauer (1989). "Convergence rates for Tikhonov regularisation of non-linear ill-posed problems". In: *Inverse problems* 5.4, p. 523.

Engl, Heinz Werner, Martin Hanke, and Andreas Neubauer (1996). *Regularization of inverse problems*. Vol. 375. Springer Science & Business Media.

Flemming, Jens (2012). "Solution smoothness of ill-posed equations in Hilbert spaces: four concepts and their cross connections". In: *Applicable Analysis* 91.5, pp. 1029–1044.

Flemming, Jens, Bernd Hofmann, and Peter Mathé (2011). "Sharp converse results for the regularization error using distance functions". In: *Inverse Problems* 27.2, p. 025006.

Fortuin, Vincent et al. (2021). "Bayesian neural network priors revisited". In: *arXiv preprint arXiv:2102.06571*.

Freedman, David (1999). "Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters". In: *The Annals of Statistics* 27.4, pp. 1119–1141.

Gelman, A et al. (2013). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Ghorbani, Behrooz et al. (2021). "Linearized two-layers neural networks in high dimension". In: *The Annals of Statistics* 49.2, pp. 1029–1054.

Ghosal, S and A W van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press.

Ghosal, Subhashis and Aad W van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press.

Gibbs, Alison L. and Francis Edward Su (2002). "On Choosing and Bounding Probability Metrics". In: *International Statistical Review / Revue Internationale de Statistique* 70.3, pp. 419–435.

Gilbarg, David et al. (1977). *Elliptic partial differential equations of second order*. Vol. 224. 2. Springer.

Giné, Evarist and Richard Nickl (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Giordano, Matteo and Richard Nickl (2020). "Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem". In: *Inverse Problems* 36.8, p. 085001.

Golub, Gene H., Michael Heath, and Grace Wahba (1979). "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". In: *Technometrics* 21.2, pp. 215–223.

Grünwald, P D and J Langford (2007). "Suboptimal behavior of Bayes and MDL in classification under misspecification". In: *Machine Learning* 66.2-3, pp. 119–149.

Grünwald, P D and N A Mehta (2020). "Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes". In: *Journal of Machine Learning Research* 21, pp. 1–80.

Grünwald, P D and T van Ommen (2017). "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It". In: *Bayesian Anal.* 12.4, pp. 1069–1103.

Grünwald, Peter (2012). "The safe Bayesian: learning the learning rate via the mixability gap". In: *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*. Springer, pp. 169–183.

Guionnet, Alice et al. (2023). *Estimating rank-one matrices with mismatched prior and noise: universality and large deviations*. eprint: `arXiv:2306.09283`.

Hadamard, Jacques (1902). "Sur les problèmes aux dérivées partielles et leur signification physique". In: *Princeton university bulletin*, pp. 49–52.

Hastie, T, R Tibshirani, and J H Friedman (2009). *The Elements of Statistical Learning*. New York, NY: Springer.

Hastie, Trevor et al. (2022). "Surprises in high-dimensional ridgeless least squares interpolation". In: *The Annals of Statistics* 50.2, pp. 949–986.

Hofmann, Bernd (2006). "Approximate source conditions in Tikhonov–Phillips regularization and consequences for inverse problems with multiplication operators". In: *Mathematical Methods in the Applied Sciences* 29.3, pp. 351–371.

Hofmann, Bernd and Peter Mathé (2007). "Analysis of profile functions for general linear regularization methods". In: *SIAM Journal on Numerical Analysis* 45.3, pp. 1122–1141.

Hohage, Thorsten and Mihaela Pricop (2008). "Nonlinear Tikhonov regularization in Hilbert scales for inverse boundary value problems with random noise". In: *Inverse Probl. Imaging* 2.2, pp. 271–290.

Hu, Hong and Yue M Lu (2022). "Sharp Asymptotics of Kernel Ridge Regression Beyond the Linear Regime". In: *arXiv preprint arXiv:2205.06798*.

Hugenberg, N R et al. (2021). "Toward improved accuracy in shear wave elastography of arteries through controlling the arterial response to ultrasound perturbation in-silico and in phantoms". In: *Phys. Med. Biol.* 66.23, p. 235008.

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31.

Jasra, A et al. (2011). "Inference for Lévy driven stochastic volatility models via Sequential Monte Carlo". In: *Scand. J. of Statist* 38.1, pp. 1–22.

Jiang, W and M A Tanner (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining". In: *Ann. Statist.* 36.5, pp. 2207–2231.

Johnstone, Iain M (2010). "High dimensional Bernstein-von Mises: simple examples". In: *Institute of Mathematical Statistics Collections* 6, p. 87.

Kaipio, J and E Somersalo (2005). *Statistical and computational inverse problems*. New York, NY: Springer.

Kantas, N, A Beskos, and A Jasra (2014). "Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier-Stokes equations". In: *SIAM/ASA Journal on Uncertainty Quantification* 2.1, pp. 464–489.

Kato, Tosio (2013). *Perturbation theory for linear operators*. Vol. 132. Springer Science & Business Media.

Kleijn, BJK and AW van der Vaart (2012). "The Bernstein-Von-Mises theorem under misspecification". In: *Electronic Journal of Statistics* 6, pp. 354–381.

Knapik, B T, A W van der Vaart, and J H van Zanten (2011). "Bayesian inverse problems with Gaussian priors". In: *Ann. Statist.* 39.5, pp. 2626–2657.

Knapik, Bartek T et al. (2016). "Bayes procedures for adaptive inference in inverse problems for the white noise model". In: *Probability Theory and Related Fields* 164.3, pp. 771–813.

Knapik, BT, AW van der Vaart, and JH van Zanten (2011). "Bayesian inverse problems with gaussian priors". In: *The Annals of Statistics* 39.5, pp. 2626–2657.

Li, Zeng, Chuanlong Xie, and Qinwen Wang (2021). "Asymptotic Normality and Confidence Intervals for Prediction Risk of the Min-Norm Least Squares Estimator". In: *International Conference on Machine Learning*. PMLR, pp. 6533–6542.

Lions, Jacques Louis and Enrico Magenes (2012). *Non-homogeneous boundary value problems and applications: Vol. 1*. Vol. 181. Springer Science & Business Media.

Lytova, A and L Pastur (2009). "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries". In: *Annals of Probability* 37.5, pp. 1778–1840.

Martin, R, R Mess, and S G Walker (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models". In: *Bernoulli* 23.3, pp. 1822–1847.

Mei, Song, Theodor Misiakiewicz, and Andrea Montanari (2019). "Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit". In: *Conference on Learning Theory*. PMLR, pp. 2388–2464.

Mei, Song and Andrea Montanari (2022). "The generalization error of random features regression: Precise asymptotics and the double descent curve". In: *Communications on Pure and Applied Mathematics* 75.4, pp. 667–766.

Mei, Song, Andrea Montanari, and Phan-Minh Nguyen (2018). "A mean field view of the landscape of two-layer neural networks". In: *Proceedings of the National Academy of Sciences* 115.33, E7665–E7671.

Miller, J W and D B Dunson (2019). "Robust Bayesian inference via coarsening". In: *J. Am. Stat. Assoc.* 114.527, pp. 1113–1125.

Monard, François, Richard Nickl, and Gabriel P Paternain (2021). "Statistical guarantees for Bayesian uncertainty quantification in nonlinear inverse problems with Gaussian process priors". In: *The Annals of Statistics* 49.6, pp. 3255–3298.

Nair, MT (2009). "On Morozov's discrepancy principle for nonlinear ill-posed equations". In: *Bulletin of the Australian Mathematical Society* 79.2, pp. 337–342.

Neerven, JMAM van et al. (2010). "γ-radonifying operators—a survey". In: *The AMSI-ANU workshop on spectral theory and harmonic analysis*. Vol. 44. Austral. Nat. Univ. Canberra, pp. 1–61.

Nickl, Richard (2020). "Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation". In: *Journal of the European Mathematical Society* 22.8, pp. 2697–2750.

— (2023). *Bayesian non-linear statistical inverse problems*. EMS press.

Nickl, Richard, Sara van de Geer, and Sven Wang (2020). "Convergence rates for penalized least squares estimators in PDE constrained regression problems". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.1, pp. 374–413.

Nickl, Richard and Gabriel P Paternain (2022). "On some information-theoretic aspects of nonlinear statistical inverse problems". In: *Proc. Int. Cong. Math*. Vol. 7, pp. 5516–5538.

Nickl, Richard and Sven Wang (2022). "On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms". In: *Journal of the European Mathematical Society*.

Ovadia, Yaniv et al. (2019). "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift". In: *Advances in neural information processing systems* 32.

Owhadi, Houman and Clint Scovel (2017). "Qualitative robustness in Bayesian inference". In: *ESAIM: Probability and Statistics* 21, pp. 251–274.

Owhadi, Houman, Clint Scovel, and Tim Sullivan (2015). "Brittleness of Bayesian inference under finite information in a continuous world". In: *Electron. J. Statist.* 9.1, pp. 1–79.

Parhi, Rahul and Robert D Nowak (2022). “What kinds of functions do deep neural networks learn? Insights from variational spline theory”. In: *SIAM Journal on Mathematics of Data Science* 4.2, pp. 464–489.

Rahimi, Ali and Benjamin Recht (2007). “Random features for large-scale kernel machines”. In: *Advances in neural information processing systems* 20.

Resnick, Sidney (2019). *A probability path*. Springer.

Richter, Gerard R (1981). “An inverse problem for the steady state diffusion equation”. In: *SIAM Journal on Applied Mathematics* 41.2, pp. 210–221.

Roy, T and M N Guddati (2021). “Shear wave dispersion analysis of incompressible waveguides”. In: *J. Acoust. Soc. Am.* 149.972, pp. 972–982.

Roy, T, M Urban, et al. (2021). “Multimodal guided wave inversion for arterial stiffness: methodology and validation in phantoms”. In: *Phys. Med. Biol.* 66, p. 115020.

Silva, Luca and Giacomo Zanella (2022). *Robust leave-one-out cross-validation for high-dimensional Bayesian models*. arXiv: 2209.09190 [stat.CO].

Stuart, A M (2010). “Inverse problems: a Bayesian perspective”. In: *Acta Numerica* 19, pp. 451–559.

Stuart, Andrew M (2010). “Inverse problems: a Bayesian perspective”. In: *Acta numerica* 19, pp. 451–559.

Syring, N and R Martin (2019). “Calibrating general posterior credible regions”. In: *Biometrika* 106.2, pp. 479–486.

Syring, Nicholas and Ryan Martin (2023). “Gibbs posterior concentration rates under sub-exponential type losses”. In: *Bernoulli* 29.2, pp. 1080–1108.

Szabó, BT, AW van der Vaart, and JH van Zanten (2015). “Frequentist coverage of adaptive nonparametric Bayesian credible sets”. In: *The Annals of Statistics* 43.4, pp. 1391–1428.

Taylor, Michael (2013). *Partial differential equations I: Basic Theory*. Vol. 115. Springer Science & Business Media.

Tikhonov, A N and V Y Arsenin (1977). *Solutions of Ill-Posed Problems*. Washington, D. C.: Winston & Sons.

Triebel, Hans (1978). *Interpolation theory, function spaces, differential operators*. Vol. 18. North-Holland.

Vaart, AW van der and JH van Zanten (2007). “Bayesian inference with rescaled Gaussian process priors”. In: *Electronic Journal of Statistics* 1, pp. 433–448.

Van Der Vaart, Aad (1991). “On differentiable functionals”. In: *The Annals of Statistics*, pp. 178–204.

Van Der Vaart, Aad W and Jon A Wellner (1996). *Weak convergence*. Springer.

Vehtari, A, A Gelman, and J Gabry (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Stat. Comput.* 27, pp. 1413–1432.

Vollmer, Sebastian J (2013). “Posterior consistency for Bayesian inverse problems through stability and regression results”. In: *Inverse Problems* 29.12, p. 125011.

Wenzel, Florian et al. (2020). “How good is the bayes posterior in deep neural networks really?” In: *arXiv preprint arXiv:2002.02405*.

Weyl, Hermann (1911). “Über die asymptotische Verteilung der Eigenwerte”. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1911, pp. 110–117.

Zou, Z et al. (2019). *Adaptive particle-based approximations of the Gibbs posterior for inverse problems*. eprint: `arXiv:1907.01551`.

Biography

I have graduated Williams College *summa cum laude* in 2019 with B.A. in English and Statistics. I have begun my Ph.D. in Statistical Science at Duke from 2019 and have been a guest researcher in the Max Planck Institute for Mathematics in the Science, Leipzig in 2023. I currently hold a student membership in the International Society for Bayesian Analysis (ISBA). The following is a list of my previous publications.

Baek, Y., Berchuck, Samuel I., & Mukherjee, S. (2023). Asymptotics of Bayesian Uncertainty Estimation in Random Features Regression. *Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Baek, Y., Aquino, W., & Mukherjee, S. (2023). Generalized Bayes Approach to Inverse Problems with Model Misspecification. *Inverse problems*, **39** 105011

Binette, O., York, S. A., Hickerson, E., Baek, Y., Madhavan, S., & Jones, C. Estimating the Performance of Entity Resolution Algorithms: Lessons Learned Through PatentsView.org. *The American Statistician*, **77**(4), 370–380.