THREE ESSAYS OF BAYESIAN INFERENCE ON DYNAMICAL SYSTEM, CONTINUOUS TIME MARKOV CHAIN, AND LOW DIMENSIONAL STRUCTURE

by

Tao Tang

Department of Mathematics
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Xiuyuan Cheng

_____
Jianfeng Lu

_____
Simon Mak

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

2023

ABSTRACT

THREE ESSAYS OF BAYESIAN INFERENCE ON DYNAMICAL SYSTEM, CONTINUOUS TIME MARKOV CHAIN, AND LOW DIMENSIONAL STRUCTURE

by

Tao Tang

Department of Mathematics
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Xiuyuan Cheng

_____
Jianfeng Lu

_____
Simon Mak

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

2023

Abstract

This dissertation propels the frontier of Bayesian inference, particularly for structurally diverse data, by developing innovative methodologies and robust theories, delineated into three distinct segments, each tackling Dynamical systems, continuous time Markov Chains, and data with low-dimensional structures.

The initial part presents the Hierarchical Shrinkage Gaussian Process (HierGP), a novel model adept at detecting structured sparse features within limited data from response surfaces. Incorporating the principles of effect sparsity, heredity, and hierarchy into a Gaussian process framework, HierGP demonstrates superior performance compared to its contemporaries. This superiority is backed by a range of numerical experiments and its successful application to dynamical system recovery.

The dissertation's second segment introduces the Bayesian Approximation Spectral Inference (BASI) method. Inspired by the necessity for pragmatic inference techniques for continuous-time Markov chains, BASI innovatively employs probabilistic matrix factorization to estimate a continuous-time Markov chain generator's low-rank representation. This approach effectively circumvents the computational challenges of assessing implicit likelihood functions inherent in previous methods, with our theoretical support for BASI showcased via the asymptotic properties of the posterior distribution.

In its final part, the dissertation delves into posterior contraction rates for Gaussian process regression in the context of data with low-dimensional structures. The conditions established enable adaptivity to any intrinsic dimension, thereby giving the optimal posterior contraction rate. This exploration is further enriched by an innovative empirical Bayes prior of bandwidth, dispensing with the necessity for prior knowledge of the intrinsic dimension.

Cumulatively, these three essays enhance Bayesian statistics for data with varying and intricate structures. The dissertation seamlessly intertwines theoretical contributions with practical numerical experiments, lending empirical weight to our innovative propositions.

Contents

List of Tables

List of Figures

Acknowledgements

I am profoundly grateful to my supervisor and committee chair, Professor David B. Dunson, whose exceptional guidance and constant support have been invaluable throughout my academic journey. My sincere thanks are extended to Professor Simon Mak for his patience, mentorship, and indispensable assistance during my Ph.D. tenure. I cannot overstate my gratitude to Professor Xiuyuan Cheng, whose guidance and motivation were paramount to the successful completion of the last part of my thesis.

I extend my heartfelt appreciation to the committee who were responsible for my preliminary examination and thesis defense. I would like to recognize Xiuyuan Cheng, Jianfeng Lu, and Sayan Mukherjee for their vital contributions to my work.

I owe a debt of gratitude to my collaborators. My sincere thanks go to Trevor Bowman, James Johndrow, Omar Melikechi, Chih-li Song, Hau-tieng Wu, Irene Ji Yi, and Alex Young for their critical roles in my research endeavors.

In conclusion, my deepest appreciation is reserved for my family and friends, who have provided unwavering support and encouragement throughout my academic journey. Their belief in me and my work has been a consistent source of motivation and strength for which I am forever grateful. I am especially thankful to my parents, Min Ao and Qing Tang, for their lifelong love and education. I would also like to extend special thanks to my girlfriend, Xizhao Wang. Her friendship, support, and love have been my rock since we first met.

Chapter 1

Introduction

In this section, we review the literature and give an outline of our contributions in each essay.

1.1 Computer code emulation and dynamical system recovery

Scientific computing is playing an increasingly important role in solving modern scientific and engineering problems, particularly with recent breakthroughs in mathematical modeling and computing. Quantities of interest that were difficult or infeasible to observe from physical experiments can now be reliably simulated from computer code. Such *computer experiments* typically involve solving a system of differential equations based on physical models and have had a wide-reaching impact in many fields, including rocket engine design [MSW+18], personalized surgical planning [CMJZ21], and universe expansions [KBH+11; JMS+21]. One critical bottleneck, however, is that these virtual experiments can be very resource-intensive for computation. For example, the full-scale simulation of a single rocket engine injector can require millions of CPU hours [MSW+18]. This in turn places heavy demand on computational resources for design exploration and optimization.

One way to address this is to train an *emulator* model that can efficiently predict (or *emulate*) the expensive computer code $f(\boldsymbol{x})$ at different parameters $\boldsymbol{x}$. The idea is to run the computer code at carefully chosen points over the parameter space $\mathcal{X}$, then use this as training data to fit the emulator model. A popular choice of emulator model is the Gaussian process (GP; see Ras03), a flexible Bayesian model

for probabilistic predictive modeling. GPs have two key advantages for computer code emulation [SWNW03]: they offer closed-form expressions for prediction and uncertainty quantification of the emulator, and provide a flexible non-parametric framework for modeling the black-box response surface.

Despite this, standard GPs have several notable limitations for emulation. One problem is that, when the training sample size n is small, the highly flexible form of a GP can become more of a vice than a virtue. In particular, with limited sample sizes, the GP posterior may distribute probability across a wide range of functions, thus resulting in not only poor predictive performance but also high predictive uncertainty. A promising solution is to elicit prior knowledge on the response surface f (for example, based on knowledge of the physics underlying the applied domain), and integrate this within the probabilistic GP model. This knowledge can take the form of mechanistic models [WDP$^+$14], boundary conditions [DMW19], or shape constraints [GBCC15; WB16], and can greatly improve the predictive performance of a GP model with limited data. However, for highly complex physical systems, the specific prior knowledge for such existing models can be difficult to elicit with confidence and too complex to integrate for probabilistic modeling. In lieu of this, alternate emulators are needed to learn useful structures for prediction in data-limited settings. In the statistics literature on GPs, some relevant existing ideas include GPs with variable selection [SVS11] and graph Laplacian GPs [DWW20], which allow the inputs to be restricted to an unknown lower-dimensional manifold.

We refer to our proposed approach as the hierarchical shrinkage GP (HierGP), which embeds this structured sparsity of effects (namely, the principles of effect hierarchy and effect heredity) via carefully-designed shrinkage priors within a Gaussian process framework. The HierGP begins with a basis expansion of a GP, then assigns hierarchical shrinkage priors on basis coefficients to capture the desired hierarchical

sparsity structure. In particular, we adapt the cumulative shrinkage priors proposed in [LDD20] (which was used for factor analysis), and show that the HierGP model with such priors embedded indeed has the desired properties of effect hierarchy and heredity. We then propose an efficient Gibbs sampler that leverages a data augmentation trick for efficient posterior computation. Under mild conditions on sparsity, we then prove posterior contraction results for the HierGP model for both fixed and randomly-sampled design points. Finally, we demonstrate the effectiveness of the proposed HierGP over existing methods in a suite of numerical experiments and an application to dynamical system recovery.

1.2 Inference of continuous-time Markov chains (CTMC)

Inference on continuous-time Markov chains (CTMC) from discretely observed data is of increasing interest in many fields ranging from biology and physics to finance ([CVE06], [MFK08], [MDJS07]). For instance, molecular dynamics employs large state space CTMCs to investigate protein folding. The state space corresponds to protein conformations, and transitions indicate folding or unfolding events. Researchers attempt to infer the CTMC parameters from discretely-observed time series data. However, substantial computational challenges arise, particularly in settings involving many states, due in part to the intractability of the likelihood function. In practice, Markov processes with large state spaces often display approximately low-rank structures ([ZLWZ22], [RZMC11], [BGL17]). This motivates us to develop low-rank Bayesian spectral approximation models, Bayesian Approximation Spectral Inference (BASI), for the estimation and reconstruction of CTMC from high-dimensional discretely observed data.

For inference in CTMC models, it is natural to focus on estimating the generator.

However, direct estimation of the CTMC generator is difficult due to the implicit nature of the likelihood function for discretely observed data. Several approaches have been proposed to evaluate the implicit likelihood and then infer the generator. One classical type of method relies on data augmentation ([BS05; BAVX21]) to evaluate the implicit likelihood. [EDB18], instead, uses the solutions of the C-K equation, based on the empirical transition probabilities, to evaluate the likelihood and further estimate the generator. [She21] evaluates the likelihood via a direct matrix exponential. All of these approaches are computationally heavy and difficult to scale up to large state spaces and long time series ([She21]).

This paper presents approximate Bayesian techniques for inference on continuous-time Markov chains (CTMC) using their spectral representation. Rather than directly estimating the generator of the CTMC, we instead estimate its spectral characteristics under a low-rank approximation. Leveraging this property, we introduce approximation methods that can identify the optimal spectral representation of the CTMC from the data. To efficiently approximate the likelihood under the low-rank spectral representation, we develop innovative probabilistic matrix factorizations. This approach results in an explicit likelihood function for the spectral features, significantly reducing computational complexity.

Similar to us, [HKWBC14] and [MDJS07] do not evaluate the implicit likelihood function of the generator. [HKWBC14] use a particle-based Monte Carlo approach, where the holding times are marginalized analytically, to estimate the generator. Our paper, compared to [HKWBC14], can achieve lower computational complexity via the low rank approximation. [MDJS07] consider matching the spectral properties of the generator via quadratic programming. [MDJS07] consider recovering the generator, while we focus on the low-rank approximation. Our methods, as Bayesian approaches, further provide natural uncertainty quantification for the reconstructed CTMC and

associated spectral characteristics.

Previous literature has discussed the model reductions and low-rank approximation of discrete-time dynamical systems ([Gia19; LWZ18]). For Markov processes, [ZW19] and [ZLWZ22] study low-rank transition matrix recovery for the discrete-time case. However, continuous-time cases have been less studied. To the best of our knowledge, our paper is the first to develop a low-rank approximation of Markov processes in continuous-time.

1.3 Data with Low Intrinsic Dimension

The curse of dimensionality in statistical learning theory refers to the significant impact of input space dimensionality on the convergence rates of nonparametric methods used in regression and classification. As a result, researchers have invested considerable effort in establishing the concept of intrinsic dimension in data, aiming to improve the dependency on the ambient space dimension for well-known learning rates. The primary objective of this approach is to achieve faster convergence rates by assuming a substantial disparity between the intrinsic dimension of the data and the dimension of its ambient space. This assumption holds practical importance for real-world datasets, as indicated by the multitude of techniques developed for dimensionality reduction. Thus, there is a compelling interest in conducting theoretical analyses to evaluate the performance of common learning methods under the assumption of low intrinsic dimension in the data [HS21].

Intrinsic dimensionality in data analysis has been extensively studied. The prevailing notion assumes the data-generating distribution is supported on a low-dimensional smooth manifold $\mathcal{M} \subset \mathbb{R}^D$, as observed in previous studies [BL07a; SN06; YD16; YZ08; YZ09; DWW22]. Specifically, previous literature on Bayesian nonparametric regression methods focus on the manifold assumption with known [CKP14] and un-

known manifold [YD16; DWW22; NDCW23]. So far, there is a lack of study using the low intrinsic dimensionality beyond the manifold assumption in Bayesian nonparametric regression setting. To fill this gap, our work uses the covering number of the subspace which considerably generalizes the manifold assumption.

Some works have explored the notion of intrinsic dimension goes beyond the traditional manifold assumption by utilizing the box-counting dimension of the distribution's support [HS21]. Existing research primarily focuses on elementary learning algorithms within notions of low intrinsic dimensionality, such as local polynomial regression, k-nearest neighbor, Nadaraya-Watson kernel regression,decision trees, and least-squares kernel ridge regression [BL07a; Kpo11; KG13; KP95; SN06; HS21]. However, the above explorations beyond the manifold assumption do not cover any Bayesian nonparametric regression setting.

In this essay, our focus is on the Gaussian process (GP) for the data with unknown low intrinsic dimension. Gaussian process (GP) is one of the most popular nonparametric Bayesian methods, and has gained considerable popularity as non-parametric priors for modeling unknown functions. In the context of nonparametric regression, Gaussian processes can be employed as priors for capturing the uncertainty associated with the unknown function. In this work, we focus on the regression model under Gaussian noise:

$$Y_i = f^*(X_i) + w_i, w_i \overset{i.i.d}{\sim} N(0, \sigma^2), i = 1, ..., n \tag{1.1}$$

for which $Y_i \in \mathbb{R}$ and $X_i \in \mathcal{X}$ where $\mathcal{X}$ be a measurable subset of $[0,1]^D$, $f^* : \mathcal{X} \to \mathbb{R}$, and σ is known. $X_1, ..., X_n$ are either fixed or i.i.d samples from a distribution $\mathbb{P}_X$ on $\mathcal{X}$. Our goal is to estimate the f^* by an estimator $\hat{f}$ and minimize the L_2 population loss $||f^* - \hat{f}||_2^2 := \int_{\mathcal{X}} (f^*(x) - \hat{f}(x))^2 \mathbb{P}_X(dx)$ or in-sample empirical loss $||f^* - \hat{f}||_n^2 := \mathbb{P}_n(|f^* - \hat{f}|^2) = \frac{1}{n}\sum_{i=1}^n (f^*(X_i) - \hat{f}(X_i))^2$, where $\mathbb{P}_n$ is the empirical measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

A few paper analyze the performance of GP under the assumption of low intrinsic dimensionality. [CKP14] rescale solutions of the heat equation to construct a nonstationary Gaussian process prior on a known manifold. It is important to note that in numerous scenarios, however, the underlying manifold is not known a priori. [YD16] introduced a methodology that achieves a minimax-optimal adaptive rate in estimating the function f on unknown manifold. This adaptive rate is designed to adapt to the manifold and smoothness of the regression function, while their order of smoothness should not be larger than 2 and dimension needs to be estimated first. In their work, [DWW22] propose a covariance construction that incorporates the intrinsic geometry of the data for building the Gaussian Process (GP). However, as acknowledged in their paper, their posterior contraction rate results are suboptimal on manifold, suggesting potential avenues for future research exploration for the optimal results.

In this study, we establish the conditions under which the posterior contraction rates adapt to low intrinsic dimensionality, as measured by the covering number. To demonstrate our findings, we introduce a Gaussian process that satisfies these derived conditions. To address the challenge of lacking prior knowledge of the intrinsic dimension, we employ a novel empirical Bayes priors approach using kernel count statistics and the K-nearest neighbor (KNN) algorithm to determine the bandwidth prior. This approach eliminates the need for prior knowledge and allows for adaptive estimation. The proposed Gaussian process exhibits optimal adaptive posterior contraction rates, up to a logarithmic constant, for any order of function smoothness $(0 < s < \infty)$ on a unknown manifold, subject to the intrinsic conditions.

Chapter 2

Hierarchical shrinkage Gaussian processes for dynamical system recovery

2.1 Introduction

In this section, we describe the HierGP and discuss connections with existing GP models. We first review standard GPs and their representation as an infinite basis expansion with random coefficients. In what follows, we let $f(\boldsymbol{x})$ denote the expensive black-box function to be emulated, where $\boldsymbol{x} \in \mathbb{R}^d$ are its input parameters. A Gaussian process [Ras03; SWNW03] adopts the following probabilistic prior on $f(\boldsymbol{x})$:

$$f(\cdot) \sim \mathrm{GP}\{\mu(\cdot), k(\cdot,\cdot)\}. \tag{2.1}$$

Here, $\mu(\cdot)$ is the mean function of the process, and $k(\cdot,\cdot)$ its covariance function. A key appeal for GP modeling is that, conditional on observed data $f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)$ from the black-box system, the posterior predictive distribution $[f(\cdot)|f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)]$ remains a GP with closed-form posterior mean and covariance. This facilitates prediction and uncertainty quantification via a flexible Bayesian nonparametric model.

In order to integrate the desired hierarchical shrinkage structure, however, we will employ an alternate representation of the GP as an infinite basis expansion. This relies on the well-known Karhunen–Loève theorem (stated below), which shows that a GP can be represented as an infinite basis expansion of orthonormal functions.

Theorem 2.1.1 (Karhunen–Loève, Theorem 5.3 of Ale15)**.** *Let $\mathcal{X} \subseteq R^d$ be the input space of interest. Let $f(\boldsymbol{x}) \sim GP\{0, k(\cdot,\cdot)\}$ be a zero-mean Gaussian process, with covariance kernel k continuous on $\mathcal{X} \times \mathcal{X}$, and $k(\boldsymbol{x}, \boldsymbol{x}') \in L^2(\mathcal{X})$. Then there exists*

an orthonormal basis $\{\phi_k(\boldsymbol{x})\}_{k=1}^{\infty}$ *of* $L^2(\mathcal{X})$ *such that:*

$$f(\boldsymbol{x}) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{x}), \tag{2.2}$$

where the coefficients $\{\lambda_k\}_{k=1}^{\infty}$ *are independent Gaussian random variables given by* $\lambda_k = \int_{\mathcal{X}} f(\boldsymbol{x})\phi_k(\boldsymbol{x})d\boldsymbol{x}$*, and satisfy* $\mathbb{E}(\lambda_k) = 0$ *and* $\mathbb{E}(\lambda_j\lambda_k) = \mathbf{1}(j = k)\mathrm{Var}(\lambda_k)$.

Here, the conditions on kernel k ensure the Gaussian process f is mean-square continuous, with $f \in L^2(\mathcal{X} \times \Omega)$ [Ale15]. (2.2) is also known as the Karhunen–Loève expansion, which is widely used in statistics [Wan08] and uncertainty quantification [Xiu10; GS91]. This expansion can be viewed as a stochastic analogy of the classical Fourier expansion.

As a tangible illustration, consider the case with dimension $d = 1$ and $f : \mathbb{R} \to \mathbb{R}$. Suppose $f(x)$ follows the Gaussian process in (2.2) with mean function $\mu(\cdot) \equiv 0$ and squared-exponential covariance function:

$$k(x, x') = \sigma^2 \exp\left\{-\frac{-(x - x')^2}{2l^2}\right\}. \tag{2.3}$$

Then one can show (see, e.g., Ras03) the corresponding Karhunen-Loève decomposition of $f(x)$ to be of the form (2.2) with

$$\lambda_k = \sqrt{\frac{2a}{A}} B^k \xi_k, \qquad \phi_k(x) = \exp\left\{-(c - a)x^2\right\} H_k(\sqrt{2c}x), \tag{2.4}$$

where $\xi_k \overset{i.i.d.}{\sim} N(0, 1)$, H_k is the k-th order Hermite polynomial (see, e.g., Xiu10), with $a^{-1} = 4\sigma_w^2$, $b^{-1} = 2l^2$, $c = \sqrt{a^2 + 2ab}$, $A = a + b + c$ and $B = b/A$.

For our motivating problem of emulation of expensive computer codes, the decomposition in (2.2) highlights a potential limitation of the standard GP as an emulator model. From a prior specification perspective, the use of independent Gaussian priors on the coefficients λ_k for each basis function $\phi_k(\boldsymbol{x})$ may not reflect the modeler's

prior belief that there are a *sparse* number of *structured* dominant features in the response surface. This is because such coefficients are all non-zero (and thus active) with probability one. Thus, when such sparse structure indeed is present in the response surface, the use of standard GPs for emulation may result in poor predictive performance and high uncertainty, both of which are undesirable. We present next a novel modification of the basis representation (2.2), which embeds the desired hierarchical sparsity structure within the prior specification of $(\lambda_k)_{k=1}^{\infty}$, thus allowing for the structured sparse learning of dominant features for the response surface $f(\boldsymbol{x})$.

2.2 The univariate HierGP model

We now present the model specification of the HierGP, which provides a sparse modification of the GP basis representation in (2.2). For simplicity, let us introduce the model first for the $d = 1$ dimensional setting. Suppose the design space is the unit interval $\mathcal{X} = [0, 1]$, and we obtain the training data $\{(x_i, y_i)\}_{i=1}^n$, where $x_1, \cdots, x_n \in \mathcal{X}$ are the training input parameters, and $y_1, \cdots, y_n$ are its corresponding outputs. We assume the outputs are obtained from the model:

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \theta^2), \quad i = 1, \cdots, n. \tag{2.5}$$

For the earlier problem of computer code emulation, the error term ϵ_i reduces to zero for deterministic simulators, since observations from f are obtained without noise. For the sake of generality, we will adopt the above noisy model specification for the remainder of the paper, and reduce the error term to zero whenever appropriate.

Following (2.2), the HierGP assumes a basis expansion model on the response surface f:

$$f(x) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x). \tag{2.6}$$

Here, $\{\phi_k(x)\}_{k=1}^{\infty}$ is a fixed L^2-orthonormal basis on $\mathcal{X}$, and $\{\lambda_k\}_{k=1}^{\infty}$ its corresponding coefficients. We then adopt the following hierarchical prior specification for the coefficients λ_k:

$$\lambda_k \sim N(0, \sigma_k^2), \quad \sigma_k^2 \sim w_k \pi_k + (1 - w_k)\delta_{\sigma_\infty^2}, \quad k = 1, 2, \cdots \tag{2.7}$$

The prior $\sigma_k^2 \sim w_k \pi_k + (1 - w_k)\delta_{\sigma_\infty^2}$ is shorthand for the spike-and-slab prior [IR05], where π_k is the slab distribution and $\delta_{\sigma_\infty^2}$ (a point mass at σ_∞^2) is the spike distribution. With probability w_k, this prior samples from π_k; otherwise, it samples a point mass at σ_∞^2. The first case can be seen as the coefficient λ_k being *active*, i.e., influential for the response surface f, whereas the latter can be seen as λ_k being *inactive*, i.e., inert for the response surface. A common choice for the slab distribution is $\pi_k = IG(a_\sigma, b_\sigma)$; we later adopt this in Section 2.4 for the proposed Gibbs sampler.

For spike-and-slab priors, it is well-known that the point mass $\delta_{\sigma_\infty^2}$ with $\sigma_\infty^2 = 0$ may cause computational instability and poor mixing of the resulting MCMC chain [SFK12]; we have encountered similar issues in our implementation. A common workaround is to set σ_∞^2 as a small (but non-zero) constant, such that σ_∞^2 is much smaller than the mean of the slab distribution $E(\pi_k)$. The specification of hyperparameters a_σ, b_σ and σ_∞^2 will be discussed in later sections.

Of course, prior to data, the probability w_k (for λ_k to be active) is typically unknown. We thus assign the following *cumulative* prior on $\{w_k\}_{k=1}^{\infty}$, adapted from the cumulative shrinkage priors in [LDD20] for sparse factor modeling.

$$w_k = \prod_{j=1}^{k}(1 - \nu_j), \quad \nu_j \overset{i.i.d.}{\sim} Beta(1, \alpha_j = \alpha), \quad w_0 = 1, \quad j = 1, 2, \cdots. \tag{2.8}$$

To complete the specification, we assign an independent Inverse-Gamma prior on the noise variance as $\theta^2 \sim IG(a_\theta, b_\theta)$. Prior (2.8) provides an appealing *cumulative* property which addresses the aforementioned need for effect hierarchy in the response

surface f. To see why, note that $\mathbb{E}(w_k) = (\alpha/(1+\alpha))^k$. Thus, for smaller indices k, one can see that the prior favors larger values of w_k; similarly, for larger indices k, it favors smaller values of w_k. This cumulative property nicely captures *effect hierarchy* [WH09], a widely-used statistical principle in experimental analysis which asserts that lower-order effects are more important than higher-order ones. The prior (2.7) directly embeds this property by placing greater weight on larger coefficients for lower-order coefficients, and on smaller coefficients for higher-order ones.

2.3 The multivariate HierGP model

With this univariate case in hand, we now present the multivariate HierGP model. Suppose the design space is the unit hypercube $\mathcal{X} = [0,1]^d$, and suppose we collect training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathcal{X}$ are training input parameters, and $y_1, \cdots, y_n$ its corresponding outputs. As before, we assume the outputs are obtained from the model:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \theta^2), \quad i = 1, \cdots, n, \tag{2.9}$$

where f takes the form

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k}}^{\infty} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \quad \boldsymbol{k} = (k_1, \cdots, k_d), \tag{2.10}$$

where $\boldsymbol{k} \in \mathbb{N}_0^d$. Here, $\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^{\infty}$ is an L^2-orthonormal basis on $\mathcal{X} = [0,1]^d$, and $\{\lambda_{\boldsymbol{k}}\}_{|\boldsymbol{k}|=1}^{\infty}$ are its coefficients. As before, we adopt independent spike-and-slab priors on the coefficients:

$$\lambda_{\boldsymbol{k}} \overset{indep.}{\sim} N(0, \sigma_{\boldsymbol{k}}^2), \quad \sigma_{\boldsymbol{k}}^2 \sim w_{\boldsymbol{k}} \pi_{\boldsymbol{k}} + (1 - w_{\boldsymbol{k}}) \delta_{\sigma_\infty^2}, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \tag{2.11}$$

where $\pi_{\boldsymbol{k}}$ and $\delta_{\sigma_\infty^2}$ are again the slab and spike distributions, respectively.

In the earlier univariate setting, the prior (2.8) on spike-and-slab probabilities embeds the first principle of effect hierarchy. We now generalize this prior to the multivariate setting to capture the second principle of *effect heredity*:

$$w_{\boldsymbol{k}} = \prod_{m=1}^{d} w_{k_m,m}, \quad w_{k_m,m} = \prod_{j=1}^{k_m}(1-\nu_{j,m}), \quad w_{0,m} = 1, \quad \nu_{j,m} \overset{i.i.d.}{\sim} Beta(1, \alpha_{j,m} = \alpha), \tag{2.12}$$

for $m = 1, \cdots, d$ and $j = 1, 2, \cdots$. Recall that the principle of *effect heredity* [HW92] (or the *marginality principle*; see MN89) presumes an interaction effect is active only when all of its component variables are also active. Here, the j-th variable is deemed "active" if its cumulative term $w_{k_m,m}$ is large and "inactive" if $w_{k_m,m}$ is small, since a larger $w_{k_m,m}$ results in a larger probability that its corresponding coefficient is non-zero and vice versa. To see why the above prior embeds this principle, note that the probability $w_{\boldsymbol{k}}$ for each multi-index $\boldsymbol{k}$ is modeled as a *product* of separate cumulative terms $w_{k_m,m}$ for each of the d dimensions. By setting $w_{\boldsymbol{k}}$ as the product form (2.12), it follows that term $\boldsymbol{k}$ is active (i.e., large) only when all of its component variables (represented by its cumulative terms) are active, which is precisely effect heredity. One can thus view the prior model (2.12) as a *structured* shrinkage prior on coefficients, which captures the desired hierarchical effect principles for response surface modeling.

The hierarchical shrinkage framework (2.12) can be easily modified in several ways to incorporate additional prior information from the modeler. First, in some cases, a modeler may have preference on a choice of basis $\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^{\infty}$ (which may not be L_2-orthonormal) for which this hierarchical sparsity is expected to hold. This non-orthonormal basis can be directly integrated within this modeling framework and the later posterior sampler (our assumption of orthonormality is, however, needed in later theoretical results). Second, if a modeler expects heavier-tailed effects for coefficients,

they can easily adopt heavier-tailed distributions (e.g., the horseshoe priors in CPS09) within the spike-and-slab prior framework (2.11). One can also integrate heavier tails via a careful specification of the hyperparameter sequence $\alpha_{j,m}$. For example, with $\alpha_{j,m}$ set as an increasing sequence (e.g., $\mathcal{O}(m^2)$), we can place greater weights on latter terms in the expansion and thus impose heavier tail behavior.

2.3.1 Alternate choices of shrinkage priors: HierGP$_2$

For certain problems, one may not have a strong prior belief for hierarchical cumulative shrinkage of the coefficients $\lambda_{\boldsymbol{k}}$, despite knowing such coefficients are likely sparse. In such a setting, an alternate prior model on $\lambda_{\boldsymbol{k}}$ may be the exchangeable sparse priors given by:

$$\lambda_{\boldsymbol{k}}|\sigma^2_{\boldsymbol{k}} \sim N(0, \sigma^2_{\boldsymbol{k}}\tau^2), \quad \sigma_{\boldsymbol{k}} \sim \Psi, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \tag{2.13}$$

where Ψ is a distribution supported on $(0, +\infty)$. Such priors are known as global-local (GL) mixtures: $\{\sigma_{\boldsymbol{k}}\}_{\boldsymbol{k}}$ are typically called the *local shrinkage* parameters, and τ the *global shrinkage* parameter. The GL prior provides a flexible framework for Bayesian shrinkage and approximate feature selection: heavy-tailed distributions on Ψ allow for identification of strong signals, and its concentration around zero provides the desired (approximately) sparse behavior. Examples of GL priors include the Horseshoe prior [CPS10], the Dirichlet–Laplace prior [BPPD15], and the generalized double Pareto prior [ADL13]. We shall call the basis expansion model (2.10) with GL priors (2.13) the HierGP$_2$ model; in later experiments, the HierGP$_2$ is implemented with Φ taken as the horseshoe prior from [CPS10]. When the underlying response surface has sparse features which do not adhere strongly to effect sparsity or effect heredity, we show later the HierGP$_2$ can also yield improved predictive performance over existing emulation models.

2.4 Posterior Sampling

With the HierGP in hand, we present next an efficient MCMC algorithm for posterior sampling of the response surface $f(\cdot)$ given data. We first present a Gibbs sampler for the univariate HierGP, then extend this to a Gibbs sampler for the multivariate HierGP.

2.4.1 Gibbs sampling for the univariate HierGP

Suppose we collect training data $\{(x_i, y_i)\}_{i=1}^n$ from model (2.5), and let $\Theta = \{(\lambda_k)_{k=1}^\infty, (v_j)_{j=1}^\infty, \theta^2\}$ be the parameter set for posterior inference. One can write the likelihood function[1] as:

$$L(\Theta|\{(x_i, y_i)\}_{i=1}^n) = [\{(x_i, y_i)\}_{i=1}^n|\Theta] = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(y_i - f(x_i))^2}{2\theta^2}}, \tag{2.14}$$

where the response surface f is a function of parameters $(\lambda_k)_{k=1}^\infty$ (see (2.6)). Conditional on data, we then wish to draw samples $\{\Theta_{[b]}\}_{b=1}^B$ from the posterior distribution:

$$[\Theta|\{(x_i, y_i)\}_{i=1}^n] \propto L(\Theta|\{(x_i, y_i)\}_{i=1}^n)\ [\Theta]. \tag{2.15}$$

Here, $[\Theta] = [(v_j)_{j=1}^\infty, (\lambda_k)_{k=1}^\infty, \theta^2]$ follows from the prior model in (2.6) and (2.7) for the univariate HierGP. With samples $\{\Theta_{[b]}\}_{b=1}^B$, posterior predictive samples on f (call this $\{f_{[b]}(\cdot)\}_{b=1}^B$) can then obtained by plugging $\{\Theta_{[b]}\}_{b=1}^B$ into (2.6).

To sample from (2.15), we will make use of a data-augmented Gibbs sampler [GCSR95], which leverages closed-form full conditional distributions for efficient posterior sampling. This sampler is similar in spirit to the Gibbs sampler in [LDD20] for factor models, but adapted for the GP setting at hand. As mentioned earlier, we adopt the choice of $\pi_k = \pi = IG(a_\sigma, b_\sigma)$ for the slab distribution in (2.7). We further

[1]Here, $[X]$ denotes the distribution of a random variable X.

employ the following truncation for $f(\cdot)$:

$$f(x) = \sum_{k=1}^{K} \lambda_k \phi_k(x) \tag{2.16}$$

for a sufficiently large choice of truncation index K; further details are provided in Appendix 2.9 on an adaptive choice of K, following [LDD20].

We now derive the required full conditional distributions. Let

$$\tilde{\boldsymbol{X}} = (\phi_1(x_i), \cdots, \phi_K(x_i))_{i=1}^n$$

be the design matrix for the data. Further, for $k = 1, \cdots, K$, let z_k be a latent categorical random variable, defined conditionally as $\mathbb{P}(z_k = l|\Theta) = \nu_l w_{l-1}$ for $l = 1, 2, \cdots, K$. It can be shown[2] that:

$$[\sigma_k^2|z_k] \sim \{1 - \mathbf{1}(z_k \leq k)\} IG(a_\sigma, b_\sigma) + \mathbf{1}(z_k \leq k)\delta_0, \tag{2.17}$$

where $\mathbf{1}(\cdot)$ is the indicator function. With this data augmentation trick, one can then easily derive the full conditional distribution of z_k as:

$$[z_k = l|-] \propto \begin{cases} \nu_l w_{l-1} \phi(\lambda_k; 0, \sigma_\infty^2), & l = 1, \cdots, k, \\ \nu_l w_{l-1} t_{2a_\sigma}(\lambda_k; 0, (b_\sigma/a_\sigma)), & l = k+1, \cdots, K, \end{cases} \tag{2.18}$$

where $\phi(\lambda; 0, \sigma_\infty^2)$ and $t_{2a}(\lambda; 0, (b_\sigma/a_\sigma))$ are the densities of the normal and t-distributions (with $2a$ degrees-of-freedom) evaluated at λ, respectively.

Consider now the full conditional distributions for the parameter set Θ. Let $\boldsymbol{D} = \text{diag}\{\sigma_k^2\}_{k=1}^K$ and $\boldsymbol{y} = (y_1, \cdots, y_n)$. For the coefficient vector $\boldsymbol{\Lambda} = (\lambda_k)_{k=1}^K$, its full conditional distribution can be shown to be:

$$[\boldsymbol{\Lambda}|-] \sim N(\boldsymbol{V}\theta^{-1}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V}), \quad \boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}. \tag{2.19}$$

[2]For brevity, the notation $[\theta|-]$ denotes the full conditional distribution of parameter θ, conditional on both the data $y_1, \cdots, y_n$ and all parameters in Θ except for the considered parameter θ.

Algorithm 1: Gibbs sampling for the univariate HierGP

Inputs: hyperparameters a_σ, b_σ, a_θ, b_θ, σ^2_∞, data $\{(x_i, y_i)\}_{i=1}^n$, number of iterations B.

Set initial parameters $\Theta_{[0]} = \{(\lambda_k^{[0]})_{k=1}^K, (v_j^{[0]})_{j=1}^K, \theta^2_{[0]}\}$.
for $b = 1, \cdots, B$ **do**
  for $k = 1, \cdots, K$ **do**
    Sample $z_k^{[b]}$ from the full conditional distribution $[z_k = l|-]$ in (2.18) with $\lambda_k = \lambda_k^{[b-1]}$.
  end for (Step 1)
  for $k = 1, \cdots, K$ **do**
    If $z_k^{[b]} \leq k$, let $\sigma_k^2 = \sigma^2_\infty$; else sample $\sigma_k^2 \sim IG(a_\sigma + 0.5, b_\sigma + 0.5(\lambda_k^{[b-1]})^2)$.
  end for (Step 2)
  for $j = 1, \cdots, K$ **do**
    Sample $v_j^{[b]} \sim \text{Beta}(1 + \sum_{k=1}^K \mathbf{1}(z_k^{[b]} = j), \alpha_j + \sum_{k=1}^K \mathbf{1}(z_k^{[b]} > j))$; update w_j in (2.12).
  end for
  Sample $\theta^2_{[b]} \sim IG(a_\theta + N/2, b_\theta + \boldsymbol{S}^T\boldsymbol{S}/2)$ where $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}_{[b-1]}\tilde{\boldsymbol{X}}$. (Step 3)
  Sample $\boldsymbol{\Lambda}_{[b]} = (\lambda_k^{[b]})_{k\leq K} \sim N(\boldsymbol{V}\theta^{-2}_{[b]}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}_{[b]}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\boldsymbol{D} = \text{diag}(\sigma_k^2)_{k\leq K}$. (Step 4)
end for

Similarly, for parameter v_j, its full conditional distribution becomes:

$$[v_j|-] \sim \text{Beta}\left(1 + \sum_{k=1}^K \mathbf{1}(z_k = j), \alpha_j + \sum_{k=1}^K \mathbf{1}(z_k > j))\right), \quad j = 1, \cdots, K. \tag{2.20}$$

Lastly, for the noise parameter θ^2, its full conditional distribution can be shown to be:

$$\left[\theta^2|-\right] \sim IG\left(a_\theta + \frac{N}{2}, \frac{b_\theta + \boldsymbol{S}^T\boldsymbol{S}}{2}\right), \quad \boldsymbol{S} = \boldsymbol{y} - \Lambda\tilde{\boldsymbol{X}}. \tag{2.21}$$

A complete derivation of these full conditional distributions is provided in Appendix 2.8.

Algorithm 1 presents the detailed steps of the Gibbs sampler, which combines the above full conditional steps (2.18) - (2.21) for sampling the desired posterior distribution $[\Theta|\{(x_i, y_i)\}_{i=1}^n]$. Here, all sampling steps are quite straightforward; for (2.18),

we sample from this K-point discrete distribution via inverse transform sampling, with probabilities given by normalizing the weights in (2.18) to sum to one.

2.4.2 Gibbs sampling for the multivariate HierGP

We now extend the above Gibbs sampler for the multivariate HierGP model[3], which leverages the full cumulative shrinkage prior (2.12). Suppose we obtain data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and let $\Theta = \left\{(\lambda_{\boldsymbol{k}})_{|\boldsymbol{k}|=1}^{\infty}, (v_{j,m})_{j=1}^{\infty}{}_{m=1}^{d}, \theta^2\right\}$ be the parameter set. As before, we adopt the choice of $\pi_k = \pi = IG(a_\sigma, b_\sigma)$ for the slab distribution. We employ the following truncation for $f(\cdot)$:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k} \leq \boldsymbol{K}} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{2.22}$$

where $\boldsymbol{K} = (K_1, \cdots, K_d)$ is the vector of truncation indices for each dimension. Again, these indices $K_1, \cdots, K_d$ should be set sufficiently large (further details are provided later in the appendix).

We now derive similar closed-form full conditional distributions of Θ for the multivariate HierGP. Let $\tilde{\boldsymbol{X}} = (\phi_{\boldsymbol{k}}(\boldsymbol{x}_i))_{\boldsymbol{k} \leq \boldsymbol{K}, i=1,\cdots,n} \in \mathbb{R}^{n \times \|\boldsymbol{K}\|}$ be the design matrix for the data, where $\boldsymbol{k} \in \mathbb{N}_0^d$ is a multi-index which iterates over $\boldsymbol{K}$ and $\|\boldsymbol{K}\| = \prod_{m=1}^d K_m$. Also let $\boldsymbol{z}_{\boldsymbol{k}}$ be the vector of latent random variables, defined conditionally as:

$$\mathbb{P}\left\{\boldsymbol{z}_{\boldsymbol{k}} = (l_1, \cdots, l_d) | w_{l_1,1}, \cdots, w_{l_d,d}\right\} = \prod_{m=1}^{d} v_{l_m,m} w_{l_m-1,m}, \tag{2.23}$$

where $w_{l_m,m}$ is as defined in (2.12). With this, we can leverage a similar data aug-

[3]We will refer to the multivariate HierGP model as simply "the HierGP model" from here on.

Algorithm 2: Gibbs sampling for the multivariate HierGP

Inputs: hyperparameters a_σ, b_σ, a_θ, b_θ, σ^2_∞, data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, number of iterations B.

Set initial parameters $\Theta_{[0]} = \{(\lambda^{[0]}_{\boldsymbol{k}})_{\boldsymbol{k}\leq \boldsymbol{K}}, (v^{[0]}_{j,m})_{j\leq K_m, m\leq d}, \theta^2_{[0]}\}$.
for $b = 1, \cdots, B$ **do**
  for $\boldsymbol{k}$ in $(1, \cdots, 1) : (K_1, \cdots, K_d)$ **do**
    Sample $\boldsymbol{z}^{[b]}_{\boldsymbol{k}}$ from $[\boldsymbol{z}_{\boldsymbol{k}} = \boldsymbol{l}|-]$ in (3.15) with $\lambda_{\boldsymbol{k}} = \lambda^{[b-1]}_{\boldsymbol{k}}$. (Step 1)
  end for
  for $\boldsymbol{k} = (k_1, \cdots, k_d)$ in $(1, \cdots, 1) : (K_1, \cdots, K_d)$ **do**
    if $\exists m$: $z^{[b]}_{\boldsymbol{k},m} \leq k_m$, then let $\sigma^2_{\boldsymbol{k}} = \sigma^2_\infty$; else $\sigma^2_{\boldsymbol{k}} \sim IG(a_\sigma + 0.5, b_\sigma + 0.5(\lambda^{[b-1]}_{\boldsymbol{k}})^2)$. (Step 2)
  end for
  for m in $1, \cdots, d$ **do**
    for j in $1, \cdots, K_m$ **do**
      Sample $v^{[b]}_{j,m} \sim \text{Beta}(1 + \sum_{\boldsymbol{k}\leq \boldsymbol{K}} \mathbf{1}(z^{[b]}_{\boldsymbol{k},m} = j), \alpha_j + \sum_{\boldsymbol{k}\leq \boldsymbol{K}} \mathbf{1}(z^{[b]}_{\boldsymbol{k},m} > j))$.
    end for
  end for
  Update $w_{j,m}$ and $w_{\boldsymbol{k}}$ from (2.12).
  Update $\theta^2_{[b]}$ from $IG(a_\theta + N/2, b_\theta + S^T S/2)$ where $S = \boldsymbol{y} - \Lambda_{[b-1]}\tilde{\boldsymbol{X}}$. (Step 3)
  Sample $\boldsymbol{\Lambda}_{[b]} = (\lambda^{[b]}_{\boldsymbol{k}})_{\boldsymbol{k}\leq \boldsymbol{K}} \sim N(\boldsymbol{V}\theta^{-2}_{[b]}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}_{[b]}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\boldsymbol{D} = \text{diag}(\sigma^2_{\boldsymbol{k}})_{\boldsymbol{k}\leq \boldsymbol{K}}$. (Step 4)
end for

mentation trick to derive the full conditional distribution of $\boldsymbol{z}_{\boldsymbol{k}} = [z_{\boldsymbol{k},1}, \cdots, z_{\boldsymbol{k},d}]$ as:

$$[\boldsymbol{z}_{\boldsymbol{k}} = \boldsymbol{l}|-] \propto \begin{cases} \left(\prod_{m=1}^{d} v_{l_m,m} w_{l_m-1,m}\right) \phi(\lambda_{\boldsymbol{k}}; 0, \sigma^2_\infty), & \text{otherwise,} \\ \left(\prod_{m=1}^{d} v_{l_m,m} w_{l_m-1,m}\right) t_{2a_\sigma}(\lambda_{\boldsymbol{k}}; 0, (b/a)), & l_1 > k_1, \cdots, l_d > k_d \end{cases} \tag{2.24}$$

by marginalizing out all $\sigma^2_{\boldsymbol{k}}$. Details on this marginalization are provided in Appendix 2.8.

Let us consider now the full conditional distributions for the parameter set Θ. Let $\boldsymbol{D} = diag(\sigma^2_{\boldsymbol{k}})_{\boldsymbol{k}\leq \boldsymbol{K}}$. For the coefficient vector $\boldsymbol{\Lambda} = (\lambda_{\boldsymbol{k}})_{\boldsymbol{k}\leq \boldsymbol{K}}$, its full conditional can be shown to be the same form as (2.19). For parameter v^m_j, its full conditional

distribution follows:

$$[v_{j,m}|-] \sim \text{Beta}\left(1 + \sum_{\boldsymbol{k} \leq \boldsymbol{K}} \mathbf{1}(z_{\boldsymbol{k},m} = j), \alpha_j + \sum_{\boldsymbol{k} \leq \boldsymbol{K}} \mathbf{1}(z_{\boldsymbol{k},m} > j)\right), \tag{2.25}$$
$$j = 1, \cdots, K_m, \ m = 1, \cdots, d.$$

Finally, the full conditional of θ^2 follows the same form as in (2.21). Algorithm 2 details the steps in this Gibbs sampling algorithm. As before, one can adopt a sufficiently large choice of truncation indices $\boldsymbol{K}$, or adopt an adaptive choice of $\boldsymbol{K}$; further details on the latter is provided in Appendix 2.9, following [LDD20].

We provide a brief analysis of computational complexity for the Gibbs sampler in Algorithm 2. The computation cost for Step 4 can be shown to be $\mathcal{O}(dn\|\boldsymbol{K}\| + (\|\boldsymbol{K}\|)^3)$, since it requires the computation of $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ from n observations and $\|\boldsymbol{K}\|$ bases, and a matrix inversion step for computing $\boldsymbol{V}$. The computational cost for Step 3 is $\mathcal{O}(|\boldsymbol{K}|)$ as we update $O(\|\boldsymbol{K}\|)$ variables and each update costs $O(1)$. The cost of Step 2 is $\mathcal{O}(\|\boldsymbol{K}\|^2)$, since it iterates through $\|\boldsymbol{K}\|$ variables, with each sampling step requiring $\mathcal{O}(\|\boldsymbol{K}\|)$ computation. The cost of Step 1 can similarly be shown to be $O(\|\boldsymbol{K}\|^2)$. Combining this, we thus have a computational complexity of $O\{B(dn\|\boldsymbol{K}\| + \|\boldsymbol{K}\|^3)\}$ for this Gibbs sampler, where B is the number of Gibbs iterations.

2.4.3 Gibbs sampling for HierGP$_2$

In situations in which one expects sparsity of basis coefficients but does not have strong prior belief of effect hierarchy or heredity, the alternative HierGP$_2$ model in Section 2.3.1 (which makes use of GL shrinkage priors) may be an appealing alternative. We present next an analogous Gibbs sampler for this alternate model with Φ taken as the horseshoe priors in [CPS10]. Here, the posterior sampling of the parameter set $\Theta = \{(\lambda_k)_{k=1}^{\infty}, \theta^2\}$ reduces to the same setting as Bayesian linear regression

Algorithm 3: Blocked Metropolis-within-Gibbs sampler for the HierGP_2

Inputs: hyperparameters a_θ, b_θ, τ, data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, number of iterations B.

1: Set initial parameters $\Theta_{[0]} = \{(\lambda_k^{[0]})_{k=1}^K, (\theta^{[0]})^2\}$.
2: **for** $b = 1, \cdots, B$ **do**
3: Sample $\boldsymbol{\Lambda}^{[b]} = (\lambda_1^{[b]}, \cdots, \lambda_K^{[b]})$ from $N(\boldsymbol{V}(\theta^{[b-1]})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}, \boldsymbol{V})$, $V = (\boldsymbol{D}^{-1} + (\theta^{[b-1]})^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\boldsymbol{D} = \text{diag}((\sigma_1^{[b-1]})^2, \cdots, (\sigma_K^{[b-1]})^2)$. (Step 1)
4: Sample $\{\sigma_k^{[b]}\}_{k=1}^K$ via Metropolis-Hastings from the density proportional to $\prod_{k=1}^K \left(\frac{1}{1+\sigma_k}\right) \exp(\frac{-(\lambda_k^{[b]})^2\tau\sigma_k}{2(\theta^{[b]})^2})$. (Step 2)
5: Update $(\theta^{[b]})^2$ from $IG(a_\theta + N/2, (b_\theta + \boldsymbol{S}^T\boldsymbol{S})/2)$ where $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}^{[b]}\tilde{\boldsymbol{X}}$. (Step 3)
6: **end for**

with horseshoe priors, for which there are existing posterior sampling algorithms. Algorithm 3 provides a direct extension of the blocked Metropolis-within-Gibbs sampler in [JOB20] for the HierGP_2 model.

2.5 Consistency results

We now present several theoretical results which confirm consistency for both the (multivariate) HierGP and HierGP_2. These results extend existing theory on consistency for high dimensional linear regression, specifically results in [SL17] and [CS07]. As mentioned before, for the HierGP, we will adopt point masses $\pi_k = \delta_{\sigma^2_{\boldsymbol{k},0}}$ for the slab distributions, where $\{\sigma^2_{\boldsymbol{k}}\}$ is pre-specified; this is commonly used in the literature for theoretical analysis. We further show a Bernstein-von-Mises-type theorem for the shape approximation of the HierGP_2 model with horseshoe priors, which extends theory from [SL17] to the current GP setting. Main theorems and corresponding assumptions are presented next, with proofs deferred to the Appendix.

2.5.1 The HierGP model

To prove consistency for the proposed HierGP model in Section 2.3, we will require some of the following assumptions:

(A1) *Basis representation*: Let $f_0(\cdot)$ denote the true function we wish to predict. We assume that f_0 takes the form

$$f_0(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in S} \lambda_{\boldsymbol{k}}^0 \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{2.26}$$

where $S \subset \mathbb{N}_0^d$ is a finite index set of size $s = \text{card}(S)$. This further implies that $f_0 \in L^2(\mathcal{X}) \bigcap C^1(\mathcal{X})$.

(A2) *Smoothness*: We further assume $a_{\boldsymbol{k}} = \sup_{\boldsymbol{x} \in \mathcal{X}} |\phi_{\boldsymbol{k}}(\boldsymbol{x})| < \infty$, $b_{\boldsymbol{k}} = \sup_{\boldsymbol{x} \in \mathcal{X}} |\phi'_{\boldsymbol{k}}(\boldsymbol{x})| < \infty$, $\sum_{\boldsymbol{k}} a_{\boldsymbol{k}} \sigma_{\boldsymbol{k},0} < \infty$ and $\sum_{\boldsymbol{k}} b_{\boldsymbol{k}} \sigma_{\boldsymbol{k},0} < \infty$. These assumptions control the smoothness of the underlying function $f_0(\boldsymbol{x})$.

(A3) *Random design*: The design points $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ are sampled i.i.d. from some probability distribution P_0 on $\mathcal{X} = [0,1]^d$. We will use this in Theorem 2.5.2.

(A4) *Fixed design*: The design points $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ satisfy the following condition. For each hypercube $\mathcal{H}$ in $\mathcal{X} = [0,1]^d$, there exists a constant $0 < K_d \leq 1$ such that, whenever its Lebesgue measure $\lambda(\mathcal{H}) \geq (K_d n)^{-1}$, the hypercube $\mathcal{H}$ contains at least one design point. This can be seen as an extension of the balance condition for digital net sampling in Quasi-Monte Carlo (see, e.g., Owe97; DKS13), which ensures the design points are uniformly spaced out over $\mathcal{X}$. This assumption will be used in Theorem 2.5.1.

(A5) *Prior modification*: In the fixed design setting, we will also require a small modification on the prior to prove L_1-consistency. Let Π_1^* denote the prior for the HierGP model in Section 2.3, and let V be a constant satisfying $V >$

$\sup_{[0,1]^d} |f_0'|_\infty$. Define $\Omega = \{f : |f'|_\infty < V\}$. We can then define the modified prior model as $\Pi_2^*(\cdot) = \Pi_1^*(\cdot \bigcap \Omega)/\Pi_1^*(\Omega)$ with $\Pi_1^*(\Omega) > 0$. This is a standard modification used for proving consistency in high-dimensional Bayesian linear regression (see, e.g., [CS07]).

Next, we introduce several types of neighborhoods for proving consistency, two for random designs and the other two for fixed designs. Here, let σ_0^2 denote the true noise variance for observations.

1. *Hellinger neighborhood*:

$$H_\epsilon = \{(f, \sigma) | d_H(p, p_0) < \epsilon\}$$

where d_H is the Hellinger distance $d_H(p_1, p_2) = \int (\sqrt{p_1} - \sqrt{p_2})^2 d\xi$. Here, p is the probability density function of $(\boldsymbol{x}, y)$ with respect to $\xi = P_0 \times \lambda$, namely $p(x, y) = \phi([y - f(x)]/\sigma)/\sigma$ where ϕ is the standard normal density and λ is the Lebesgue measure.

2. *Empirical measure neighborhood*:

$$W_{\epsilon,n} = \left\{(f, \sigma) : \int |f(\boldsymbol{x}) - f_0(\boldsymbol{x})| \, dP_n(\boldsymbol{x}) < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}$$

where $P_n(x) = n^{-1} \sum_{i=1}^n I_{x_i}(x)$ is the empirical measure.

3. *L_1-neighborhood*:

$$L_\epsilon = \left\{(f, \sigma) : \int |f(x) - f_0(x)| dx < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}$$

4. *P_0-neighborhood*:

$$U_\epsilon = \left\{(f, \sigma) : d_{P_0}(f_0, f) < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}$$

where $d_{P_0}(f, g) = \inf\{\epsilon : P_0(\{\boldsymbol{x} : \ |f(\boldsymbol{x}) - g(\boldsymbol{x})|_\infty > \epsilon\}) < \epsilon\}$.

We can now prove consistency results, first for fixed designs then for random designs.

Theorem 2.5.1 (Consistency of HierGP, fixed design)**.** *Let f follow the modified HierGP prior Π_2^* above. Let Q_0 be the conditional distribution of the data $\{y_i\}_{i=1}^n$ given fixed design points $\{\boldsymbol{x}_i\}_{i=1}^n$. Suppose Assumptions* (A1)*,* (A2)*,* (A4)*, and* (A5) *hold. Then:*

(a) For any $\epsilon > 0$, we have:

$$\Pi_2^*\left((f,\sigma) \in W_{\epsilon,n}^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n\right) \to 0, \quad [Q_0] \text{ almost surely.} \tag{2.27}$$

(b) We also have:

$$\Pi_2^*(L_\epsilon^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n) \to 0, \quad [Q_0] \text{ almost surely.} \tag{2.28}$$

Theorem 2.5.2 (Consistency of HierGP, random design)**.** *Let f follow the HierGP prior Π_1^* in Section 2.3. Let Q_0 denote the joint distribution of $\{\boldsymbol{x}_n, y_n\}_{n=1}^\infty$. Suppose Assumptions* (A1)*,* (A2) *and* (A3) *hold. Then, for any random design following measure P_0, we have*

$$\Pi_1^*(U_\epsilon^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n) \to 0, \quad [Q_0] \text{ almost surely,} \tag{2.29}$$

and

$$\Pi_1^*(H_\epsilon^c | \{y_i, \boldsymbol{x}_i\}_{i=1}^n) \to 0, \quad [Q_0] \text{ almost surely.} \tag{2.30}$$

Theorems 2.5.1 and 2.5.2 show that, under regularity conditions, the posterior distribution of the regression function and variance parameter indeed converge to the truth under appropriate topologies, as sample size goes to infinity. This establishes posterior consistency of the proposed model, and guarantees that the employed

shrinkage prior provides enough support for predicting the class of functions outlined in Assumption (A1). The proofs of these theorems extend results in [CS07]; details are provided in the Appendix.

We provide some insight on why only consistency is shown for the HierGP model. While there exists a rich literature on contraction rates for standard shrinkage priors, the employed hierarchical cumulative shrinkage prior in the HierGP is quite new. Standard analysis tools for high-dimensional Bayesian linear regression (see, e.g., CSHVdV15, SL17, JG20) are not suitable in this setting, due to the highly structured nature of our shrinkage prior. We thus focus on establishing consistency of our model for this novel prior setting, and defer the more complex question of contraction rates to future work.

2.5.2 The HierGP$_2$ model

We now present some results for the HierGP$_2$ from Section 2.3.1 and 2.4.3. We let $(\boldsymbol{\lambda^0}, \theta_0) = ((\lambda_1^0, \cdots, \lambda_K^0), \theta_0)$ be the true coefficients and variance of the model, and $\boldsymbol{\lambda}_{S'}, \tilde{\boldsymbol{X}}_{S'}$ denote the coefficients and design matrix for the index subset $S' \in (1, \cdots, K)$. Here, we require the following assumptions:

(B1) *Truncation level*: Let $K_{[n]}$ be the truncation level employed given a sample size of n. We assume that $K_{[n]} \geq n$, i.e., the truncation level is always larger than the sample size.

(B2) *Design*: There exists constant integer C (depending on n and $K_{[n]}$) and scalar λ_0, such that $C > s$, and $\lambda_{\min}(\tilde{\boldsymbol{X}}_{S'}^T \tilde{\boldsymbol{X}}_{S'}) \geq n\lambda_0$ for any subset $S' \subset \{1, \cdots, K_{[n]}\}$ where $|S'| \leq C$, and $s = \text{card}(S)$ is the size of the active index set S for the true model.

(B3) *Sparsity*: We further assume that $s \log(K_{[n]}) \leq n$. This can be seen as a sparsity

assumption on the true function f_0.

(B4) *Coefficient magnitude*: Suppose $\max|\lambda_k^0/\theta_0| \leq \gamma_3 E_n$ for some fixed $\gamma_3 \in (0,1)$, where E_n is non-decreasing with respect to n. This provides an upper bound on the magnitude of basis coefficients.

With this in hand, we can then show the following contraction and Bernstein-von-Mises (BvM) shape approximation result for the HierGP$_2$ model.

Theorem 2.5.3 (Contraction & BvM theorem for HierGP$_2$)**.** *Let f follow the HierGP$_2$ prior Π in Section 2.3.1. Let P^* be the joint measure of $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ under the true function $f_0(\cdot)$ with noise variance θ_0^2. Suppose Assumptions* (B1)-(B5) *hold. Then:*

(a) (Posterior contraction) *We have*

$$P^*\left(\Pi\left(||f-f_0||_2 \geq c_1\theta_0\epsilon_n|\{\boldsymbol{x}_i,y_i\}_{i=1}^n\right) \geq e^{-c_2 n\epsilon_n^2}\right) \leq e^{-c3n\epsilon_n^2}, \tag{2.31}$$

for some positive constants c_1, c_2 and c_3, where $\epsilon_n = M\sqrt{s\log(K_{[n]})/n}$ is the contraction rate and M is a fixed constant.

(b) (BvM shape approximation) *The posterior distribution $\Pi([f-f_0](\boldsymbol{x}), \theta^2|\{\boldsymbol{x}_i, y_i\}_{i=1}^n)$ converges in total variation to the distribution*

$$\left[\sum_{k=1}^{K_{[n]}} \lambda_k\phi_k(\boldsymbol{x})\Big|\theta^2\right][\theta^2], \tag{2.32}$$

where

$$\begin{aligned}[\{\lambda_k\}_{k=1}^{K_{[n]}}|\theta^2] &= \phi\left(0; \hat{\lambda}_S, \theta^2(\tilde{X}_S^T\tilde{X}_S)^{-1}\right)\prod_{k\notin S}\pi(\lambda_k|\theta^2),\\ [\theta^2] &\sim IG\left(\frac{n-s}{2}, \frac{\hat{\theta}^2(n-s)}{2}\right),\end{aligned} \tag{2.33}$$

$\pi(\lambda_k|\theta^2)$ *is the conditional prior distribution in* (2.13)*, and* $\hat{\lambda}_S$ *and* $\hat{\theta}^2$ *are the MLEs of* λ_S *and* θ^2 *given data* $\{\tilde{\boldsymbol{X}}_S, y_i\}_{i=1}^n$.

Theorem 2.5.3 implies that, under regularity conditions, the posterior distribution of the regression function and variance parameter concentrates around the truth at a rate of ϵ_n, and is asymptotically normal under proper topologies as sample size goes to infinity. The latter is known as a "Bernstein-von-Mises" type theorem. Such results are stronger than the earlier posterior consistency results for the HierGP, since they provide not only an explicit posterior contraction rate, but also shed light on an adaptive choice of truncation level which varies by sample size. The proof of this theorem extends results in [SL17].

2.6 Numerical Experiments

We now explore the proposed HierGP and HierGP$_2$ in a suite of numerical experiments. We first investigate their performance for computer code emulation, and then demonstrate effectiveness for recovery of dynamical systems.

2.6.1 Computer code emulation

For our experiments on *computer code emulation*, we will consider a suite of test functions and compare the proposed models (HierGP and HierGP$_2$) with several popular and/or related GP-based emulators. This includes the standard GP emulator with Matérn-3/2 kernel [Ste99]; the additive GP model in [LBH22], which builds off of recent work [DNR11] on leveraging additive low-dimensional structure; the "least-squares" model, which makes use of a least-squares fit of the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ using a pre-specified basis matrix $\tilde{\boldsymbol{X}}$; and the sparse least-squares fit, which makes use l_1-regularized estimates under the same set-up (with penalty parameters tuned via cross-validation). The latter two are akin to surrogate models used in polynomial chaos; see

[LMS21] for a comprehensive review. Both models provide useful benchmarks for the HierGP in highlighting potential advantages of embedding effect sparsity, hierarchy and heredity within the GP. For a fair comparison, both models make use of the same basis matrix as the HierGP.

Consider first the simple setting where test functions are simulated from the HierGP prior in Section 2.4.2, in $d = 2$ and $d = 3$ dimensions. This simulation is performed with parameters $\alpha = 6, a_\sigma = 1, b_\sigma = 1, \sigma_\infty = 0$, a truncation limit of $\boldsymbol{K} = (8, 8)$ and $(4, 4, 4)$ for $d = 2$ and $d = 3$, respectively, and sinusoidal basis $\phi_{\boldsymbol{k}}(\boldsymbol{x}) = \Pi_{m=1}^{d} \sin(2\pi k_m x_m)$. The simulated functions thus capture the presumed effect sparsity, heredity, and hierarchy principles. For model training, we use $n = 70$ uniformly-sampled design points. For the HierGP, least-squares, and sparse least-squares, we assume the perfectly-specified setting where the basis matrix and truncation levels are set to be the same as the simulation model; we will explore a misspecified setting next. This simulation is then replicated 50 times to measure error variability.

Figures 2.1 and 2.2 show boxplots of the prediction errors for 400 uniformly-sampled testing points, for each approach in $d = 2$ and 3 dimensions. We see that the HierGP yields improved predictions over competing models. This is not surprising – when structured sparsity is present in f, the HierGP successfully leverages such structure for improved predictions. The HierGP$_2$ (which adopts a less structured horseshoe prior on basis coefficients) also does quite well, but has slightly higher errors since it does not integrate prior information on effect heredity and hierarchy. The standard Matérn GP and the additive GP both yield significantly worse predictive performance; this is intuitive since such models do not embed structured sparsity in f. Finally, the two least-squares fits (even with a perfectly specified basis) also yield poor performance, which is expected since such fits also do not capture the desired

structured sparsity, despite having a perfectly-specified basis.

Table 2.1 compares the uncertainty quantification performance of these models, by reporting the empirical coverage rates of 95% posterior predictive credible intervals and corresponding average predictive interval widths. Here, we compared only the two HierGP models with the standard GP, as the other methods are not fully probabilistic. We see that, for both $d = 2$ and $d = 3$, the empirical coverage rates for the HierGP models are noticeably higher than that for the standard GP, which can dip below the nominal 95% rate. Furthermore, the credible interval widths for the HierGP models are significantly smaller than that for the standard GP. This suggests that, when f has the presumed hierarchical sparsity structure, the proposed models can indeed leverage such structure to provide more *precise* probabilistic predictions with improved coverage over standard GPs, as desired.

Consider next the emulation of the two synthetic test functions in the literature, the Branin function [SFK08]:

$$f(\mathbf{x}) = a(x_2 - bx_1 + cx_1 - r)^2 + s(1-t)\cos(x_1) + s,$$

where $a = 1$, $b = 5.1/(4\pi^2)$, $c = 5/\pi$, $r = 6$, $s = 10$ and $t = 1/\pi$, and the Cheng & Sandu function [CS10]:

$$f(\mathbf{x}) = \cos(x_1 + x_2)\exp(x_1 x_2).$$

These two functions vary in their degree of adherence to the effect heredity and hierarchy principles. Here, the HierGP, least-squares, and sparse least-squares methods make use of the above sinusoidal basis with truncation limit $\boldsymbol{K} = (8, 8)$. This provides a good test for how robust the proposed models are when there are minor violations of the effect principles with respect to the chosen basis functions. As before, $n = 70$ uniformly-sampled design points are used for training.

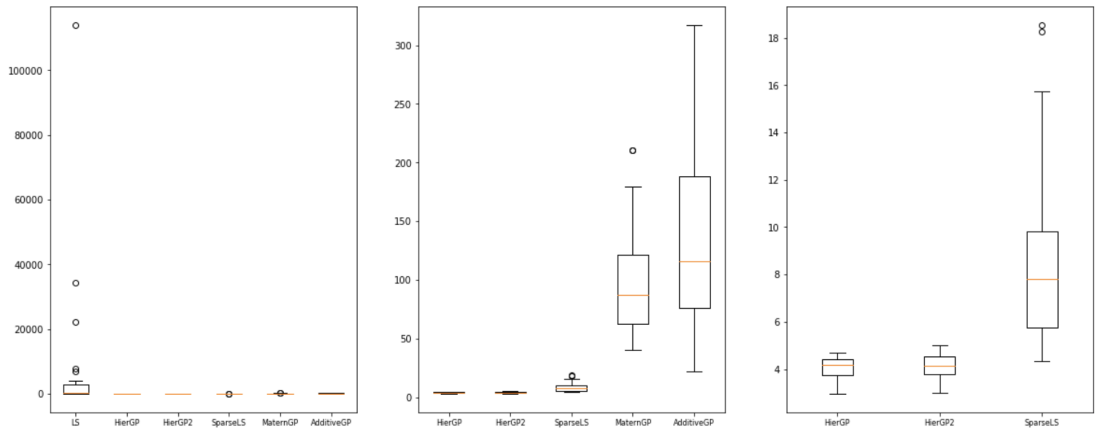

Figure 2.1: Prediction error boxplots for $d = 2$ dimensions. From left to right are zoom-in versions of the boxplots.

Figure 2.3 shows boxplots of prediction errors for 400 uniformly-sampled testing points for each test function. For the Branin function, the HierGP provides the best predictive performance of all the considered models, with the HierGP$_2$ a close competitor. Upon further inspection, this is not surprising since its functional form suggests some form of effect heredity and hierarchy is present. For the Cheng & Sandu function, we see that the HierGP$_2$ provides the best predictive performance, with the HierGP a close competitor. This can be explained by the more complex interaction structure present in its functional form, which may be difficult to identify with the presumed structured sparsity in the HierGP. Regardless, the above experiments suggest that, when effect hierarchy and heredity are present in f (even with minor violations), the proposed models can learn and integrate such structure for improved predictive performance.

2.6.2 Recovery of dynamical systems

We now further investigate the HierGP for the problem of dynamical system recovery and prediction, which is widely used in climatology, ecology and finance (see, e.g.,

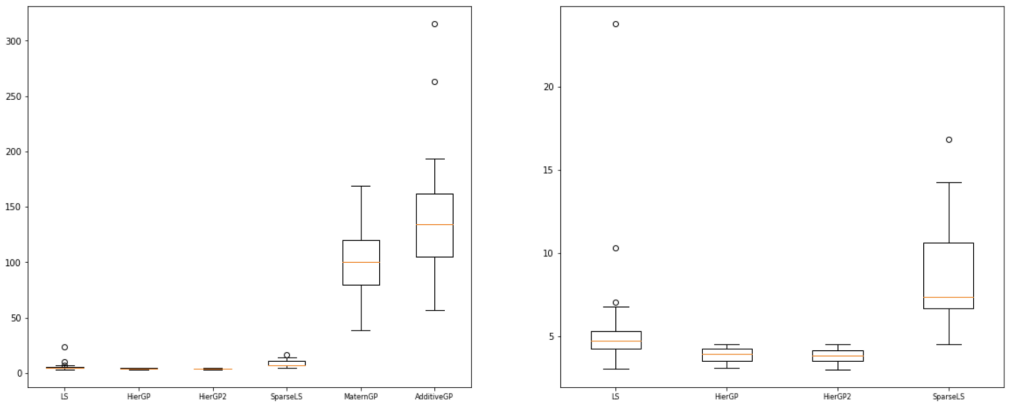

Figure 2.2: Prediction error boxplots $d = 3$ dimensions. From left to right are zoom-in versions of the boxplots.

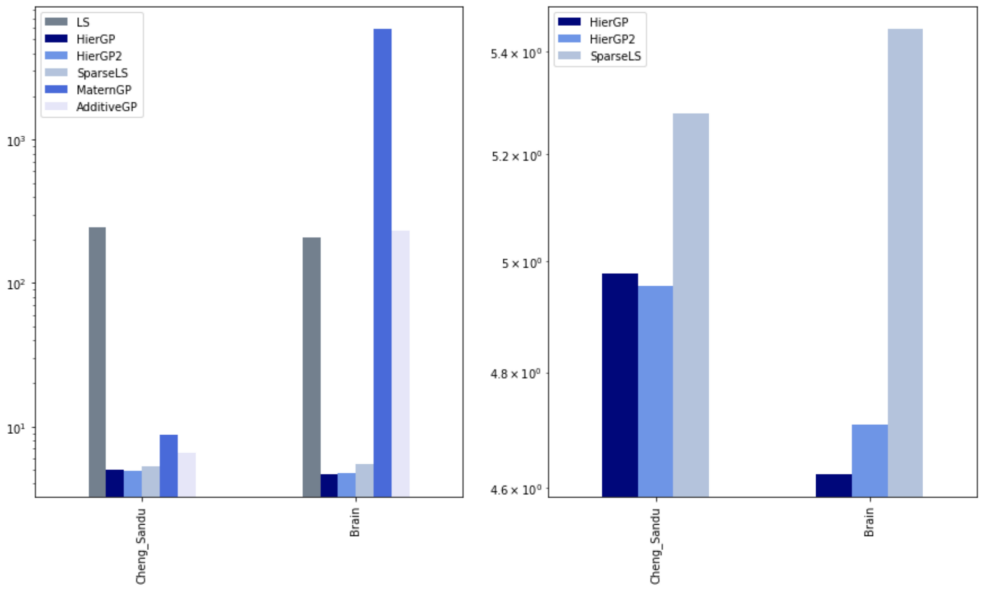

Figure 2.3: Barplots of mean prediction error the Branin and Cheng & Sandu function. From left to right are zoom-in versions of the barplots.

Table 2.1: Empirical coverage rates for 95% posterior predictive credible intervals and corresponding average interval widths.

	HierGP	**HierGP2**	**MatérnGP**
Empirical coverage rate ($d = 2$)	100.0%	100.0%	91.0%
Empirical coverage rate ($d = 3$)	97.7%	99.1%	92.2%
Average credible interval width ($d = 2$)	1.721	1.313	4.062
Average credible interval width ($d = 3$)	1.277	1.798	7.250

GE22, LOT$^+$11, Mud19). We first provide a brief review of this problem, following [BPK16]. We consider here dynamical systems [GH13] which take the form:

$$\frac{d}{dx}\boldsymbol{x}(t) := \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)), \tag{2.34}$$

Here, $\boldsymbol{x}(t) \in \mathbb{R}^q$ denotes the states of the system at time t, and $\boldsymbol{f}(\boldsymbol{x}(t))$ captures dynamical constraints which govern the equations of motion for system states. The formulation (2.34) covers a broad range of dynamical systems used in ecology, biology, and other scientific disciplines [DK22].

Consider now the setting where data $\{(\boldsymbol{x}(t_i), \dot{\boldsymbol{x}}(t_i)\}_{i=1}^n$ are observed on the system states, where $t_1, \cdots, t_n$ are the sampled time points. We can rearrange this into the state matrices:

$$\boldsymbol{X} = (\boldsymbol{x}^T(t_1), \cdots, \boldsymbol{x}^T(t_n))^T = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_d(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_d(t_2) \\ \cdots & \cdots & \ddots & \vdots \\ x_1(t_n) & x_2(t_n) & \cdots & x_d(t_n) \end{pmatrix} \in \mathbb{R}^{n\times d},$$

$$\dot{\boldsymbol{X}} = (\dot{\boldsymbol{x}}^T(t_1), \cdots, \dot{\boldsymbol{x}}^T(t_n))^T = \begin{pmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_d(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_d(t_2) \\ \cdots & \cdots & \ddots & \vdots \\ \dot{x}_1(t_n) & \dot{x}_2(t_n) & \cdots & \dot{x}_d(t_n) \end{pmatrix} \in \mathbb{R}^{n\times d}.$$

With this, we will then construct a "library" of candidate functions for recovering the function $\boldsymbol{f}$ in (2.34), thus recovering the underlying system dynamics. Suppose

these candidate functions $\mathcal{F} = \{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{\boldsymbol{k}}$ are parametrized by the multi-index $\boldsymbol{k} = (k_1, \cdots, k_d)$, $\boldsymbol{k} \leq \boldsymbol{K}$. Given the sampled time points, this library can be represented by the model matrix:

$$\Phi(\boldsymbol{X}) = \begin{pmatrix} | & \cdots & | & \cdots & | \\ \Phi(\boldsymbol{X})^{[\mathbf{1}]} & \cdots & \Phi(\boldsymbol{X})^{[\boldsymbol{k}]} & \cdots & \Phi(\boldsymbol{X})^{[\boldsymbol{K}]} \\ | & \cdots & | & \cdots & | \end{pmatrix} \in \mathbb{R}^{n \times \|\boldsymbol{K}\|},$$

where $\|\boldsymbol{K}\| = \prod_{m=1}^{d} K_m$ is the total number of basis functions in $\mathcal{F}$, and $\Phi(\boldsymbol{X})^{[\boldsymbol{k}]}$ is the model matrix consisting of the basis functions in $\mathcal{F}$ with multi-index $\boldsymbol{k}$.

The dynamical system (2.34) can then be represented by the linear system of equations:

$$\dot{\boldsymbol{X}} = \Theta(\boldsymbol{X})\boldsymbol{\Xi} \tag{2.35}$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_d) \in \mathbb{R}^{\|\boldsymbol{K}\| \times d}$ is the matrix of coefficients for $\boldsymbol{f}$, and $\boldsymbol{\xi}_m \in \mathbb{R}^{\|\boldsymbol{K}\|}$ is the coefficient vector for the m-th component of $\boldsymbol{f}$.

Given data matrices $\dot{\boldsymbol{X}}$ and $\boldsymbol{X}$, the goal of recovering $\boldsymbol{f}$ can be viewed as a regression problem on estimating the coefficient matrix $\boldsymbol{\Xi}$. In a seminal paper, [BPK16] argued that, since for many physical systems there are only a few dominant terms that govern the underlying dynamics, the coefficients in $\boldsymbol{\Xi}$ should be estimated in a sparse manner. To achieve this, they proposed a method called Sparse Identification of Nonlinear Dynamics (SINDy), which makes use of compressed sensing algorithms for sparse estimation of $\boldsymbol{\Xi}$, thus allowing for a sparse identification of the system $\boldsymbol{f}$. Since then, there has been further developments on SINDy via sparse regression and deep learning; see [CZA+20] and [BCSK21].

A potential limitation with the above SINDy-based methods is that, as mentioned in section 2.1, sparsity in physical systems is often structured via the principles of effect hierarchy and heredity [HW92]: main effects typically have greater influence

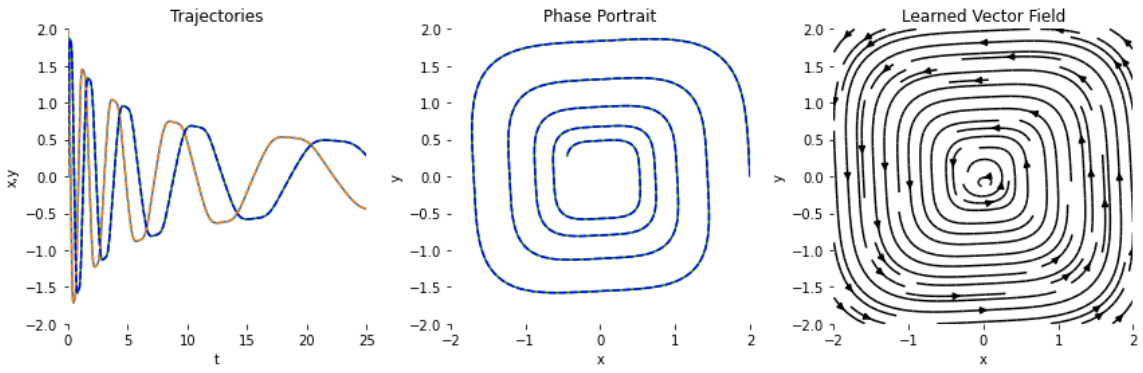

Figure 2.4: Visualizing the true and predicted dynamical system trajectories from the 2D cubic system.

than interactions, and interactions are only present when component main effects are present. One way to capture such structure is to assign the proposed hierarchical cumulative priors (2.11) and (2.12) independently over each row of the coefficient matrix $\boldsymbol{\Xi}$. With these priors, the resulting linear system (2.35) can be viewed as fitting m independent HierGP models, with basis functions taken from the function library $\mathcal{F}$. The *recovery* of governing equations can thus be performed via posterior sampling of the coefficient matrix $\boldsymbol{\Xi}$ given data $\{(\boldsymbol{x}(t_i), \dot{\boldsymbol{x}}(t_i)\}_{i=1}^n$, using the Gibbs sampler in Section 2.4.2. With posterior samples $\{\boldsymbol{\Xi}_1, \cdots, \boldsymbol{\Xi}_B\}$ generated, one can then *predict* and *quantify uncertainty* on the dynamical system via forward solves of (2.34) using each coefficient matrix sample $\boldsymbol{\Xi}_b$, $b = 1, \cdots, B$.

In problems when structured sparsity exists in the governing equations, it is intuitive to expect that the integration of such structure within the HierGP can yield improved dynamical system recovery with greater certainty, particularly with limited data. We explore this below in numerical comparisons with existing methods on two dynamical systems.

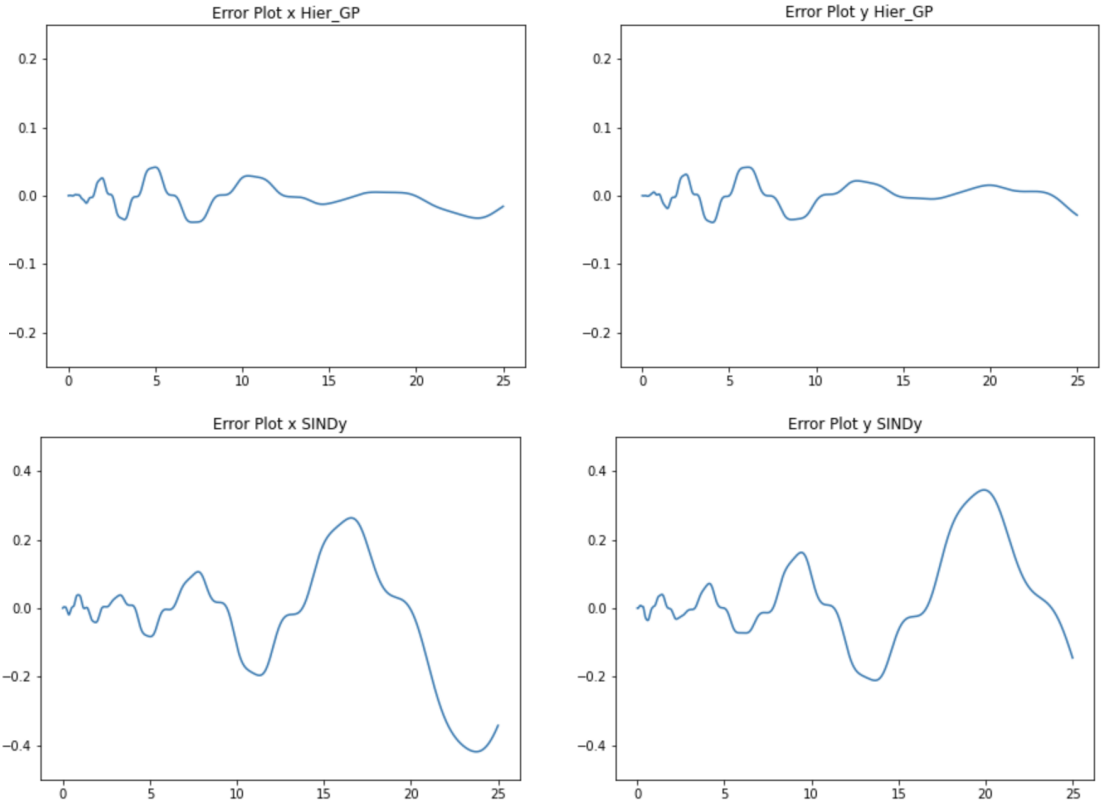

Figure 2.5: Prediction errors (in x and y coordinates) of the HierGP (top) and SINDy (bottom) for the 2D cubic system with 500 time-steps.

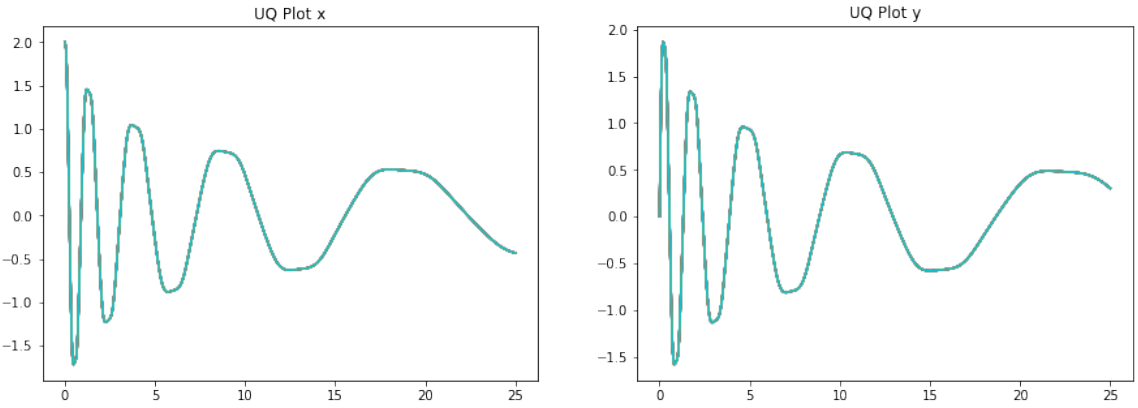

Figure 2.6: Visualizing the forward runs of 50 posterior sample draws of the HierGP for the 2D Lorenz System in x-y-z coordinates (from left to right).

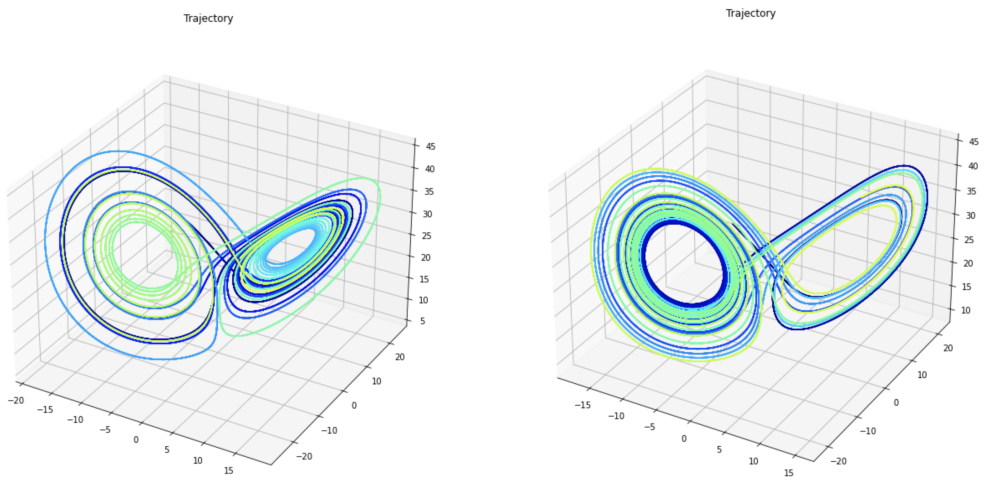

Figure 2.7: Visualizing the true (right) and recovered (left) 3D Lorenz system using the HierGP.

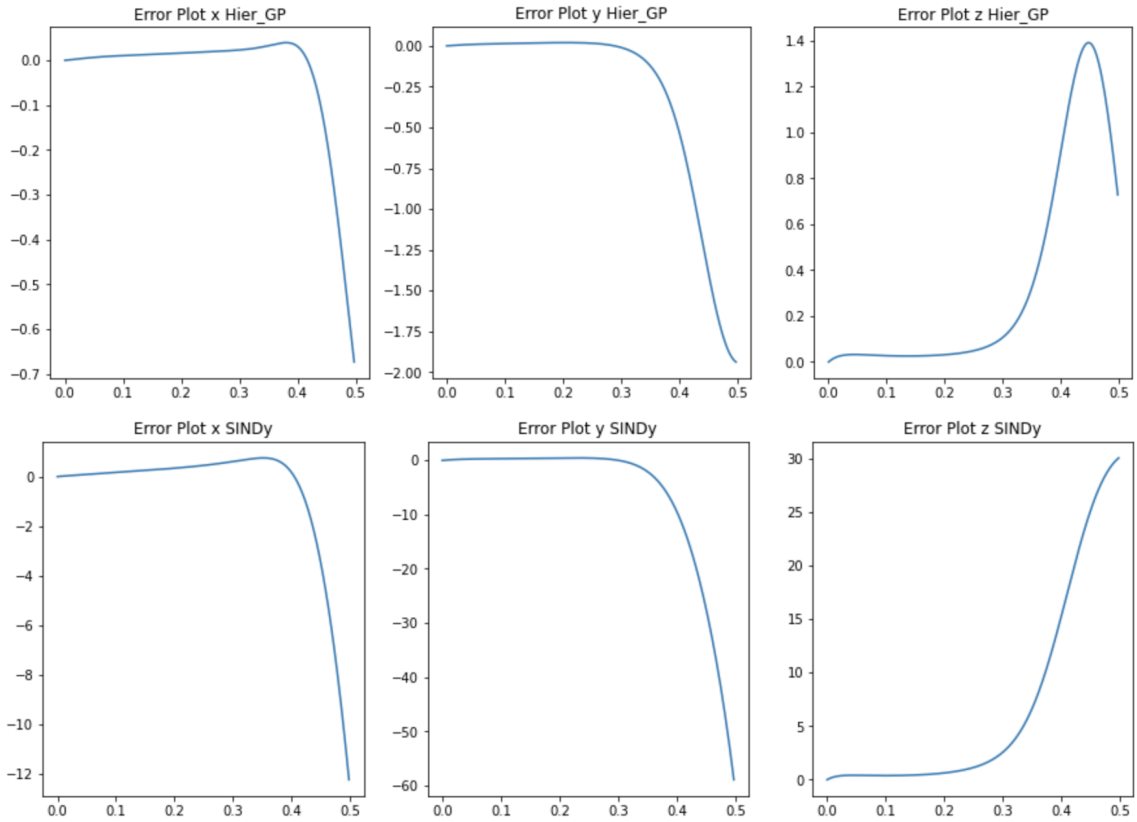

Figure 2.8: Prediction errors (in x-, y- and z-coordinates) of the HierGP (top) and SINDy (bottom) for the 3D Lorenz system.

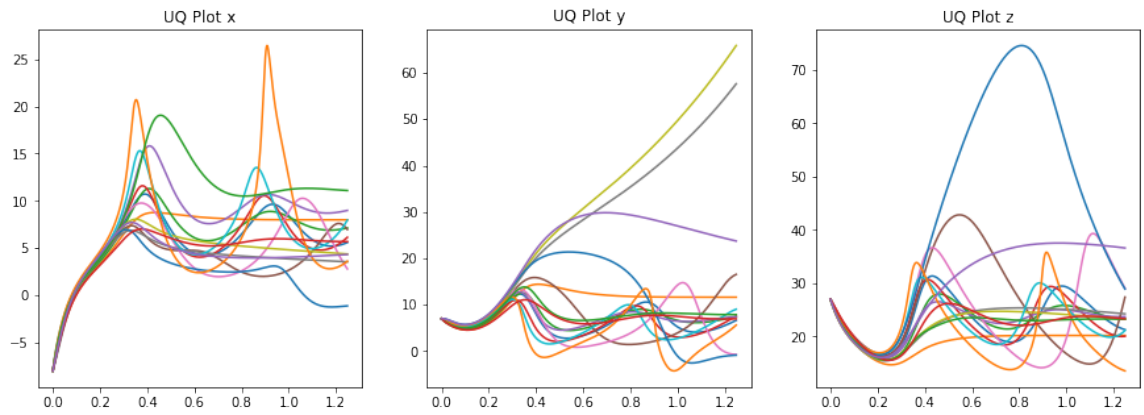

Figure 2.9: Visualizing the forward runs of 50 posterior sample draws of the HierGP for the 3D Lorenz system in x-y-z coordinates (from left to right).

2D cubic equations

Consider the following two-dimensional planar dynamical system (see [BPK16] for further details):

$$\begin{aligned} \frac{dx(t)}{dt} &= -ax(t)^3 + by(t)^3, \\ \frac{dy(t)}{dt} &= -bx(t)^3 + ax(t)^3, \end{aligned} \tag{2.36}$$

where a and b are constant parameters. While this system is simple, it has two appealing features which allow interesting comparison of recovery methods. First, its derivative functions capture effect sparsity and hierarchy as they depend on only two basis functions of relatively low order. Secondly, one can show for any initial point $(x(0), y(0))$, the system will always converge to a stationary point $(0, 0)$ with quasi-periodic behavior (BPK16), thus providing stability and predictability to numerical solutions. In the following experiment, we set the true parameters as $a = 0.1$ and $b = 2$, with initial condition $(x_0, y_0) = (2, 0)$. We then generate the training data by numerically solving the dynamical system (2.36), then sampling $n = 500$ observations (with a time step of 0.04) from one trajectory corrupted with i.i.d. Gaussian noise

(with variance 0.01). Finally, we set $\boldsymbol{K} = (5, 5)$ for the HierGP.

Figure 2.4 shows the trajectory of the true dynamical system forward simulations of recovered systems from SINDy and the proposed HierGP. For the latter, we first performed posterior sampling on model parameters Θ, then used its posterior mean as parameters for a forward solve of the system (2.36). Visually, we see that the HierGP trajectory captures well the desired periodic and asymptotic behavior of the cubic system. Figure 2.5 shows the corresponding prediction error of the recovered systems for the HierGP and SINDy in each of the two coordinates. We see that the HierGP indeed yields noticeably improved predictions over SINDy; this shows that when such structured sparsity is present in the governing equations, integrating such structure within the prior specification indeed allows for improved system recovery. The errors for both methods are relatively small, which is unsurprising since the true dynamical system is quite simple. These errors do grow slightly with time; this is again intuitive since recovery errors should propagate in time given estimation errors for dynamical system coefficients.

Figure 2.6 further explores the uncertainty quantification of the proposed method, by showing the forward runs of 50 posterior sample draws for Θ for the 2D system (2.36) in x- and y-coordinates. The existing SINDy method [BPK16] does not provide such a quantification of uncertainty. We see the recovered trajectory from the HierGP not only recovers the true system well but also does so with relatively high certainty. We do note that the posterior uncertainty from our model grows gradually in time; this is not surprising, since it is known that uncertainties in model parameters accumulate over time for such systems.

Chaotic Lorenz system

Consider next the following three-dimensional chaotic Lorenz system [Lor63], given by:

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma(y(t) - x(t)), \\ \frac{dy(t)}{dt} &= x(t)(\rho - z(t)) - y(t), \\ \frac{dz(t)}{dt} &= x(t)y(t) - \beta z(t), \end{aligned} \tag{2.37}$$

where σ, β and ρ are constant parameters. Although these equations have rich and chaotic dynamics that evolve on a strange attractor [BPK16], the modeled equations for each derivative are sparse and of relatively low order, thus satisfying the three effect principles. In particular, the derivative functions are typically influenced by only a few low-order terms which have sparse and hierarchical structure [BPK16]. In the following, we set the true parameters as $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$ with initial conditions $(x(0), y(0), z(0)) = (-8, 7, 27)$. As before, the training data are generated by solving the dynamical system (2.37) and then sampling $n = 200$ (with time step 0.05 unit time) observations corrupted with Gaussian noise (with variance 0.01) from the resulting solution with $\boldsymbol{K} = (5, 5, 5)$.

Figure 2.7 shows the trajectory of the true Lorenz system, as well as the trajectories of the recovered system from both SINDy and the proposed HierGP. Visually, we see that both recovered systems capture the desired strange attractor behavior and short-time dynamics: all trajectories move locally and predictably initially, but more globally and chaotically as time progresses, constrained within the region with complex geometric structure [Lor63]. Figure 2.8 shows the corresponding prediction errors of the recovered system in each of the three coordinates. We again see that the HierGP yields noticeably improved performance over SINDy, which again supports

the fact that when structured sparsity is present, the integration of such structure can indeed aid in more accurate system recovery. We note that the errors grow rapidly as time progresses, which suggests recovery becomes increasingly difficult over time; this is not too surprising given the chaotic nature of this system.

Figure 2.9 further investigates the uncertainty quantification for the HierGP, by showing the forward runs of 50 posterior sample draws on Θ for the 3D Lorenz system in $x/y/z$ -coordinates. Initially, we see the recovered trajectory using the HierGP has little uncertainty, with all sample paths quite close to each other. However, as time progresses, we see much greater posterior uncertainty, with sample paths growing further apart as uncertainty accumulates over time. This is not surprising given the chaotic nature of the system and its error propagation over time, and again suggests that long-term prediction of such systems is a challenging problem.

2.7 Technical Proofs

2.7.1 Proof of Theorem 2.5.1

Proof. For notational simplicity, we will prove this result first for the *univariate* HierGP, then extend this argument for the multivariate HierGP. We use the following notation below: θ denotes (f, σ) with $\theta_0 = (f_0, \sigma_0)$. The density $f_i(x, \sigma)$ is the normal density with mean $f(x_i)$ and variance σ^2. The parameter space Θ is a product space of a function space Θ_1 and $\mathbb{R}^+$. the prior on θ is the product measure $\Pi = \Pi_1 \times \Pi_2$.

The theorem is analogous to Theorem 4 in [CS07] for our prior, and it is enough to check the conditions in Theorem 1 of the same paper for the prior Π. According to the equivalence conditions given in Section 4.2 of that paper, it is enough to check:

(1) *Prior Positivity*: $\Pi(B_\delta) > 0$, $B_\delta = \{(f, \sigma) : |f - f_0|_\infty < \delta, |\frac{\sigma}{\sigma_0} - 1| < \epsilon\}$ for any $\delta > 0$

(2) *Probability of* Θ_n^c: $Pr(\Theta_n^c) \leq C_2 e^{-c_2 n}$, $\Theta_n^c = \Theta_{n,0}^C \bigcup \Theta_{n,1}^C$, $\Theta_{n,0}^C = \{f : |f|_\infty > M_n\}$,$\Theta_{n,1}^C = \{f : |f'|_\infty > M_n\}$ where $M_n = O(n^\alpha)$ for $1/2 < \alpha < 1$.

(3) *Existence of tests*: See Equation 3 and Theorem 2 in [CS07].

We first show that (1) is true. As $\Pi_2(|\frac{\sigma}{\sigma_0} - 1| < \epsilon)$ are always positive, it's enough to show $\Pi_1(|f - f_0|_\infty < \delta) > 0$. Notice that $\{|f - f_0|_\infty < \delta\} \supset \{|\sum_{k\in S}(\lambda_k \phi_k(x) - \lambda_k^0 \phi_k(x)|_\infty < \delta/2\} \bigcap \{|\sum_{k\notin S} \lambda_k \phi_k(x)|_\infty < \delta/2\} = A \bigcap B$. Here we know that conditional on w, A and B are independent as λ_k are conditionally independent. So $P(A \bigcap B|w) = P(A|w)P(B|w)$.

For the set $C_\epsilon = \{w_k > \epsilon, k \in S\}$ where ϵ is small enough, we claim that $\Pi_1(C_\epsilon) > 0$ and $P(A|w) > t(\epsilon, \lambda_0) > 0$ for some function t and any $w \in C_\epsilon$. The first claim is clear as $w_k > 0$ almost surely and S is finite; The second claim is true as $D = \{\lambda_k \sim N(0, \sigma_k^0)$ for $k \in S\}$ with probability large than $\epsilon^{|n|}$(the probability of choosing slab part for all covariates in S) on C, and $P(A|w, \{\sigma_k\}_k, k \in S) = P(A|\{\sigma_k\}_k = \{\sigma_{k,0}\}_k, k \in S) = P(|\sum_{k\in S}(\lambda_k \phi_k(x) - \lambda_k^0 \phi_k(x)|_\infty < \delta/2|\{\sigma_k\}_k = \{\sigma_{k,0}\}_k, k \in S) \geq \Pi_{k\in S} P(a_k|\lambda_k - \lambda_k^0| < \delta/(2|S|)|\{\sigma_k\}_k = \{\sigma_{k,0}\}_k, k \in S) = s > 0$, then by marginizing out σ_k we prove the second claim: $P(A|w) > s * \epsilon^{|S|} = t(\epsilon, \lambda_0) > 0$.

Now we claim that $P(B|w) > 0$ almost surely. If this is true, then $\Pi_1(A \bigcap B) = \mathbb{E}_{\Pi_1}(P(A \bigcap B|w)) = \mathbb{E}_{\Pi_1}(P(A|w)P(B|w)) \geq \mathbb{E}_{\Pi_1}(P(B|w) \cdot P(A|w) 1_{C_\epsilon}) \geq \mathbb{E}_{\Pi_1}(P(B|w) \cdot t(\epsilon, \lambda^0) 1_{C_\epsilon}) > 0$. The last step is due to $P(B|w) > 0$ almost surely and $\Pi_1(C_\epsilon) > 0$.

With this shown, notice that $P(B|w) = \mathbb{E}(P(B|w, \{\sigma_k\}_{k=1}^\infty)|w)$. Further note that we have $P(B|w, \{\sigma_k\}_{k=1}^\infty)$ almost surely, from the definition of $B = \{|\sum_{k\notin S} \lambda_k \phi_k(x)|_\infty < \delta/2\}$ and the fact $Y = \sum_{k\notin S} \lambda_k \phi_k(x)|w, \{\sigma_k\}_{k=1}^\infty$ follows a Gaussian process with $\sum_{k\notin S} a_k \sigma_{k,0}$. (For this Gaussian process, we always have $P(|Y|_\infty < \delta) > 0$, see the argument outlined in [CS07]). It thus holds that $P(B|w) > 0$ almost surely. So we prove the claim and (1) is true.

Next, we show that (3) is true. This follows by Theorem 2 of [CS07] and our assumption on the sampling of design points.

Finally, we show that (2) holds. We follow below a similar argument as in Example 6.1 of [CS07].

Since $|\sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{x})| < \sum_{k=1}^{\infty} a_k |\lambda_k|$, it follows by Markov's inequality, Chernoff bounds, and the fact that $|\lambda_k| \leq |N(0, \sigma_{k,0}^2)|$, $i > 0$ in distribution sense, that:

$$\begin{aligned}
\Pi_1 \left(\sup |\sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{x})| > M_n \right) &\leq \Pi_1 \left(\sum_{k=1}^{\infty} a_k |\lambda_k| > M_n \right) \\
&\leq \exp(-tM_n) \mathbb{E}(\exp(\sum_{k=1}^{\infty} a_k |\lambda_k|)) \\
&= \exp(-tM_n) \mathbb{E}(\mathbb{E}(\exp(\sum_{k=1}^{\infty} a_k |\lambda_k|)|w)) \\
&\leq \exp(-tM_n) \mathbb{E}(\mathbb{E}(\exp(\frac{1}{2} t^2 \sum_{k=1}^{\infty} a_k^2 \sigma_{k,0}^2) 2\Phi(a_j \sigma_{k,0} t)|w)) \\
&= \exp(-tM_n) \mathbb{E}(\exp(\frac{1}{2} t^2 \sum_{k=1}^{\infty} a_k^2 \sigma_{k,0}^2) 2\Phi(a_j \sigma_{k,0} t)).
\end{aligned}$$

Now the left side is precisely the same as Example 6.1 of [CS07]. We can thus follow the arguments there to show

$$\Pi_1 \left(\sup |\sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{x})| > M_n \right) \leq \exp\left(-\frac{n}{4 \sum_{k=1}^{\infty} a_k^2 \sigma_{k,0}^2} \right).$$

In the same way, we can show $\Pi_1(\sup |\sum_{k=1}^{\infty} \lambda_k \phi_k'(\boldsymbol{x})| > M_n) \leq \exp(-\frac{n}{4 \sum_{k=1}^{\infty} b_k^2 \sigma_{k,0}^2})$, thus showing (2).

A similar argument as the above can then be extended for the *multivariate* HierGP. It is enough to check the same three conditions (1)-(3), with normal random

variables changed to multivariate normal random variables. Analogous proofs can be used to show these conditions. □

2.7.2 Proof of Theorem 2.5.2

Proof. This theorem is precisely Theorem 6 in [CS07] for our prior, and thus it is enough to check the conditions in Theorem 1 of the same paper for the prior Π. According to the equivalence conditions given in Section 4.2 of that paper, it's enough to check:

(1) *Prior Positivity*: $\Pi(B_\delta) > 0$, $B_\delta = \{(f, \sigma) : |f - f_0|_\infty < \delta, |\frac{\sigma}{\sigma_0} - 1| < \epsilon\}$ for any $\delta > 0$

(2) *Probability of* Θ_n^c: $Pr(\Theta_n^c) \leq C_2 e^{-c_2 n}$, $\Theta_n^c = \Theta_{n,0}^C \bigcup \Theta_{n,1}^C$, $\Theta_{n,0}^C = \{f : |f|_\infty > M_n\}$,$\Theta_{n,1}^C = \{f : |f'|_\infty > M_n\}$ where $M_n = O(n^\alpha)$ for $1/2 < \alpha < 1$.

(3) *Existence of Tests*: See Equation 3 and Theorem 2 of [CS07].

The verification of such conditions is the same as the last Theorem. □

2.7.3 Proof of Theorem 2.5.3

Proof. By Theorems 2.1, 2.4, 3.1 of [SL17], we get the consistency and shape approximation for $\{\lambda_k\}_{k=1}^{H_n}$. As $\{\phi_k\}_{k=1}^{H_n}$ are orthonormal basis in L^2, by the isometry of $l_2 - L_2$ and the contraction of total variation under coefficients-function mapping, the result is proven. □

2.8 Derivation of the Gibbs sampler for the HierGP

The derivations present here are akin to the derivation in [LDD20], but we provide them below for completeness.

For the univariate HierGP, define the independent indicators z_k with probabilities $p(z_k = l|\{\nu_k, w_k\}_{k=1}^K) = \nu_l w_{l-1} (k, l = 1, \cdots, K)$, and the conditional probabilities of σ_k given by:

$$(\sigma_k|z_k) \sim (1 - 1_{z_k \leq k}) IG(a_\sigma, b_\sigma) + 1_{z_k \leq k} \delta_{\theta_\infty}. \tag{2.38}$$

Note that marginalizing out the z_k gives the original prior distribution. To get the full conditional probability $p(z_k|-)$, first we get the joint conditional distribution $p(z_k, \sigma_k|-)$ by the distribution of $p(z_k = l|\{\nu_k, w_k\}_{k=1}^K)$, (B.1) and $\lambda_k|\sigma_k$. Then marginalizing the z_k we can get the $p(z_k|-)$. Specifically, $\sigma_k \sim IG(a_\sigma, b_\sigma)$ when $z_k > k$, and marginalizing out σ_k gives us $p(z_k = l|-) \propto \nu_l w_{l-1} \cdot \int p(\lambda_k|\sigma_k) p(\sigma_k|-) d\sigma_k \propto \nu_l w_{l-1} \cdot t_{2a_\sigma}(\lambda_k; 0, (b_\sigma/a_\sigma))$ where we use the fact that normal density integrating with inverse gamma leads to density of t-distribution. Similarly, when $z_k \leq k$ we have $p(z_k = l|-) \propto \nu_l w_{l-1} \cdot N(\lambda_k; 0, \sigma_\infty)$, as in this case $\sigma_k = \theta_\infty$ which is a constant.

For the multi-variate case, we define the multi-index independent indicators $p(z_{k_1 \cdots k_d} = (l_1 \cdots l_d)|\{\nu_{l_1}^1, \cdots, \nu_{l_d}^d\}\{w_{l_1}^1, \cdots, w_{l_d}^d\}) = \Pi_{m=1}^d \nu_{l_m}^m w_{l_m - 1}^m$, and define the conditional distribution

$$(\sigma_{\boldsymbol{k}}|z_{\boldsymbol{k}}) \sim (1 - 1_{z_{\boldsymbol{k}} \leq \boldsymbol{k}}) IG(a_\sigma, b_\sigma) + 1_{z_{\boldsymbol{k}} \leq \boldsymbol{k}} \delta_{\theta_\infty}. \tag{2.39}$$

Note again that the marginalization of the $z_{\boldsymbol{k}}$ gives the original prior distribution. As above, we first get the joint distribution $p(z_{\boldsymbol{k}}, \sigma_{\boldsymbol{k}}|-)$ and then marginalizing out $z_{\boldsymbol{k}}$ by (B.2), which finally gives us

$$p(z_{\boldsymbol{k}} = \boldsymbol{l}|-) \propto \begin{cases} (\Pi_{m=1}^d \nu_{l_m}^m w_{l_m - 1}^m \cdot N(\lambda_{\boldsymbol{k}}; 0, \sigma_\infty), & \text{otherwise,} \\ (\Pi_{m=1}^d \nu_{l_m}^m w_{l_m - 1}^m \cdot t_{2a_\sigma}(\lambda_{\boldsymbol{k}}; 0, (b_\sigma/a_\sigma)), & \text{for } l_1 > k_1, \cdots, l_d > k_d. \end{cases}$$

2.9 Adaptive truncation limits for the HierGP

We provide here an adaptive implementation of the Gibbs sampler for fitting the proposed HierGP model. The adaptivity here relates to the increase in the truncation limit $\boldsymbol{K}$ as more data are observed from f. We adopt the approach in [LDD20], which iteratively (with probability decaying with time) removes the indexes for inactive basis functions and adds indices for additional bases. This probabilistic approach enables the HierGP to adaptively update the truncation limit as we collect more data on the response surface f, similar to the approach in [LDD20] for factor models. Below we outline such an adaptive Gibbs sampler first for the univariate HierGP, then for the general multivariate HierGP.

Algorithm 4: One cycle of the adaptive Gibbs sampler for the univariate HierGP

Inputs: number of iterations until adaptivity $\bar{B}$, initial truncation K, constants α_0 and α_1.

for $b = 1, \cdots, B$ **do**
  Perform one iteration of the Gibbs sampler in Algorithm 3.1.
end for
if $b \geq \bar{B}$ **then**
  With probability $p(b) = \exp(\alpha_0 + \alpha_1 b)$ ...
  if $K^* = \sum_k 1_{z_k^{[b]} > k} \leq K$ **then**
    Set $K \leftarrow K^* + 1$.
    Drop the inactive columns in Λ together with the associated parameters in $\Lambda, \sigma, \omega, \nu$.
    Add a final component to Λ, σ, w, ν sampled from the corresponding priors.
  else
    Set $K \leftarrow K + 1$.
    Add a final column sampled from the spike to Λ, together with the associated parameter in σ, w, ν, sampled from the corresponding priors.
  end if
end if

Algorithm 5: One cycle of the adaptive Gibbs sampler for the multivariate HierGP

Inputs: number of iterations until adaptivity $\bar{B}$, initial truncation $\boldsymbol{K}$, constants α_0 and α_1.

for $b = 1, \cdots, B$ **do**
    Perform one cycle of the Gibbs sampler in Algorithm 3.2.
end for
if $b \geq \bar{B}$ **then**
    With probability $p(b) = \exp(\alpha_0 + \alpha_1 b)$ ...
    for m in $1 : d$ **do**
        if $K_m^* = \sum_{\boldsymbol{k}} 1_{z_{\boldsymbol{k},m}^{[b]} > k_m} \leq K_m$ **then**
            set $K_m = K_m^* + 1$.
            Drop the inactive columns in Λ together with the associated parameters in σ, w, ν.
            Add a final component to Λ, σ, w, ν from the corresponding priors.
        else
            Set $K_m = K_m + 1$.
            Add a final column sampled from the spike to Λ, together with the associated parameter in σ, w, ν, sampled from the corresponding priors
        end if
    end for
end if

2.10 Conclusion

We proposed in this work a novel hierarchical shrinkage Gaussian process (HierGP), which embeds the well-known principles of effect sparsity, heredity and hierarchy [HW92] within carefully-constructed cumulative shrinkage priors in a Gaussian process model. Similar to the use of such principles for classical analysis of experiments (see, e.g., WH09), the embedding of this structured sparsity structure within a Bayesian nonparametric predictive model allows for improved predictive performance given limited experimental data. We then derived efficient posterior sampling algorithms for model training and prediction, and proved desirable consistency results for the HierGP. Numerical experiments confirmed the improved performance of the HierGP over existing models, for both computer code emulation and dynamical system recovery.

Despite promising results, there are many avenues for fruitful future work. One direction is in establishing posterior contraction rates for the HierGP. In the case where f has the presumed structured sparsity, it would be interesting to explore whether the HierGP rates improve upon standard contraction rates for GPs, which are known to suffer from a curse-of-dimensionality [vdVvZ08]. Another direction is in exploring cumulative shrinkage priors which can capture a weaker form of effect heredity (see, e.g., WH09; MW19), where interactions effects can be active when at least one component effect is active. This can provide a more flexible model in cases where there may be minor violations to the effect principles.

Chapter 3

Bayesian Spectrum Inference for Continuous-time Markov chain

3.1 Introduction

In this chapter, we focus on continuous-time Markov chains on a finite state space $\mathbf{S}$. A CTMC is defined by (1) its state space $\mathbf{S}$, (2) a transition rate matrix (generator) $\mathbf{L}$ with dimensions equal to that of $\mathbf{S}$, and (3) an initial distribution $X_0 \sim \pi_0$. For $p, q \in \mathbf{S}, p \neq q$, we denote the non-diagonal elements of $\mathbf{L}$ as $\mathbf{L}_{pq}$ or $\mathbf{L}(p, q)$, and they are non-negative and describe the rate of the process transitions from state p to state q. The diagonal elements $\{\mathbf{L}_{pp}\}_{p\in\mathbf{S}}$ are chosen to be negative for mathematical convenience to restrict $\sum_{q\in\mathbf{S}} \mathbf{L}_{pq} = 0$ for all p.

For a CTMC, the generator can be defined as

$$\lim_{t\to 0+} \frac{\mathbb{E}_p f(X_t) - f(p)}{t} = \mathbf{L}f(p) = \sum_{q\in\mathbf{S}} \mathbf{L}(p, q)f(q)$$

for all suitable test functions $f : \mathbf{S} \to \mathbb{R}$, where $\mathbb{E}_p$ denotes expectation conditional on $X_{t=0} = p$ and X_t is a CTMC sample path.

We can define the transition probability matrix $\mathbf{P}^t$ as

$$\mathbf{P}^t(p, q) = P(X_t = q|X_0 = p) = \lim_{N\to\infty} \frac{\sum_{j=-N}^{N} 1_{X_{jt}=p} 1_{X_{(j+1)t}=q}}{\sum_{j=-N}^{N} 1_{X_{jt}=p}},$$

where $1_{(\cdot)}$ is the indicator function. We can write the Chapman-Kolmogorov equation as

$$\frac{d\mathbf{P}^t}{dt} = \mathbf{L}\mathbf{P}^t.$$

By spectral theorems, the solution of this equation can be expressed as

$$\mathbf{P}^t = e^{t\mathbf{L}} = \sum_{n=0}^{\infty} \frac{(t\mathbf{L})^n}{n!},$$

computation of which is often intractable [CVE06].

Discretely observed model. In this paper, we consider observing the data discretely $\{X_{t_i}\}_{i=1}^n$ at equally-spaced times $t_i = i\Delta t$ with $\mathbf{S} = \{1, \ldots, N\}$. From such data, our goal is to infer the parameters characterizing the CTMC, including uncertainty quantification in estimation. Parameter estimation is more challenging than in the discrete-time case due to the intractability of the likelihood function. The issue is that we do not observe the transition times but only have observations of the state of the process at different time steps. The intractability of the likelihood, and the need for dimensionality reduction to improve the quality of parameter estimates from sparse data, motivates us to define an alternative representation of a CTMC instead of relying directly on the generator.

3.2 Spectrum decomposition and approximation for generator of CTMC

Defining a spectral decomposition of the generator $\mathbf{L}$ in terms of right and left eigenvectors, we have

$$\mathbf{L}\phi_k^T = \lambda_k \phi_k^T, \quad \psi_k \mathbf{L} = \lambda_k \psi_k, \quad \psi_k \phi_l^T = \delta_{kl} \tag{3.1}$$

with $\psi_k = (\psi_k(1), ..., \psi(N))$, $\phi_k = (\phi_k(1), ..., \phi_k(N))$, and order $Re(\lambda_k) \geq Re(\lambda_{k+1})$. Assuming uniqueness of the first eigenvector ($k = 1$), the first eigenvector encodes the invariant distribution $\mu(x)$:

$$\psi_1 = \mu, \quad \phi_1^T = (1, ..., 1)^T, \quad \lambda_1 = 0. \tag{3.2}$$

If we further assume that all eigenvalues have multiplicity one, $\mathbf{L}$ has spectral representation

$$\mathbf{L}(p,q) = \sum_{k=1}^{N} \lambda_k \phi_k(p) \psi_k(q). \tag{3.3}$$

This spectral representation of $\mathbf{L}$ can simplify computation of the exponential of $\mathbf{L}$. Indeed, the transition probability matrix $\mathbf{P}^{\Delta t}$ can be represented as

$$\mathbf{P}^{\Delta t}(p,q) = \sum_{k=1}^{N} \Lambda_k(\Delta t) \phi_k(p) \psi_k(q), \tag{3.4}$$

where $\Lambda_k(\Delta t) = \exp(\lambda_k \Delta t)$. Conversely, if we know the eigenmodes of $\mathbf{P}^{\Delta t}$ for which the above decomposition holds, then $\mathbf{L} = \frac{1}{\Delta t} \log(\mathbf{P}^{\Delta t})$ can be represented as in (3.3) with $\lambda_k = (\Delta t)^{-1} \log\{\Lambda_k(\Delta t)\}$. Letting $t_i = i\Delta t$, we can express the likelihood function as

$$\begin{aligned}
\mathbb{L}(\{X_{t_i}\}_{i=1}^n | P^{\Delta t}) &= \mathbb{L}(\{X_{t_i}\}_{i=1}^n | \{\lambda_k, \phi_k, \psi_k\}_{k=1}^N) \\
&= \prod_{i=1}^{n} \mathbf{P}^{\Delta t}(X_{t_{i-1}}, X_{t_i}) \\
&= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{N} \exp(\lambda_k \Delta t) \phi_k(X_{t_{i-1}}) \psi_k(X_{t_i}) \right\}.
\end{aligned} \tag{3.5}$$

Although low-rank approximations of dynamical systems and Markov processes are of increasing interest ([ZLWZ22; BTBK14]), there has been limited consideration of the continuous-time case. For a discrete-time Markov chain, a common approach for reducing computational and statistical complexity in estimating the transition probability matrix is to rely on a low-rank approximation ([ZW19; GGK+23]). However, such approximation methods cannot be directly applied to CTMCs, as we are not only concerned with the fixed-time transition matrix $\mathbf{P}^{\Delta t}$, but also the underlying CTMC and its generator. For CTMCs, we need to consider the relations between the generator and the transition matrix in the approximation. To

the best of our knowledge, in the previous literature, no low-rank approximation method has been established for CTMC. In this paper, we use new approximations for both the generator and the transition matrix as $\mathbf{L}(p,q) = \sum_{k=1}^{d} \lambda_k \phi_k(p)\psi_k(q)$ and $\mathbf{P}^{\Delta t}(p,q) = \sum_{k=1}^{d} \Lambda_k(\Delta t)\phi_k(p)\psi_k(q)$ with $d \ll N$.

As in [CVE06], we focus on equally-spaced observations in time, but extensions to unequally-spaced data are straightforward, as we will discuss later. In our setting, the low-rank approximation is natural for two reasons. First, in many real applications, the generators for the underlying CTMC models are approximately low-rank [GGK$^+$23]. Second, when Δt is large, the transition matrix $\mathbf{P}^{\Delta t}$ will become closer to low rank as the larger eigenvalues decay to 0 faster. This is similar to the two-time-scale approximation of Markov chains ([YZ05]): for low-frequency observations, the process can be approximated well by the small eigenvalues and eigenvectors (low-frequency terms), and large eigenvalues/eigenvectors can be ignored. We rely on likelihood (3.5) along with a low-rank approximation. If we can obtain an approach for inferences on the eigenmodes $\{\psi_k, \phi_k, \lambda_k\}_{k=1}^{d}$, then we can directly infer the stationary distribution μ through the first eigenvector ψ_1, the eigengap through the second eigenvalue λ_2, and the eigenvalue decay $\lambda_k - \lambda_{k-1}$. Motivated by a desire to characterize statistical uncertainty in estimating the spectral representation of the generator, we propose to take a Bayesian approach to inference. This is facilitated by a careful reparameterization described in the next subsection.

3.3 Spectral Approximation of the Likelihood

Potentially, we may start by choosing priors for the eigenmodes $\{\psi_k, \phi_k, \lambda_k\}_{k=1,\ldots,d}$ and then update these priors with the likelihood function (3.5) to obtain a posterior for the eigenmodes. Then, for each sample from this posterior, we can calculate the low rank approximation of the generator $\mathbf{L}$ to obtain corresponding posterior samples

of $\mathbf{L}$.

However, this direct approach is difficult to implement in practice. First, we have discussed in Section 3.1 that the likelihood is difficult to evaluate if we directly parameterize $\mathbf{L}$. Computing $\mathbf{P}^{\Delta t}$ from $\mathbf{L}$ will be costly and numerically unstable. Second, regarding parameterization in terms of $\{\psi_k, \phi_k, \lambda_k(\Lambda_k)\}_{k=1}^d$, while we get a closed form likelihood from (3.5), it is hard to use it directly in practice. Without carefully imposing constraints, the estimated transition matrix will likely not be valid [She21]. However, implementing appropriate constraints in estimating $\mathbf{L}$ is difficult due to the Markov chain embedding problem [CVE06]. Third, while directly estimating $\mathbf{P}^{\Delta t}$ is manageable, the estimated $\mathbf{P}^{\Delta t}$ does not necessarily belong to the set $\{\mathbf{P}^{\Delta t} | \exists \mathbf{L} \; such \; that \; (1) \sum_{j=1}^N \mathbf{L}_{ij} = 0, (2) \mathbf{L}_{ij} \geq 0 \; \forall i \neq j, \; and \; (3) \mathbf{P}^{\Delta t} = \exp(\Delta t \mathbf{L})\}$. This is again due to the embedding problem [CVE06].

To tackle these difficulties, we propose a different inferential strategy. We aim to estimate the transition probability $\mathbf{P}^{\Delta t}$ and the corresponding eigenmodes $\{\psi_k, \phi_k, \lambda_k(\Lambda_k)\}_{k=1}^d$ at the same time using the following pseudo log-likelihood:

$$
\begin{aligned}
&\log \mathbb{L}(\{X_{t_i}\}_{i=1}^n | \mathbf{P}^{\Delta t}, \{\psi_k, \phi_k, \lambda_k\}_{k=1}^d) \\
&= \log(\Pi_{i=1}^n \mathbf{P}^{\Delta t}(X_{t_{i-1}}, X_{t_i})) - \alpha/2 ||\mathbf{P}^{\Delta t} - \sum_{k=1}^d \exp(\lambda_k \Delta t) \phi_k \psi_k||_F^2 \\
&= \sum_{i=1}^n \log(\mathbf{P}^{\Delta t}(X_{t_{i-1}}, X_{t_i})) - \alpha/2 \sum_{p,q=1}^N \left| \mathbf{P}^{\Delta t}(p,q) - \sum_{k=1}^d \exp(\lambda_k \Delta t) \phi_k(p) \psi_k(q) \right|^2
\end{aligned}
\tag{3.6}
$$

with $\alpha > 0$ and $|| \cdot ||_F$ the Frobenius norm. Equation (3.6) has two parts: (1) the sum of the log transition probabilities between states at adjacent time observations, and (2) the Frobenius norm of the difference between transition probability $\mathbf{P}^{\Delta t}$ and its spectral approximation. Term (2) is equivalent to the log likelihood under a matrix factorization model that expresses the mean as low rank and assumes i.i.d

Gaussian measurement errors across the cells. Hence, (3.6) combines log likelihoods for transition estimation and low-rank decomposition together, leading to a useful pseudo log likelihood, which we rigorously justify in Section 3.5.

3.4 Pseudo Posterior Inference for BASI

In this section, we specify prior probability distributions for the unknowns in our model and propose a (pseudo) Bayesian approach for statistical inference. The proposed approach uses a modular strategy for inference relying on (3.6). We start by conducting inferences on the first module corresponding to the first term of (3.6) to obtain an initial posterior for the transition probability matrix that relies on a conjugate Dirichlet prior that does not exploit the low rank approximation. We then proceed to conduct inferences on the second module corresponding to the terms in the eigen decomposition conditionally on the inferred transition matrix. Under appropriate priors, Gibbs sampling can be used to generate samples from our pseudo posterior as described in Section 3.4.1, while Section 3.4.2 sketches Markov chain Monte Carlo sampling under arbitrary priors.

3.4.1 Gibbs sampling

Let $\boldsymbol{\phi} = (\phi_1^T, ..., \phi_d^T) \in \mathbf{R}^{n\times d}$ and $\boldsymbol{\psi} = (\psi_1^T, ..., \psi_d^T) \in \mathbf{R}^{n\times d}$. Letting $\mathbf{P}^{\Delta t}(p, \cdot) = (\mathbf{P}^{\Delta t}(p, q))_{q=1}^N \overset{i.i.d.}{\sim} Dir(\alpha, ..., \alpha), p = 1, ..., N$, the first module posterior is

$$\mathbf{P}^{\Delta t}|\{X_{t_i}\}_{i=1}^n = \otimes_{p=1}^N \mathbf{P}^{\Delta t}(p, \cdot)|\{X_{t_i}\}_{i=1}^n \sim \otimes_{p=1}^N Dir(\alpha + \boldsymbol{c}_p), \tag{3.7}$$

where $\boldsymbol{c}_p = (c_p^q)_{q=1}^N$, with $c_p^q = (\sum_{i=1}^{N-1} 1_{\{X_{t_i}=p, X_{t_{i+1}}=q\}})$ the empirical transition counts. Hence, the first stage posterior is independent Dirichlet on the rows of the transition matrix.

Proceeding to the second module, we put priors on the components $\{\psi_k, \phi_k, \lambda_k\}_{k=1}^d$. Prior specification and posterior computation are made more difficult by the orthogonality constraint. For recent related work, refer to [JHD20] and [JHD21]. To simplify computation, we follow [DYND20] and [MHD22], and define a prior that relaxes the orthogonality constraint by shrinking strongly towards the Stiefel manifold. Recalling $\psi_1 = \mu$, $\phi_1^T = (1, ..., 1)^T$, and $\lambda_1 = 0$ ($\Lambda_1 = 1$), the resulting logarithm of the joint prior density is (up to a constant):

$$\begin{aligned}
&\log \mathbb{P}(\{\psi_k, \phi_k, \Lambda_k\}_{k=1}^d) = \\
&(||\phi_1||^2 - ||\phi||_F^2)/2\sigma_\phi^2 - ||\psi||_F^2/2\sigma_\psi^2 - \log(P(\{\Lambda\}_{k=1}^d)) - \sum_{j,k=1}^d (\delta_{jk} - \phi_j\psi_k)^2/2\sigma_c^2 \\
=&\sum_{k=2}^d ||\phi_k||^2/2\sigma_\phi^2 - \sum_{k=1}^d ||\psi_k||^2/2\sigma_\psi^2 - \log(P(\{\Lambda\}_{k=1}^d)) - \sum_{j,k=1}^d (\delta_{jk} - \phi_j\psi_k)^2/2\sigma_c^2.
\end{aligned} \tag{3.8}$$

This expression implies multivariate Gaussian priors for $\{\psi_k\}_{k=2}^d, \{\phi_k,\}_{k=1}^d$. For $\{\Lambda_k\}_{k=1}^d$, we choose a uniform prior subject to an identifiability constraint on the eigenvalues:

$$P(\{\Lambda_k\}_{k=1}^d) \propto 1_A, \quad A = \{\{\Lambda_k\}_{k=1}^d | 1 = \Lambda_1 > \Lambda_2 \geq ... \geq \Lambda_d > 0\}.$$

The last term in (3.8) shrinks towards orthogonality to an extent determined by the value of σ_c^2; in practice we choose $\sigma_c^2 \approx 0$ to obtain near orthogonality. By our approximation $\{\psi_k, \phi_k, \Lambda_k\}_{k=1}^d | \{X_{t_i}\}_{i=1}^n, \mathbf{P}^{\Delta t} = \{\psi_k, \phi_k, \lambda_k\}_{k=1}^d | \mathbf{P}^{\Delta t}$, the full conditional distribution of ψ_k is

$$\begin{aligned}
\log \mathbb{P}(\psi_k | \{\phi_k, \Lambda_k\}_{k=1}^d, \psi_{[-k]}, \mathbf{P}^{\Delta t}) \propto -\alpha/2 \sum_{p,q=1}^N |\mathbf{P}^{\Delta t}(p,q) - \sum_{k=1}^d \Lambda_k \phi_k(p)\psi_k(q)|^2 \\
- \sum_{k=1}^d ||\psi_k||^2/2\sigma_\psi^2 - \sum_{j,k=1}^d (\delta_{jk} - \phi_j\psi_k)^2/2\sigma_c^2
\end{aligned} \tag{3.9}$$

and we have $(\psi_k|-) \sim N(\mu_{\psi_k}, \Sigma_{\psi_k})$ where

$$\Sigma_{\psi_k}^{-1} = (1/\sigma_\psi^2 + \alpha||\Lambda_k\phi_k||^2)I_n + \boldsymbol{\phi}\boldsymbol{\phi}^T/\sigma_c^2, \quad \mu_{\psi_k} = \Sigma_{\psi_k}(\alpha\Lambda_k(\boldsymbol{Y}_k^0)^T\phi_k^T + \boldsymbol{\phi}(Y_k^1)^T/\sigma_c^2)$$

$$\boldsymbol{Y}_k^0(p,q) = \mathbf{P}^{\Delta t}(p,q) - \sum_{j=1}^{d} \Lambda_j\phi_j(p)\psi_j(q) + \Lambda_k\phi_k(p)\psi_k(q), Y_k^1 = (\delta_{jk})_{j=1}^d.$$

Similarly, the full conditional distribution of ϕ_k is

$$\begin{aligned}\log \mathbb{P}(\phi_k|\{\psi_k, \Lambda_k\}_{k=1}^d, \phi_{[-k]}, \mathbf{P}^{\Delta t}) \propto -\alpha/2 \sum_{p,q=1}^{N} |\mathbf{P}^{\Delta t}(p,q) - \sum_{k=1}^{d} \Lambda_k\phi_k(p)\psi_k(q)|^2 \\ - \sum_{k=1}^{d} ||\phi_k||^2/\sigma_\phi^2 - \sum_{j,k=1}^{d} (\delta_{jk} - \phi_j \cdot \psi_k)^2/\sigma_c^2\end{aligned} \tag{3.10}$$

and we have $(\phi_k|-) \sim N(\mu_{\phi_k}, \Sigma_{\phi_k})$ where

$$\Sigma_{\phi_k}^{-1} = (1/\sigma_\phi^2 + \alpha||\Lambda_k\psi_k||^2)I_n + \boldsymbol{\psi}\boldsymbol{\psi}^T/\sigma_c^2, \quad \mu_{\phi_k} = \Sigma_{\phi_k}(\alpha\Lambda_k\boldsymbol{Y}_k^0\psi_k^T + \boldsymbol{\psi}(Y_k^1)^T/\sigma_c^2).$$

Finally, we have the full conditional distributions of Λ_k as

$$\begin{aligned}\log \mathbb{P}(\Lambda_k|\{\psi_k, \phi_k\}_{k=1}^d, \{\Lambda_k\}_{[-k]}, \mathbf{P}^{\Delta t}) \propto -\alpha/2 \sum_{p,q=1}^{N} |\mathbf{P}^{\Delta t}(p,q) - \sum_{k=1}^{d} \Lambda_k\phi_k(p)\psi_k(q)|^2 \\ - \log(P(\{\Lambda\}_{k=1}^d)\end{aligned} \tag{3.11}$$

and we have $\Lambda_k \sim N(\mu_{\Lambda_k}, \sigma_{\Lambda_k})1_{(\Lambda_{k+1}, \Lambda_{k-1})}$ where

$$\sigma_{\Lambda_k}^2 = 1/\alpha(||\phi_k||^2 * ||\psi_k||^2), \mu_{\Lambda_k} = (\alpha \sum_{p,q=1}^{N} \phi_k(p)\psi_k(q)\boldsymbol{Y}_k^0(p,q)) * \sigma_{\Lambda_k}^2$$

and we set $\Lambda_{N+1} = 0$ for convenience.

All the full conditional distributions turn out to be Gaussian or truncated Gaussian. Thus, they all can be sampled directly within a Gibbs sampling algorithm;

Algorithm 6: Gibbs sampler for generating samples from our pseudo posterior

Inputs: hyperparameters $\alpha, q_0, \sigma_\psi, \sigma_\phi, \sigma_c$, data $\{X_{t_i}\}_{i=1}^n$, number of iterations B.

1: Set initial parameters $\Theta_{[0]} = \{\mathbf{P}^{\Delta t,[0]}, (\psi_k^{[0]}, \phi_k^{[0]}, \Lambda_k^{[0]})_{k=1}^d\}$.
2: **for** $b = 1, \cdots, B$ **do**
3: Sample $\mathbf{P}^{\Delta t,[b]}|\{X_{t_i}\}_{i=1}^n$ from Dirichlet distributions $\otimes_{p=1}^N Dir(\alpha + \boldsymbol{c}_p)$. (Step 1)
4: Sample $\{\psi_k^{[b]}\}_{k=1}^d|-, \{X_{t_i}\}_{i=1}^n$ from Gaussian distribution 3.9. (Step 2)
5: Sample $\{\phi_k^{[b]}\}_{k=1}^d|-, \{X_{t_i}\}_{i=1}^n$ from Gaussian distribution 3.10. (Step 3)
6: Sample $\{\Lambda_k^{[b]}\}_{k=1}^d|-, \{X_{t_i}\}_{i=1}^n$ from the truncated Gaussian 3.11 by simulating $\mu_{\Lambda_k} + \sigma_{\Lambda_k} N(0,1)1_{(a,c)}$ where $a = \frac{\Lambda_{k+1} - \mu_{\Lambda_k}}{\sigma_{\Lambda_k}}$, $c = \frac{\Lambda_{k-1} - \mu_{\Lambda_k}}{\sigma_{\Lambda_k}}$. (Step 4)
7: **end for**

refer to Algorithm 6, where we use "$-$" to denote all other parameters and use "$\cdot|-, \{X_{t_i}\}_{i=1}^n$" to denote the full conditional distribution.

3.4.2 Estimation of low-rank representation for generator with MCMC

In this section, we consider the non-conjugate priors for the eigenvectors and eigenvalues. When we assign the non-conjugate priors, we can use the general MCMC methods like MALA to sample from the posterior.

Specifically, we can propose the prior of $P^{\Delta t}(p, \cdot)$ as

$$\mathbf{P}^{\Delta t}(p, \cdot) = (\mathbf{P}^{\Delta t}(p, q))_{q=1}^N \overset{i.i.d.}{\sim} \Pi(\mathbf{P}^{\Delta t}), \forall p = 1, ..., N$$

for some $\Pi(\cdot)$ such that $\Pi(\mathbf{P}^{\Delta t}(x, y) > q_0/N) = 1$ for some $1 > q_0 > 0$ (which means its support on the probabilistic matrices that are bounded below). The priors of $\{\psi_k, \phi_k, \Lambda_k\}_{k=1}^d$ are

$$\begin{aligned} \log \mathbb{P}(\{\psi_k, \phi_k, \Lambda_k\}_{k=1}^d) = &- \sum_{k=2}^d ||\phi_k||^2/\sigma_\phi^2 - \sum_{k=1}^d ||\psi_k||^2/\sigma_\psi^2 \\ &- \sum_{k=1}^d (\Lambda_k - \Lambda_k^0)^2/\sigma_\Lambda^2 - \sum_{k=1}^d (\delta_{ij} - \phi_k\psi_k)^2/\sigma_c^2 - \beta \sum_{p=1}^N |\sum_{q=1}^d \sum_{k=1}^N \lambda_k \phi_k(p)\psi_k(q)|^2 \end{aligned} \tag{3.12}$$

Algorithm 7: MCMC sampler for general priors

Inputs: hyperparameters $\alpha, q_0, \sigma_\psi, \sigma_\phi, \sigma_\Lambda, \sigma_c, \beta$, data $\{X_{t_i}\}_{i=1}^n$, number of iterations B.

1: Set initial parameters $\Theta_{[0]} = \{\mathbf{P}^{\Delta t,[0]}, (\psi_k^{[0]}, \phi_k^{[0]}, \Lambda_k^{[0]})_{k=1}^d\}$.
2: **for** $b = 1, \cdots, B$ **do**
3: Sample $\Theta_{[b]} = \{\mathbf{P}^{\Delta t,[b]}, (\psi_k^{[b]}, \phi_k^{[b]}, \Lambda_k^{[b]})_{k=1}^d\}$ via Metropolis-Hastings from the density proportional to equation 3.13.
4: **end for**

The log-posterior is

$$\begin{aligned}\log \mathbb{P}(\mathbf{P}^{\Delta t}, \{\psi_k, \phi_k, \lambda_k\}_{k=1}^d | \{X_{t_i}\}_{i=1}^n) \propto \log \mathbb{P}(\{\psi_k, \phi_k, \Lambda_k\}_{k=1}^d) + \log \mathbb{P}(\mathbf{P}^{\Delta t}) \\ + \log \mathbb{L}(\{X_{t_i}\}_{i=1}^n | \mathbf{P}^{\Delta t}, \{\psi_k, \phi_k, \lambda_k\}_{k=1}^d)\end{aligned} \tag{3.13}$$

Then, we can do the posterior sampling by general MCMC algorithms, for example, MALA or HMC, using the log-posterior 3.13. The sampling algorithm is given in algorithm 7.

3.4.3 Discussion

We have developed two types of algorithms for spectrum inferences with different priors. For the conjugate priors, the Gibbs updating provides efficient sampling for the posterior. For the general priors, we can do the inferences in one stage and can integrate the property of the generator into the priors of the spectrum. For the hyperparameter α, we should be careful to set it with respect to n (for example $\alpha = O(n)$), as it represents how much confidence we have in estimated $P^{\Delta t}$. For $\{\sigma_\psi, \sigma_\phi, \sigma_\Lambda, \sigma_c\}$, we can further put priors on them (for example, inverse gamma) to give our models more flexibility.

3.5 Error estimates and theoretical justification

In this section, we present the theoretical results which confirm the posterior consistency and error estimates for our methods. These results extend the existing theory on consistency for the transition matrix of the Markov Chain as well as the eigenvectors and eigenvalues. We first prove a Bernstein-von Mises (BvM) theorem for discrete-observed CTMC and then give consistent results for eigenvalues and eigenvectors based on the BvM theorem.

Let $\theta = (\mathbf{P}^{\Delta t}) \in \Theta \subset [0,1]^{d^2}$ be the parameter of a CTMC. Let θ_0 be the underlying true value which locates in the interior of Θ, and denote the local parameter $h = \sqrt{n}(\theta - \theta_0)$. When there is no confusion, we denote $X_{t_i} = X_{i\Delta t}$. Following [Con14] and [Con16], we first state the BvM theorem for CTMC. We list out the following assumptions for the theorem. We also give the detailed proof for the theorem and validity of assumptions for our models in the appendix.

(A1) *Dominated family*: For each n, the likelihood functions $\mathbb{L}(\{X_i\}_{i=1}^n|\theta)$ are dominated by a common σ-finite measure. $\hat{\mathbb{L}}(\{X_i\}_{i=1}^n, \theta)$ is the density with respect to the common dominating measure.

(A2) *Prior mass*: The prior $\Pi(\theta)$ is absolutely continuous with respect to the Lebesgue measure in a neighborhood of $\theta_0 \in \mathbb{R}^{d^2}$ with a continuous positive density at θ_0.

(A3) *Uniformly consistent test*: For every $\epsilon > 0$, there is a sequence of uniformly consitent test φ_N such that

$$\mathbb{E}_{\theta_0}(\varphi_n) \overset{n\to\infty}{\longrightarrow} 0 \quad \textit{and} \quad \sup_{||\theta-\theta_0||\geq\epsilon} \mathbb{E}_{\theta^*}(1-\varphi_n)) \overset{n\to\infty}{\longrightarrow} 0$$

(A4) *Local asymptotic normality*: There is an invertible matrix I and a sequence of random variables $\Delta_n = \frac{1}{\sqrt{n}} \nabla_{\theta_0} \log(\mathbb{L}(\{X_i\}_{i=1}^N|\theta))$ converging weakly to $\mathcal{N}(0, I)$ under θ_0 such that, for every $h_n \to h$:

$$\log(\frac{dp_{h_n}^n}{dp_0^n}) = h'\Delta_n - \frac{1}{2}h'Ih + o_{p_o^n}(1)$$

where p_h^n is the likelihood function $\mathbb{L}(\{X_i\}_{i=1}^n|\theta)$ under local parameter.

(A5) *Local linear lower bound for the score*: Let $s_n = \frac{1}{\sqrt{n}}\Delta_n$ where Δ_n is defined in assumption (A4). There is $n_0 \in \mathbb{N}, \delta < 1$ and $c > 0$ such that for any $||\theta - \theta_0|| \leq \delta$, andy $n > n_0$:

$$||E_\theta(s_n) - E_{\theta^*}(s_n) \geq c||\theta - \theta_0||$$

(A6) *Large deviation inequality for the score*: The two sided McDiarmid-type inequality for s_n holds. That means, there is $n_0 \in \mathbb{N}$ and $c < \infty$ such that:

$$P_\theta(||s_n - E_\theta(s_n)|| > u) \leq 2\exp(-\frac{1}{2}\frac{u^2}{c/n})$$

for any θ and $n > n_0$.

(A7) *Large deviation inequality for blocks*: We assume a one-sided Hoeffding-type inequality for non-overlapping consecutive blocks of X_i. To be explicit, let $R \in \mathbb{N}$ and $\hat{X}_s^R$ be the non-overlapping blocks of R consecutive X_i, ie $\hat{X}_1^R = (X_i)_{i=1}^R$, $\hat{X}_2^R = (X_i)_{i=R+1}^{2R}$, etc. Write $g_s = g(\hat{X}_s^R)$. Then there is $n_0 \in \mathbb{N}$ and $c_R < \infty$ such that for any bounded function $0 \leq g \leq 1$, any θ, any $n > n_0$:

$$P_\theta(\frac{1}{S}\sum_{s=1}^S g_s < \mathbb{E}(\frac{1}{S}\sum_{s=1}^S g_s) - u) \leq \exp(-\frac{1}{2}\frac{u^2}{c_R/S})$$

Our assumptions are identical to those in [Con14] with minor notations changes. Assumptions (A1)-(A4) are standard for i.i.d BvM theorems while (A5)-(A7) guarantee that the data $(\{X_{t_i}\}_{i=1}^n)$ is "weakly dependent" in terms of the score function. In [Con14], the BvM theorem for weakly dependent data is proved in a similar spirit to the i.i.d case. The main technical difficulty in [Con14] is to show the existence of exponential powerful tests from the uniformly consistent tests with the assumption (A5)-(A7). We refer interested readers to [Con14] for further details.

In our models, we further need the following assumptions for the underlying discrete-observed CTMC:

(B1) *Mixing property of embedded discrete-time Markov chain*: For any $P^{\Delta t}(x,y) = \theta \in \Theta$, the Markov chain $\{X_{t_i}\}_{i=1}^n$ is reccurent and mixing.

(B2) *Identification*: If $\theta, \theta' \in \Theta$ induce the same joint stationary distribution on $X_{-\infty:\infty}$, then $\theta = \theta'$.

(C1) *Full support of true embedded discrete-time Markov chain*: the finite time transition matrix $P^{\Delta t,0}(x,y)$ has full support, or equivalently, bounded below as $P^{\Delta t,0}(x,y) > q^0/N > 0$.

(C2) *Prior Modification for Conjugate priors*: The prior in sec 3.1 is modified as

$$(\mathbf{P}^{\Delta t}(p,q))_{q=1}^N \overset{i.i.d.}{\sim} Dir(\alpha,...,\alpha)1_{P^{\Delta}(x,y)>p^0/N}$$

for some small $p^0 > 0$.

(B1) and (C1) pertain to ensuring uniform mixing for the discrete-time Markov chain and guaranteeing the regularity of the parameter region which we study. They are natural assumptions for the embedded Markov chain if the underlying CTMC is ergodic and irreducible. (B2) represents the identification condition. In our proof,

we show that (B2) is satisfied naturally for our models. (C2) is a standard minor modification in prior to ensure the consistency for conjugate priors in sec 3.1.

Now we can state our BvM theorem:

Theorem 3.5.1. *(Bernstein-von Mises theorem for (discrete-observed) CTMC) Under assumption (C1) (C2), the assumptions (A1) to (A7), and (B1) to (B2) hold for the (discrete-observed) CTMC BASI models in sec 3.1 and 3.2. As a corollary of the main theorem in [Con14], we have the Berstein-von Mises theorem for the transition matrix $\theta = P^{\Delta t}$:*

$$||q^n(h) - \mathcal{N}^n(h)||_{TV} \xrightarrow{P_{\theta_0}} 0 \tag{3.14}$$

where $q^n(h)$ is the local version of the posterior $p(\theta|\{X_i\}_{i=1}^n) \propto L(\{X_i\}_{i=1}^n|\theta)p(\theta)$ and $\mathcal{N}^n(h) = N(I^{-1}\Delta_n, I^{-1})$ for I in (A4).

Roughly speaking, the above theorem shows that the posterior of $P^{\Delta t}$ converges to $P_0^{\Delta t} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \frac{2}{n}I)$ when the model is well-specified. This distribution is in the same form as that in the noisy matrix completion problem. Thus, it is reasonable to use the probabilistic matrix factorization in our approximation algorithm. Our proof is based on [Con16] and postpone to the appendix.

To further justify the consistency and efficiency of our algorithms, we state the consistency of eigenvalues and eigenvectors based on the BvM theorem in the blow theorem 3.5.2, which is in a similar spirit to [CVE06].

Theorem 3.5.2. *(Posterior approximation for the eigenvalues and eigenvector) In the same setting of Theorem 3.5.1, we have the following convergence estimates of the eigen decomposition $(\tilde{\psi}_k, \tilde{\phi}_k, \tilde{\lambda}_k)_{k=1}^N$ for the empirical generator $\tilde{L} = \frac{1}{\Delta t}\log(\tilde{P}^{\Delta t})$ with a sample from the posterior $\tilde{P}^{\Delta t} \sim P^{\Delta t}|\{X_{t_i}\}_{i=1}^n$:*

$$\sqrt{n}(\tilde{\lambda_k} - \lambda_k^0) \to \frac{1}{\sqrt{\Delta t}} \exp(-\lambda_k^0 \Delta t)\psi_k^0 Q^{\Delta t}\phi_k^0$$

$$\sqrt{n}(\tilde{\psi}_k - \psi_k^0) \to \sqrt{\Delta t}\psi_k^0 R_k^{\Delta t} Q^{\Delta t}$$

$$\sqrt{n}(\tilde{\phi}_k - \phi_k^0) \to \sqrt{\Delta t} R_k^{\Delta t} Q^{\Delta t} \phi_k^0$$

in probability, where $Q^{\Delta t} \sim \mathcal{N}(0, 2I^{-1})$ *and* $R_k^{\Delta t}(x, y) = \sum_{j \neq k} (e^{\lambda_k^0 \Delta t} - e^{\lambda_j^0 \Delta t})^{-1} * \phi_j^0(x)\psi_j^0(y)$ *with the true transition matrix* $P^{\Delta t,0}$ *and its eigen decomposition* $(\psi_k^0, \phi_k^0, \Lambda_k^0 = \exp(\lambda_k^0 \Delta t))_{k=1}^N$.

The theorem 3.5.2 shows that the direct calculations obtained from the eigen decomposition of the posterior transition matrix are reliable. Our algorithms accomplish the eigen decomposition by probabilistic matrix factorization while being more computationally efficient and offering a natural uncertainty quantification.

In our algorithms, the coverage credit interval for the transition matrix ($P^{\Delta t}$) is equal to two times of the inverse Fisher information when the model is well-specified. Thus, our approximation methods are asymptotically consistent and efficient for well-specified cases. When the model is misspecified, it is known that, for i.i.d case, the posterior variance no longer matches the minimal lower bound on the variance of unbiased estimators ([Boc22]). We can expect that the similar phenomenon happens in the Markovian cases. However, to the best of our knowledge, there is no existing results for BvM theorem for misspecified models beyond i.i.d case. [Boc22] presents several approaches to adjust the posterior variance to make the Bayesian inference efficient for misspecified models. In our algorithms, to get the correct uncertainty quantification when the model is misspecified, the simplest solution is to use maximum likelihood estimation (MLE) for the transition matrix instead of its posterior. To summarize, the credit interval for our approaches is accurate for well-specified models, and we have adjustment methods for the misspecified cases. As the result, theorems 3.5.1 and 3.5.2 justify the reasonability of our methods both in consistency and efficiency.

3.6 Numerical Simulation

In this section, we explore the proposed two-stage methods in a suite of numerical experiments. We investigate its performance for simulations from exact/non-exact CTMC simulations.

3.6.1 Simulations from exact CTMC

We first consider the data simulated from exact CTMC with generator:

$$Q = \begin{pmatrix} -0.6 & 0.4 & 0.2 & 0 \\ 0.5 & -1.2 & 0.4 & 0.3 \\ 0.3 & 0 & -0.6 & 0.3 \\ 0 & 0.1 & 0.2 & -0.3 \end{pmatrix}$$

with $\Delta t = 1$. From figure 3.1 we see the convergence and uncertainty quantification of the posteriors are close to the true eigenvectors and eigenvalues.

3.6.2 Simulations from non-exact time series

Next, we consider the data simulated from non-exact CTMC without a generator, similar to the example considered in [CVE06]. We take the transition matrix:

$$P^h(x,y) = \begin{cases} \dfrac{c_x}{|x-y|} & \text{if } x \neq y \\ \\ c_x & \text{otherwise} \end{cases} \tag{3.15}$$

where c_x is a normalizing constant such that $\forall x, \sum_y P^h(x,y) = 1$. For simulation, we take $h = 1$ and $N = 10$. From figure 3.2 we see the convergence and uncertainty quantification of the posteriors are close to the true eigenvectors and eigenvalues. As the models are non-exact, we have the true leading eigenvalues to be negative. The posterior of the leading eigenvalues concentrate around zero, and the posterior of the eigenvalues and eigenvectors give a reasonable approximation continuous time model for the non-exact model.

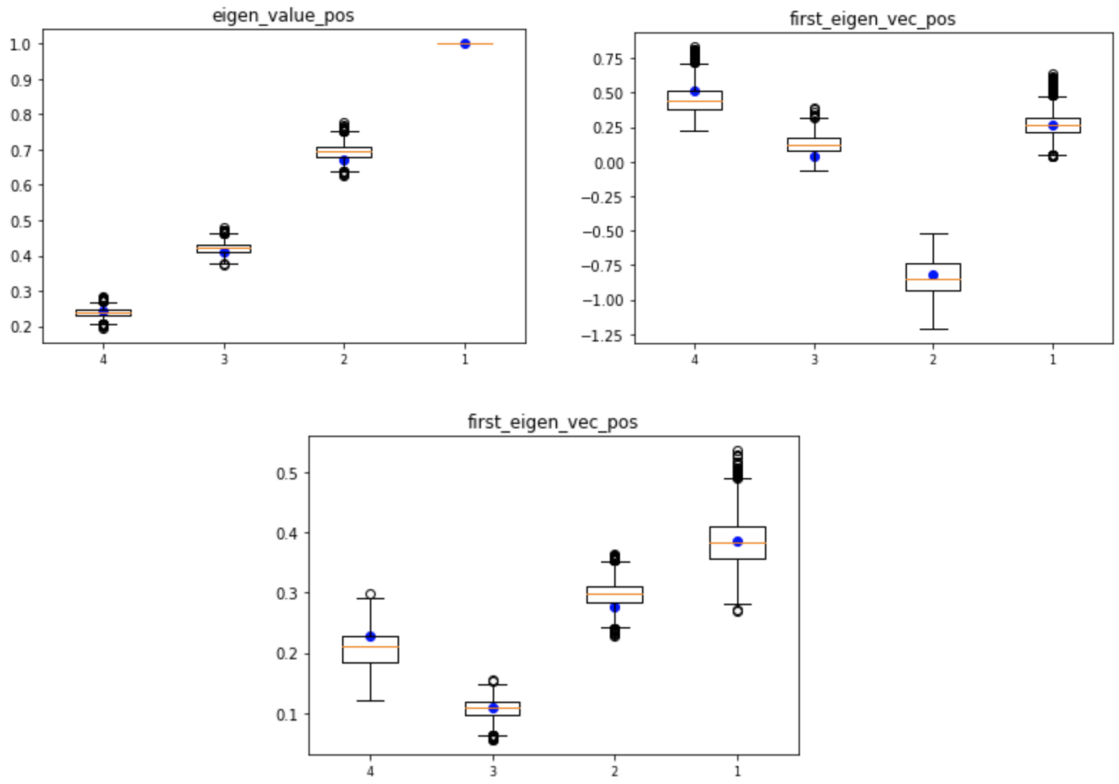

Figure 3.1: Posterior of eigenvalues, first left and second right eigenvectors for exact CTMC data.

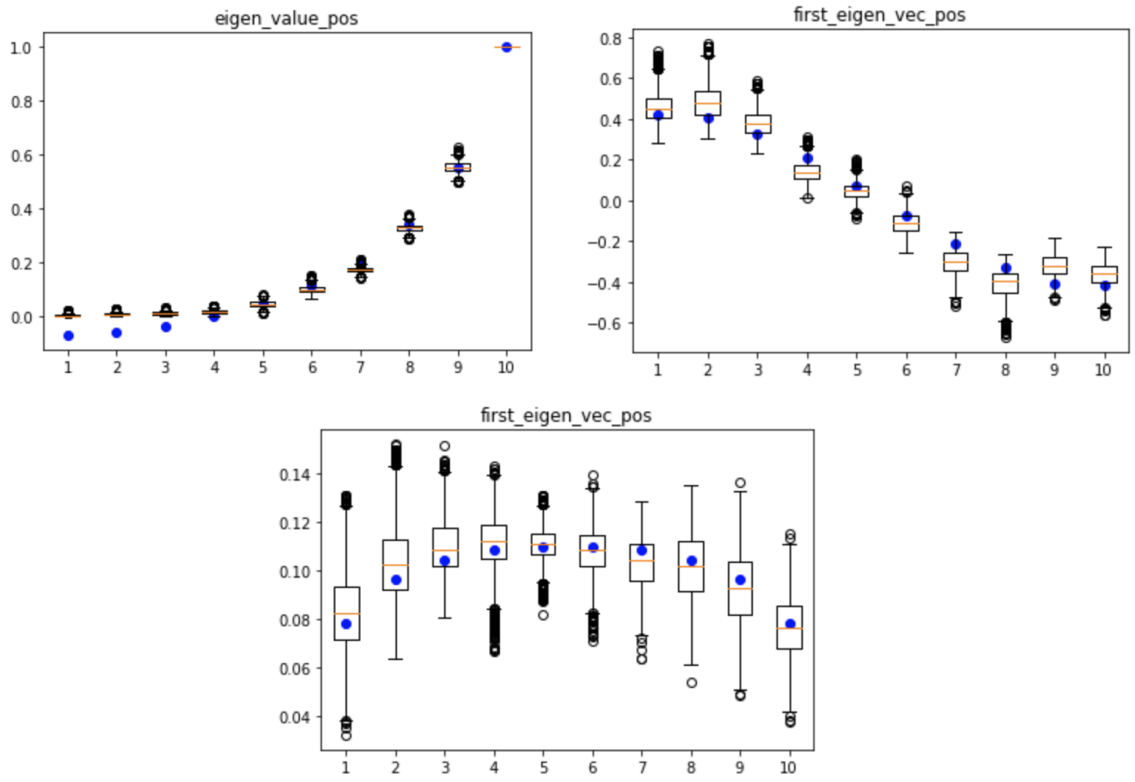

Figure 3.2: Posterior of eigenvalues first left and second right eigenvectors for non-exact CTMC data.

3.7 Technical Proofs

3.7.1 Proof of Theorem 3.1

To prove Theorem 3.5.1, we follow closely from [Con16]. We first verify each of the assumptions (B1)- (B2), and then use them to verify the assumptions (A1) - (A7). We first state the following lemmas needed.

The first lemma states the concentration inequalities for Markov chain with mixing:

Lemma 3.7.1. *Concentration inequalities for Markov Chain ([Pau15])*

Let X_i be a discrete-time Markov chain, and let $\tau(\epsilon)$ be the ϵmixing time. Let $\tau_{min} := \inf_{0\leq\epsilon<1} \tau(\epsilon)(\frac{2-\epsilon}{1-\epsilon})^2$. Let f be a function of n arguments with bounded differences:

$$f(X_1, ..., X_n) - f(X_1', ..., X_n') \leq \sum_{i=1}^{n} c_i 1_{X_i \neq X_i'}$$

. Then, the following one-sided and two-sided concentration inequality hold:

$$P(f \leq E(f) - u) \leq \exp(-\frac{1}{2}\frac{u^2}{\tau_{min}\sum_{i=1}^{n} c_i^2})$$

$$P(f \geq E(f) + u) \leq \exp(-\frac{1}{2}\frac{u^2}{\tau_{min}\sum_{i=1}^{n} c_i^2})$$

$$P(||f - E(f)|| \geq u) \leq 2\exp(-\frac{1}{2}\frac{u^2}{\tau_{min}\sum_{i=1}^{n} c_i^2})$$

Proof. See the corollary 2.11 of [Pau15] □

Next lemma derives the mixing property of (possibly non-homogeneous) Markov chain under the minorization conditions:

Lemma 3.7.2. *Theorem 4.9 of [Sen06]*

If the transition probability of a (not necessarily homogeneous) Markov Chain X_i is uniformly minorized in the sense that there are probability distribution ν_i and $\underline{a} > 0$ such that for any value of X_i, X_{i+1}

$$P(X_{i+1}|X_i) \geq \underline{a}\nu(X_{i+1})$$

Then X_i is mixing:

$$d_{TV}(\mathcal{L}(X_i|Z_0 \sim \mu_1), \mathcal{L}(X_i|Z_0 \sim \mu_2)) < (1 - \underline{a})^i$$

Proof. See the Theorem 4.9 of [Sen06]. □

Next we verify the (B1) and (B2):

Theorem 3.7.3. *Assumptions (B1) and (B2) hold for the models in sections 2.4 and 2.5*

Proof. (B1) is the direct consequence of Lemma A.2 and the transition matrixs $P^{\Delta t}$ in priors are bounded below. (B2) is clear by the way we parametrize the Markov chain, or using the convergence of transition count statistics: $\frac{\sum_{i=1}^{N} 1_{X_i=p, X_{i+1}=q}}{\sum_{i=1}^{N} 1_{X_i=p}}$. □

Now we prove X_i is uniformly mixing:

Lemma 3.7.4. *For any $P^{\Delta t} = \theta \in \Theta$, there is $\bar{\tau}_{min} < \infty$ such that:*

$$\forall \theta \in \Theta, \tau_{min}(\theta) < \bar{\tau}_{min}$$

Proof. We know that Θ is compact and that for any θ, X is recurrent and mixing. Putting (3.12) of p. 16 together with bound (2.9) p. 10 in Paulin (2014), we have the following bound:

$$\tau_{\min}(\theta) \leq 4\frac{1 + 2\log(2) + \log 1/\mu_{\min}(\theta)}{1 - \lambda(\theta)}$$

where $\mu_{\min}(\theta)$ is the minimal of stationary distribution, and $\lambda(\theta)$ is the second largest eigenvalues of the multiplicative reversibilization of $\theta = P^{\Delta t}$ ([Fil91]). Stationary distributions and eigenvalues are continuous functions of θ, and thus $\lambda(\theta)$ is bounded away from 1 and $\mu_{\min}(\theta)$ is bounded away from 0. Then we know we have a uniform low bound of $\bar{\tau}_{min}$ as in the statement.

Similarly, we can show the blocks of X_i are also uniformly mixing. Let $\hat{X}_i^R = (X_j)_{j=(i-1)R}^{iR}$ non-overlapping consecutive blocks of R. X_i is uniformly mixing and thus $\hat{X}_i$ us uniformly mixing. And we denote $\bar{\tau}_{\min,\hat{X}}$ as the low bound for mixing time. □

Now we prove the uniform law of large number for log-likelihood, observed information, and central limit theorem for score function similar to [Con16]. Without loss of generality, we assume the Markov Chain X_i is stationary in the following two lemmas: $X_0 \sim \mu$ where μ is the stationary measure. The non-stationary results can be extended from the stationary ones via the coupling argument similar to A.12.5.2 of [Con16] thanks to boundness.

Lemma 3.7.5. *ULLN for log-likelihood and observed information*

For discrete-observed CTMC models under the assumption (C1), we have

$$||\frac{1}{n}\log(\mathbb{L}(\{X_i\}_{i=1}^n|\theta) - L(\theta))||_\infty \xrightarrow{P_{\theta_0}} 0$$

$$||\frac{1}{n}\nabla^2\log(\mathbb{L}(\{X_i\}_{i=1}^n|\theta) - i(\theta))||_\infty \xrightarrow{P_{\theta_0}} 0$$

where $L(\theta) = E_{\theta_0}(\log(\mathbb{L}(X_1|X_0,\theta)))$ *and* $i(\theta) = E_{\theta_0}(\nabla^2\log(\mathbb{L}(X_1|X_0,\theta)))$

Proof. We first prove the uniform law of large number for the log-likelihood function. By Markov property, we can further write the log-likelihood function as:

$$\frac{1}{n}\log(\mathbb{L}(\{X_i\}_{i=1}^n|\theta) = \frac{1}{n}\sum_{i=1}^n \log(\mathbb{L}(X_i|X_{i-1},\theta))$$

Now we can prove the ULLN via the almost sure ergodic theorem in function space. Note the log-likelihood function $\log(\mathbb{L}(X_i|X_{i-1},\theta))$ is continuous and bounded in Θ. Now define $\mathcal{C}(\Theta)$ be the set of continuous functions from Θ to $\mathbb{R}$, the shift operator $s : \{X_i\}_{i\in\mathbb{Z}} \to \{X_{i+1}\}_{i\in\mathbb{Z}}$, and mapping $l : \{X_i\}_{i\in\mathbb{Z}} \to \mathbb{L}(\theta) = \log(\mathbb{L}(X_i|X_{i-1},\theta))$. Then we have

$$\frac{1}{n}\sum_{i=1}^{n}\log(\mathbb{L}(X_i|X_{i-1},\theta)) = \frac{1}{n}\sum_{i=1}^{n}s^i l(\{X_i\}_{i\in\mathbb{Z}})$$

where $\{X_i\}_{i\in\mathbb{Z}}$ is stationary ergodic. Since $(\mathcal{C}(\Theta), ||\cdot||)$ is separable by compactness, we have a vector version almost sure ergodic theorem (see A 2.14 of [Con16] and theorem 2.1, p.167 in[Kre85]). And we have

$$||\frac{1}{n}\sum_{i=1}^{n}\log(\mathbb{L}(X_i|X_{i-1},\theta)) - L(\theta)|| \xrightarrow{P_{\theta_0}} 0$$

where $L(\theta) = E_{\theta_0}(\log(\mathbb{L}(X_1|X_0,\theta)))$.

Similarly, for each element of $\frac{1}{n}\nabla^2\log(\mathbb{L}(\{X_i\}_{i=1}^n|\theta)$ we get the ULLN of the observed information. □

Lemma 3.7.6. *CLT for score functions*

For discrete-observed CTMC models under the assumption (C1), we have

$$\Delta_n = \frac{1}{\sqrt{n}}\nabla_{\theta_0}\log(\mathbb{L}(\{X_i\}_{i=1}^n|\theta)) \xrightarrow{P_{\theta_0}} \mathcal{N}(0, I)$$

where $I = i(\theta_0)$

Proof. Let $V_i = \nabla_{\theta_0}\log(\mathbb{L}(X_i|X_{i-1},\theta)$, then it is ergodic stationary. We show that it is also a martingale difference with respect to the filtration $\mathcal{F}_i = \sigma(\{X_j\}_{j=1}^i)$. By

dominated convergence theorem, we have

$$\begin{aligned} E_{\theta_0}(V_i|\mathcal{F}_{i-1}) &= E_{\theta_0}(\nabla_{\theta_0} \log(\mathbb{L}(X_i|X_{i-1},\theta)|\mathcal{F}_{i-1}) \\ &= \nabla_{\theta_0}(E_{\theta_0}(\log(\mathbb{L}(X_i|X_{i-1},\theta)|\mathcal{F}_{i-1})) \\ &= 0 \end{aligned}$$

where the last equality comes from that kl divergence is minimized at the probability itself. And by the central limit theorem of ergodic stationary martingale difference, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \xrightarrow{P_{\theta_0}} \mathcal{N}(0, I)$$

where $I = E_{\theta_0}(V_1 V_1') = i(\theta_0)$ □

Then we prove the local asymptotic normality for the Markov chain (A4):

Theorem 3.7.7. *(A4) is satisfied: There is an invertible matrix I and a sequence of random variables Δ_n converging weakly to $\mathcal{N}(0, I)$ under θ^* such that, for every $h_n \to h$:*

$$\log(\frac{dp^n_{h_n}}{dp^n_0}) = h'\Delta_n - \frac{1}{2}h'Ih + o_{p^n_o}(1)$$

Proof. As we have lemma A.5 and lemma A.6, the (uniform) LAN condition is an easy corollary of the limit theorems of log-likelihood/score/information by [VdV00] □

One can also use the theorem 1.24 of [HJL90] for checking the LAN conditions while we follow [Con16]'s approach for consistency.

Next, we show there are uniformly consistent tests.

Lemma 3.7.8. *(A3) is satisfied: For every $\epsilon > 0$, there is a sequence of uniformly consitent test φ_N such that*

$$\mathbb{E}_{\theta_0}(\varphi_n) \xrightarrow{n\to\infty} 0 \quad \textit{and} \quad \sup_{||\theta-\theta_0||\geq\epsilon} \mathbb{E}_{\theta^*}(1-\varphi_n)) \xrightarrow{n\to\infty} 0$$

Proof. we prove the existence of tests via the construction of uniformly consistent estimators. Define the consecutive non-overlapping block $\hat{X}_i^2 = (X_{2i}, X_{2i+1})$, and its stationary distribution π. Notice π can identify the $P^{\Delta t}$ as $P^{\Delta t}(p,q) = \frac{\pi(p,q)}{\sum_q \pi(p,q)}$. We then define the empirical estimator of π

$$\hat{\pi}_n(\{\hat{X}_i^2\}_{i=1}^n; \hat{x}) = \sum \frac{1}{n} 1_{\hat{X}_i^2 = \hat{x}} \overset{n\to\infty}{\longrightarrow} \pi(\hat{x})$$

almost surely. We will show it is also a uniformly consistent estimator, that is

$$\forall \epsilon, \quad \sup P_\theta(d_{TV}(\hat{\pi}_n, \pi) \geq \epsilon) = 0$$

.

First, as $d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi) - d_{TV}(\hat{\pi}_n(\{\tilde{X}_i\}_{i=1}^n), \pi) \leq \frac{1}{2}\frac{1}{n-2}\sum_{i=1}^n 1_{X_i \neq \tilde{X}_i}$, Lemma A.1 gives us

$$P_{\theta,\mu}(d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi) > E_{\theta,\mu}(d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi)) + u) \leq \exp(-\frac{1}{2}\frac{u^2}{\tau_{\min}\frac{n}{4(n-2)^2}})$$

where μ is the initial distribution of X_i. Then we first bound the $E_{\theta,\mu}(d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi))$ for the stationary Markov chain ($\mu = \mu_0$) and then extend it to the non-stationary case. As a corollary of (3.31) of [Pau15], we have

$$E_{\theta,\mu_0}(d_{TV}(\hat{\pi}_n, \pi)) \leq \sqrt{\frac{1}{n\gamma/2}} \sum_{\hat{x}} \sqrt{\pi(\hat{x})} \leq \sqrt{\frac{N^2}{(n-2)\gamma}}$$

where γ is the pseudo spectral gap and N^2 is the dimension of $\hat{X}_i$'s state space. We can futher bound

$$|E_{\theta,\mu_0}(d_{TV}(\hat{\pi}_n, \pi)) - E_{\theta,\mu}(d_{TV}(\hat{\pi}_n, \pi))| \leq \frac{2c}{(n-2)}(1-\rho)^n$$

where c, ρ are parameter of mixing for X_i. Gather the bounds together and we get

$$\begin{aligned}
&P_{\theta,\mu}(d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi) \geq \epsilon) \\
&= P_{\theta,\mu}(d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi) \geq E_{\theta,\mu}(d_{TV}(\hat{\pi}_n, \pi)) + (\epsilon - E_{\theta,\mu}(d_{TV}(\hat{\pi}_n, \pi)))) \\
&\leq \exp(-\frac{1}{2}\frac{(\epsilon - E_{\theta,\mu}(d_{TV}(\hat{\pi}_n, \pi)))^2}{\tau_{\min}\frac{n}{4(n-2)^2}}).
\end{aligned}$$

This bound is independent of θ, and thus we have a Dvorestzky-Kiefer-Wolfowitz type inequality

$$\forall \theta, \quad P_\theta(d_{TV}(\hat{\pi}_n(\{X_i\}_{i=1}^n), \pi) \geq \epsilon) \leq \alpha(n)$$

. Therefore, $\hat{\pi}$ is a uniformly consistent estimator. We can construct the tests as $\phi_n = 1_{d_{TV}(\hat{\pi},\pi)\geq\epsilon/2}$ and it is easy to verify they are uniformly consistent (see A.12.7.2 of [Con16]). Then we finish the proof. □

Then we state the proofs of (A5)-(A7):

Lemma 3.7.9. *(A5) is satisfied: There is $n_0 \in \mathbb{N}, \delta < 1$ and $c > 0$ such that for any $||\theta - \theta_0|| \leq \delta$, andy $n > n_0$:*

$$||E_\theta(s_n) - E_{\theta^*}(s_n) \geq c||\theta - \theta_0||$$

Proof. We have the Taylor expansion

$$E_\theta(s_n) = E_{\theta_0}(s_n) + \nabla_{\theta_0} E_{\theta_0}(s_n)(\theta - \theta_0) + (\theta - \theta_0)^T h(\bar{\theta})(\theta - \theta_0)$$

where $\bar{\theta} = t\theta_0 + (1-t)\theta$ for some $0 \leq t \leq 1$ and $h(\theta) = \nabla_\theta^2 E_\theta(s_n)$. As $\nabla_{\theta_0} E_{\theta_0}(s_n) \to I$ as $n \to \infty$ (see Lemma A.6), I is invertible and h is bounded over Θ by smoothness and compactness, there is $n_0, \delta < 1$ and c such that for any $||\theta - \theta_0|| < \delta$, $n > n_0$ the statement is true:

$$||E_\theta(s_n) - E_{\theta^*}(s_n) \geq c||\theta - \theta_0||$$

□

Lemma 3.7.10. *(A6) is satisfied: There is $n_0 \in \mathbb{N}$ and $c < \infty$ such that:*

$$P_\theta(||s_n - E_\theta(s_n)|| > u) \leq 2\exp(-\frac{1}{2}\frac{u^2}{c/n})$$

for any θ and $n > n_0$.

Proof. As the score is bounded by assumption, the above statement is an easy corollary of Lemma A.1. Let s_n^k be the $k-$th component of score, and c_k be the difference bound of $(\nabla_\theta \log(\mathbb{L}(X_i|X_{i-1}, \theta)))^k$. We have

$$P_\theta(|s_n^k - E(s_n^k)| > u) \leq 2\exp(-\frac{1}{2}\frac{u^2}{\tau_{\min}(\theta)\sum_{i=1}^n (c_k/n)^2}) \leq 2\exp(-\frac{1}{2}\frac{u^2}{\bar{\tau}_{\min} c_k/n})$$

And let $\bar{c} = max_{1\leq k\leq N} c_k$, we then have the vector concentration

$$P_\theta(||s_n - E_\theta(s_n)|| > u) \leq 2\exp(-\frac{1}{2}\frac{u^2}{\bar{\tau}_{\min}\bar{c}/n})$$

. which means the statements is true for $c = \bar{\tau}_{\min}\bar{c}$ □

Lemma 3.7.11. *(A7) is satisfied: There is $n_0 \in \mathbb{N}$ and $c_R < \infty$ such that for any bounded function $0 \leq g \leq 1$, any θ, any $n > n_0$:*

$$P_\theta(\frac{1}{S}\sum_{s=1}^S g_s < \mathbb{E}(\frac{1}{S}\sum_{s=1}^S g_s) - u) \leq \exp(-\frac{1}{2}\frac{u^2}{c_R/S})$$

Proof. As g is bounded ($0 \leq g \leq 1$) and $\hat{X}_i^R$ is uniformly mixing with mixing time bound by $\bar{\tau}_{\min,\hat{X}}$, the statement is direct corollary of Lemma A.1:

$$P_\theta(\frac{1}{S}\sum_{s=1}^S g_s < \mathbb{E}(\frac{1}{S}\sum_{s=1}^S g_s) - u) \leq \exp(-\frac{1}{2}\frac{u^2}{\bar{\tau}_{\min,\hat{X}}/S})$$

So the statement is true with $c_R = \bar{\tau}_{\min,\hat{X}}$ □

Finally, we can have the proof of Theorem 3.5.1.

Proof. As we have checked assumption (B1)-(B2), it is enough to check the assumption (A1) to (A7) by [Con14],. (A1) is true for the finite state Markov chain and (A2) is true as the way we set priors in sec 3.4.2 and 3.4.1 with the assumption (C1). (A3)-(A7) are proved in Lemma/Theorem A.8, A.7, A.9, A.10, A.11 separately. And by the main theorem of [Con14], we have the BvM approximation:

$$||q^n(h) - \mathcal{N}^n(h)||_{TV} \xrightarrow{P_{\theta_0}} 0$$

in theorem 3.1. That concludes the proof.

□

3.7.2 Proof of Theorem 3.5.2

Proof. Proof of the theorem 3.5.2 is contained in the above lemmas by similar computations of equation (24) in [CVE06]. □

3.8 Conclusion and outlook

In this essay, we develop an innovative method, the Bayesian Approximation Spectral Inference (BASI), designed to infer continuous-time Markov chains (CTMCs) from discretely observed data. This development was stimulated by the mounting interest in CTMCs across various disciplines and the computational hurdles encountered when dealing with large state spaces. We observed that CTMCs frequently exhibit a nearly low-rank structure in such situations. BASI capitalizes on this low-rank structure to approximate the generator of the CTMC, focusing on its spectral characteristics rather than a direct estimation. By approximating the likelihood within the low-rank spectral representation, BASI significantly diminishes the computational

complexity associated with high-dimensional data. Simultaneously, it provides a natural framework for uncertainty quantification for the reconstructed CTMC and its corresponding spectral characteristics.

While our results thus far are encouraging, there are abundant opportunities for future research. One potential avenue involves developing similar spectral methods for CTMCs with non-uniform time observation cases. The extension to a finite number of choices for Δt seems straightforward, although the general time duration case may necessitate additional work. Another promising research direction is to explore more structures and constraints of the CTMC generator, such as sparsity and symmetry in the generator matrix. Inferring the generator by taking into account spectral information and considering the embedding problem is also a promising direction.

Chapter 4

Adaptive Gaussian process for Data with Low Intrinsic Dimension

4.1 Introduction

In this section, we set up the notations and review background information on the Gaussian process (GP), Reproducing kernel Hilbert space (RKHS), and RKHS on the general measurable subset $\mathcal{X}$ in $\mathbb{R}^D$. Throughout the work, we focus on the case of compact $\mathcal{X}$, and assume that $\mathcal{X}$ is a bounded set inside $[0,1]^D$.

4.1.1 Basic review for Gaussian process

Gaussian process (GP) are widely used as non-parametric prior for unknown functions. For instance, in the context of nonparametric regression, one can utilize a Gaussian Process (GP) as a prior for the unknown function. Let $\mathbf{f}^* \in \mathbb{R}^n$ be the discretization of a continuous function f^* over $X_1, X_2, \ldots, X_n$ such that $\mathbf{f}^* = f^*(X_i)$. A Gaussian process (GP) prior for f implies that the conditional distribution of $\mathbf{f}^*$ given $X_1, X_2, \ldots, X_N$ is denoted as $p(\mathbf{f}^*|X_1, X_2, \ldots, X_m)$, which follows a Gaussian distribution $\mathcal{N}(0, \Sigma_{\mathbf{f}^*\mathbf{f}^*})$. Here, $\Sigma_{\mathbf{f}^*\mathbf{f}^*} \in \mathbb{R}^{n\times n}$ represents the covariance matrix induced from the kernel k of the GP. The (i,j) element of $\Sigma_{\mathbf{f}^*\mathbf{f}^*}$ corresponds to the covariance between $f(X_i)$ and $f(X_j)$, denoted as $\mathrm{cov}(f(X_i), f(X_j)) = k(X_i, X_j)$, where $1 \le i, j \le n$. By combining the prior distribution $\mathcal{N}(0, \Sigma_{\mathbf{f}^*\mathbf{f}^*})$ with the likelihood function under equation 1.1, we can obtain the posterior distribution. This posterior distribution serves as the foundation for conducting inference and making predictions.

4.1.2 Basic review for Reproducing kernel Hilbert space

Reproducing kernel Hilbert space (RKHS) is a powerful tool to study the theoretical properties of GP. We will rely on some lemmas of RKHS to prove posterior behaviour for our prior. We provide a brief overview of key definitions and concepts of RKHS, which can be found in section 4 of [SC08]. A symmetric function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is considered a kernel on a non-empty set $\mathbb{X}$ if it is positive definite, satisfying the condition that for all $\forall n \in \mathbb{N}$ and choices $x_1, ..., x_n \in \mathbb{X}$ and $\alpha_1, ..., \alpha_n \in \mathbb{R}$, the inequality $\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0$ holds.

A reproducing kernel Hilbert space (RKHS) is a Hilbert space H of functions $f : \mathbb{X} \to \mathbb{R}$ in which the evaluation functional $H \to \mathbb{R}$ is continuous for every $x \in \mathbb{X}$. Each RKHS H has a unique reproducing kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, satisfying the property

$$\langle f, k(\cdot, x) \rangle_H \quad \textit{for all, } f \in H, x \in \mathbb{X} \tag{4.1}$$

as stated in equation (4.1).

Conversely, for any given kernel k, there exists a unique reproducing kernel Hilbert space (RKHS) H in which k serves as the reproducing kernel. An alternative way to define a kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is by considering the existence of a Hilbert space H_0 and a mapping $\Phi : \mathbb{X} \to H_0$, such that $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{H_0}$ for all $x, y \in \mathbb{X}$. Here, H_0 is referred to as the feature space, and Φ is known as the feature map. It is worth noting that we can always select the RKHS H as the feature space using the canonical feature map $\Phi : \mathbb{X} \to H$, defined as $\Phi(x) := k(x, \cdot)$, where the inner product on H is entirely determined by the chosen kernel.

4.1.3 Gaussian process and RKHS on the general set $\mathcal{X}$

In this work, we focus on Gaussian process and RKHS on the general set $\mathcal{X}$. Recall that $\mathcal{X}$ is a bounded set inside $[0, 1]^D$. Let us define the squared exponential kernel

on $\mathcal{X} \times \mathcal{X}$:

$$h_\epsilon(x, y) = \exp(-\frac{||x-y||^2}{2\epsilon}) = h(\frac{||x-y||^2}{2\epsilon}) \tag{4.2}$$

where $||\cdot||$ is the Euclidean distance on the ambient space $[0,1]^D$. In fact, as we use the Euclidean distance, this kernel can be extended to be defined on the $[0,1]^D$. Using Bochner's theorem, we can express the kernel function $h_1(x, y)$ in terms of its spectral representation on the D-dimensional Euclidean space. Specifically, $h_1(x, y)$ can be represented as $\int_{\mathbb{R}^D} e^{-i(\lambda, x-y)}\mu(d\lambda)$, where μ is the spectral measure of $h_1(x, y)$ and has a sub-Gaussian tail. A sub-Gaussian tail satisfies a sub-exponential decay requirement, which means

$$\int e^{\delta||\lambda||}\mu(d\lambda) < \infty. \tag{4.3}$$

for some (in fact, any) $\delta > 0$.

The GP with squared exponential covariance/kernel $h_\epsilon(x, y)$ is frequently used. The law of the centered GP, $\{W^\epsilon(x) : x \in \mathcal{X}\}$ with squared exponential covariance $h_\epsilon(x, y)$, is entirely determined by its kernel. Although our results mainly focus on suqare exponential kernel and its associated GP, we can extend $h(x) : [0, \infty) \to [0, \infty)$ to a general function satisfies some regularity conditions. We leave the discussion in the appendix and we take $h(x) = \exp(-x)$ from now on.

We assign $\Pi(f)$, a squared exponential Gaussian process with bandwidth prior on $\mathcal{X}$, as the prior for f:

(P) Gaussian process on $\mathcal{X}$:

$$W^t|t \sim GP(0, h_t), t \sim p(t) \tag{4.4}$$

We will give conditions under which priors can give good posterior behaviour in the next section. Finally we define the estimator $\hat{f}(x)$ for f^* as the posterior mean:

$$\hat{f}(x) = \int f d\Pi(f|\{X_i, Y_i\}_{i=1}^n). \tag{4.5}$$

Following the definition of RKHS on general set from section 4 of [SC08], RKHS assoaciated with h_ϵ on $\mathbb{X} = [0,1]^D$ and $\mathbb{X} = \mathcal{X}$ is defined as the completion of the linear space $\tilde{\mathcal{H}} = \{f : \mathbb{X} \to \mathbb{R} | f = \sum_{i=1}^m a_i h_\epsilon(x_i, x),\ \textit{where}$
$a_1, ..., a_m \in \mathbb{R}, x_1, ..., x_m \in \mathbb{X}, m \in \mathbb{N}\}$ under the norm $\langle h_\epsilon(x,\cdot), h_\epsilon(y,\cdot)\rangle = h_\epsilon(x,y)$. We denote the associated reproducing kernel Hilbert space of h_ϵ on $[0,1]^D$ as $\mathbb{H}_\epsilon$ and denote the associated RKHS on $\mathcal{X}$ as $\tilde{\mathbb{H}}_\epsilon$. The properties of $\mathbb{H}_\epsilon$ and $\tilde{\mathbb{H}}_\epsilon$ are reviewed in the appendix.

4.2 Posterior contraction rates on low-dimensional set

In this section, we give conditions on the posterior contraction rate of GP on general set $\mathcal{X}$. Following [GGVDV00], [vdVvZ08], and [YD16], the posterior contraction rate of the GP prior is said to be at least ε_n under a semimetric d_n if

$$\Pi(d_n(f, f^*) > \varepsilon_n | \{X_i, Y_i\}_{i=1}^n) \to 0 \quad \textit{in probability as } n \to \infty. \tag{4.6}$$

where $\{X_i, Y_i\}_{i=1}^n$ denotes the dataset and $\Pi(A|\{X_i, Y_i\}_{i=1}^n)$ is the posterior of an event A. The semimetric d_n is used to measure the discrepancy between f and the truth f^*.

4.2.1 General assumptions

First, we state the assumptions on data and underlying truth for the regression model equation (1.1):

(A1) Low intrinsic dimension via covering number: We assume $\mathcal{X} \in [0,1]^D$. There exists a constant $C_\mathcal{X}, \varrho > 0$ and $\varepsilon_0 \in (0,1)$ such that for all $0 < \varepsilon \leq \varepsilon_0$ we have

$$N(\varepsilon, \mathcal{X}, ||\cdot||_\infty) \leq C_\mathcal{X} \varepsilon^{-\varrho}. \tag{4.7}$$

(A2) There exist $s, C_1, C_2, \epsilon_0 > 0$, such that $\forall \epsilon < \epsilon_0$, we can find a function $F^\epsilon \in \tilde{\mathbb{H}}_\epsilon$ to let the below equations hold

$$\sup_{x \in \mathcal{X}} |F^\epsilon(x) - f^*(x)| \leq C_1 \epsilon^{s/2} \tag{4.8}$$

$$||F^\epsilon(x)||^2_{\tilde{\mathbb{H}}_\epsilon} \leq C_2 \epsilon^{-\varrho/2} \tag{4.9}$$

Condition (A1) describes the complexity of data input through the covering number while the condition (A2) require that f^* can be well approximated by the RKHS on $\mathcal{X}$. There can be multiple ϱ satisfy condition (A1). The infimum over all ϱ that satisfy the condition (A1) alongside a finite constant $C_{\mathcal{X}}$ corresponds to the upper box-counting dimension of the $\mathcal{X}$, which is defined as

$$\limsup_{\varepsilon \to 0} \frac{\log(N(\varepsilon, \mathcal{X}, ||\cdot||_\infty))}{\log(\frac{1}{\varepsilon})} \tag{4.10}$$

The lower box-counting dimension is defined as the common limit when the lim sup is substituted with lim inf in equation . If these values coincide, the resulting limit represents the box-counting dimension of $\mathcal{X}$. As mentioned in [HS21], one advantage of assumption (A1) is that the upper box-counting dimension can be defined for any bounded subset $\mathcal{X} \in \mathbb{R}^D$ and is at most D. Essentially, this means we impose no requirements for smoothness or regularity on $\mathcal{X}$, which differs from the manifold assumption found in the literature.

As for the condition (A1), there are many natural examples shown as below. (A2) is usually satisfied by functions with nice properties, and we will discuss more in the next section.

Example 4.2.1. *Let $\mathcal{X} = [0,1]^D$. By proposition 1.3.1 of [SC08], there exists constants $c, C > 0$ such that*

$$c\varepsilon^{-D} \leq N(\varepsilon, \mathcal{X}, ||\cdot||_\infty) \leq C\varepsilon^{-D} \tag{4.11}$$

for all $\epsilon \in (0,1)$. As a result, in fact, any bounded $\mathcal{X} \in \mathbb{R}^D$ with non-empty interior will satisfies the assumption (A1) with $\varrho = D$.

Example 4.2.2. *Let $\mathcal{X} = \mathcal{M} \subset [0,1]^D$ be a $d-$dimensional compact connected smooth Riemannian manifold (without boundary) isometrically embedded in $[0,1]^D$ with $d \in \mathbb{Z}^+$. One can prove that $\mathcal{M}$ satisfies assumption (A1) with $\varrho = d$. We leave the proof in our appendix A.*

Example 4.2.3. *A stratified space is a mathematical object that generalizes the concept of a manifold by allowing for more complicated geometric structures. It is a space that is divided into different "strata" or "layers," each having its own well-behaved geometric properties. The covering dimension of a stratified space measures its topological complexity. It is the maximum dimension among the strata. A stratified space has finite covering dimension if and only if all of its strata have finite dimensions. A very simple example is $\mathcal{X} = \mathcal{M}_1 \bigcup \mathcal{M}_2$ where $\mathcal{M}_1, \mathcal{M}_2 \subset [0,1]^D$, a d_1/d_2-dimensional compact connected smooth Riemannian manifold with $d_1, d_2 \in \mathbb{Z}^+$. One can check that $\mathcal{X}$ satisfies assumption (A1) with $\varrho = \max(d_1, d_2)$.*

4.2.2 Convergence rate with well-specified prior

In this subsection, we show the convergence rate with well-specified prior. We further state the following condition on our prior (P):

(A3) Well-specified prior support on bandwidth: There exists $c_1, c_2, c_3, a_1, a_2, K_1, K_2, C_1, C_2 > 0$, such that

$$\forall t \in [c_1 n^{\frac{-2}{2s+\varrho}} \log^{\frac{2(1+D)}{2s+\varrho}}(n), c_2 n^{\frac{-2}{2s+\varrho}} \log^{\frac{2(1+D)}{2s+\varrho}}(n)], p(t) \geq C_1 t^{-a_1} \exp(-\frac{K_1}{t^{\varrho/2}}), \tag{4.12}$$

$$\forall t \in (0, c_3 n^{\frac{-2}{2s+\varrho}} (\log(n))^{\frac{-4(1+D)}{(2+\varrho/s)\varrho}}], p(t) \leq C_2 t^{-a_2} \exp(-\frac{K_2}{t^{\varrho/2}}) \tag{4.13}$$

Roughly speaking, the condition (A3) says that, there is enough probability mass of $p(t)$ on the interval around the order $n^{\frac{-2}{2s+\varrho}}$, while the probability $p(t)$ when t is smaller than the order $n^{\frac{-2}{2s+\varrho}}$ can be controlled. The below example, the frequently used inverse gamma prior with correct rescaling [vdVvZ08], satisfies (A3) for any $s \in \mathbb{R}^+$.

Example 4.2.4. *The rescaled inverse gamma prior is frequently used in adatpive GP. It is defined as*

$$\frac{1}{t^{\varrho/2}} \sim Ga(a_0, b_0) \tag{4.14}$$

where ϱ is the intrinsic dimension of $\mathcal{X}$, $Ga(a_0, b_0)$ is the gamma distribution with p.d.f $p(t) \propto t^{a_0-1}e^{-b_0 t}$ where $a_0, b_0 > 0$.

One drawback of the rescaled inverse gamma prior is that one has to estimate ϱ before processing the regression, and the estimation error on dimension ϱ might lead to a sub-optimal rate [YD16]. In the next section, we introduce an empirical Bayes prior which can achieve optimal contraction rate without knowing the intrinsic dimension.

Now we can give the posterior contraction rate results for prior (P) under the metrics $||\cdot||_n$ and $||\cdot||_2$. First, we state the posterior contraction rate results for fixed design:

Theorem 4.2.1 (Bayesian contraction rate in $||\cdot||_n$)**.** *Under the assumptions (A1)(A2) (A3) with the prior (P), the posterior contraction rate with respect to $||\cdot||_n$ will be at least $\varepsilon_n = Cn^{-s/(2s+\varrho)}(\log n)^{D+1}$, for any $s \in \mathbb{R}^+$ and some positive constant C, that is*

$$\Pi(||f - f^*||_n > \varepsilon_n | \{X_i, Y_i\}_{i=1}^n) \to 0 \quad \textit{in probability as } n \to \infty. \tag{4.15}$$

To get the posterior contraction rate for randomly sampled data, we further need post-process the posterior as

(B1) Posterior post-processing: Let $|f^*|_\infty < M$. For any f, denote its truncation $f_M = \min(M, \max(-M, f))$. Then our posterior after post-processing is the posterior of f_M and we denote the corresponding estimator $\hat{f}_M(x) = \int f_M d\Pi(f|\{X_i, Y_i\}_{i=1}^n)$.

Now we can state the rate for random design.

Theorem 4.2.2 (Bayesian contraction rate in $||\cdot||_2$)**.** *Under the assumptions (A1)(A2) (A3), and post-processing (B1) with the prior (P), let ε_n be a multiple of $n^{-s/(2s+\varrho)} \log(n)^{D+1}$. Then*

$$\Pi(\max(||f_M - f^*||_n, ||f_M - f^*||_2) > \varepsilon_n | \{X_i, Y_i\}_{i=1}^n) \to 0 \quad \textit{in probability as } n \to \infty. \tag{4.16}$$

Moreover, with probability tending to one, the following holds:

$$\max(||\hat{f}_M - f^*||_n, ||\hat{f}_M - f^*||_2) \leq C\varepsilon_n \lesssim n^{-s/(2s+\varrho)} (\log(n))^{D+1} \tag{4.17}$$

with ε_n given above and C a positive constant.

Remark. *These theorems suggest that the prior (P) can adapt to the low dimensional structure of $\mathcal{X}$. The ambient dimension D appears on the logarithm terms of the rate in the above theorems. When $\mathcal{X}$ is a manifold, the logarithm term in the above theorems can be further improved to $(\log(n))^{\varrho+1}$ with minor modification of the assumption (B1) on logarithm terms. The proofs' details are discussed in the section 4.5.*

We extend the previous posterior contraction results of the Gaussian process ([vdVvZ08], [CKP14] and [YD16]) to the general set $\mathcal{X}$ via a covering bound of RKHS, which is derived by the covering number of $\mathcal{X}$ ([HS21]).

4.2.3 Convergence rate with miss-specified prior

We next show that, even when the prior is not well specified, under some conditions, one could hope to have a posterior contraction result, while the rate can be sub-optimal. This is shown in the below proposition 4.2.3. To get the proposition 4.2.3, we first modify assumption (A3) to (A3').

(A3') Miss-specified prior support on bandwidth: There exist positive constants c_1, c_2, c_3, a_1, a_2, K_1, K_2, C_1, C_2, and $\varrho_+ \geq \varrho_- > 0$, $\varrho_+ \geq \varrho$ such that

$$\forall t \in [c_1 n^{\frac{-2}{2s+\varrho_+}} \log^{\frac{2(1+D)}{2s+\varrho_+}}(n), c_2 n^{\frac{-2}{2s+\varrho_+}} \log^{\frac{2(1+D)}{2s+\varrho_+}}(n)], \quad p(t) \geq C_1 t^{-a_1} \exp(-\frac{K_1}{t^{\varrho_+/2}}), \tag{4.18}$$

$$\forall t \in (0, c_3 n^{\frac{-2\varrho_+}{(2s+\varrho_+)\varrho_-}} (\log(n))^{\frac{-4(1+D)}{(2+\varrho_+/s)\varrho_-}}], \quad p(t) \leq C_2 t^{-a_2} \exp(-\frac{K_2}{t^{\varrho_-/2}}). \tag{4.19}$$

Remark. *The condition $\varrho_+ \geq \varrho_-$ is to ensure the equation (16) and (17) are consistent with each other, i.e, the intervals will not overlap. The requirement $\varrho^+ \geq \varrho$ is natural as the exponential power of the tail probability $t^{-\varrho_+/2}$ would degenerate to $t^{-\varrho/2}$ when $\varrho > \varrho_+$.*

This can be directly seen through the proof of the following proposition and we leave the details in section 4.5.

Theorem 4.2.3. *Under assumptions (A1)(A2)(A3') with the prior (P), if*

$$\varrho_- > \frac{\varrho_+}{2s + \varrho_+}, \tag{4.20}$$

then the posterior contraction rate with respect to $||\cdot||_n$ will be at least a multiple of $n^{-r(\varrho,\varrho_\pm,s)}(\log(n))^k$ where

$$r(\varrho, \varrho_\pm, s) = \frac{1}{2} - \frac{\varrho_+}{(\rho_- \wedge \varrho)} \frac{\varrho}{2(2s + \varrho_+)}, \tag{4.21}$$

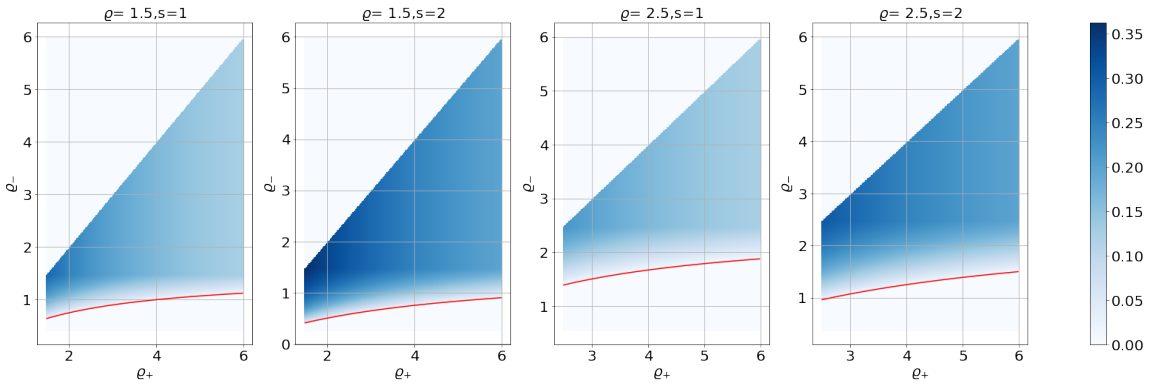

Figure 4.1: $r(\varrho, \varrho_\pm, s)$ changes with $\varrho_\pm$ given ϱ, s. The red lines denote the lower bound $\varrho_- = \frac{\varrho_+}{2s+\varrho_+}$.

$k = (1+D)/(2+\varrho_+/s)$ *if* $\varrho \leq \varrho_-$, *and* $k = \frac{\varrho(1+D)}{2\varrho_- + \varrho_+\varrho_-/s} + \frac{1+D}{2}$ *if* $\varrho > \varrho_-$.

If additionally (B1) holds, then the posterior contraction rate with respect to $||\cdot||_n$ *and* $||\cdot||_2$ *will be at least a multiple of* $n^{-r(\varrho,\varrho_\pm,s)}(\log(n))^k$, *with* $k = (1+D)/(2+\varrho_+/s)$ *if* $\varrho \leq \varrho_-$, *and* $k = \frac{\varrho(1+D)}{2\varrho_- + \varrho_+\varrho_-/s} + \frac{1+D}{2}$ *if* $\varrho > \varrho_-$.

Remark. *This theorem suggests that the prior (P) can adapt to the low dimensional structure of* $\mathcal{X}$ *even when the prior is miss-specified. The rate* $n^{-r(\varrho,\varrho_\pm,s)}$ *is slower than the well specification rate where the well specification rate is* $n^{-\frac{s}{2s+\varrho}} = n^{-(\frac{1}{2}-\frac{\varrho}{2s+\varrho})}$. *The rate* $n^{-r(\varrho,\varrho_\pm,s)}$ *gets better when* ϱ_+ *and* ϱ_- *get closer and finally recover the well specification rate when* $\varrho_\pm = \rho$. *When* $\varrho_- = \varrho_+$, *we fully recover the results in theorem 2.2 of [YD16], where they give the miss-specification rate with miss-specified dimension of the gamma prior as in example 4.2.4. We require the* ϱ_- *is not too small to make sure the rate* $n^{-r(\varrho,\varrho_\pm,s)}$ *is meaningful. Larger* s *and smaller* ϱ_+ *will give less constraints on the feasible* ϱ_-. *Illustration of the feasible region and the change of* $r(\varrho, \varrho_\pm, s)$ *with* $\varrho_\pm$ *are shown in Figure 4.1. The proof of proposition 4.2.3 is discussed in section 4.5.*

4.3 Adaptive rates for Hölder functions on Manifold

In this section, we focus on the special case when $\mathcal{X}$ is manifold. From now on we let $\mathcal{X} = \mathcal{M}$ be a $d-$dimensional compact connected smooth Riemannian manifold (without boundary) isometrically embedded in $[0,1]^D \subset \mathbb{R}^D$ via $\iota : \mathcal{M} \to \mathbb{R}^D$. We will see that we can get optimal contraction rate results for Holder functions with any smoothness $s > 0$ adaptively without knowledge of intrinsic dimension. The key novelties in this section are two folds. Firstly, we propose a new empirical Bayes prior based on kernel count statistics and KNN, which enables us to get the optimal rate without the knowledge of intrinsic dimension. Secondly, we use a new approximation scheme in RKHS to avoid the low-order approximation error induced by curvatures. These allow us to extend the results to any smoothness order of f. The main results are presented in this section while more details of the proofs are left in section 4.5.

4.3.1 Improved approximation in RKHS on Manifold

The key in our proof is to find an approximation of f^* in RKHS and to bound the RKHS norm of the approximation. To get the approximation, we develop an analysis of kernel integral on manifold which extends the previous analysis in literature (lemma 8 in [CL06]). The new analysis, Lemma 4.3.1, applies to any function smoothness order $0 < s < \infty$. Before stating the theorem, we need some further notations. We denote $G_\epsilon(f)(x) = (2\epsilon)^{-d/2}\pi^{-d/2}\int_{\mathcal{M}} h(\frac{||x-y||^2}{2\epsilon})f(y)dV(y) = h_\epsilon * f$. Let $x \in \mathcal{M}$ be a fixed point on manifold $\mathcal{M}$. $T_x\mathcal{M}$ is the tangent space to $\mathcal{M}$ at x and $(e_1, ..., e_d)$ is the fixed orthonormal basis of $T_x\mathcal{M}$. Let us start by introducing normal coordinates as our first consideration. Using the exponential map exp_x, we can generate a group of orthogonal geodesics $(\gamma_1, ..., \gamma_d)$ that intersect at x and have velocities $(e_1, ..e_d)$. For any point y within a sufficiently small neighborhood of x, there exists a corresponding

set of normal coordinates $(s_1, ..., s_d)$ along these geodesics. Thus, for any function f defined on $\mathcal{M}$ in the neighborhood of x, we can view the function f as $\tilde{f} : (s_1, ..., s_d) \to f(y(s_1, ..., s_d))$.

We use the notation ρ to denote the exponential map exp_x and ρ^{-1} to denote its inverse. Specifically, we have $\rho^{-1}(x) = (0, ..., 0)$. Let u be coordinates of y in $(e_1, ..., e_d)$, i.e, $u_i = \langle y - x, e_i \rangle$. Then we have the submanifold locally parametrized as $y = (u, g(u))$ where $g : \mathbb{R}^d \to \mathbb{R}^{D-d}$. Locally, we can have the following diagrams:

$$(s_1, ..., s_d) \overset{exp_x/\rho}{\longleftrightarrow} y \overset{proj}{\longleftrightarrow} u.$$

We will use the multi-index $\boldsymbol{j}$ to express the multi-derivative of a function $\tilde{f}$(denoted as$\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}$) or multi-integral $\int_{\mathbb{R}^d} u^{\boldsymbol{j}} h(||u||^2)du$ when it is needed. We denote the generic homogeneous polynomials of degree i with variables $v = (v_1, ..., v_d)$ as $P_{x,i}(v)$. Let $s > 0$, and suppose $f \in C^s(\mathcal{M}) = C^{\lfloor s \rfloor, \beta}(\mathcal{M})$ where $\lfloor s \rfloor \in \mathbb{N}$, $0 < \beta < 1$, and $\lfloor s \rfloor + \beta = s$. We say that a function f is *polynomially determined* by some other functions $\{g_i\}_{i=1}^N$ if it can be expressed by a multivariate polynomial of $\{g_i\}_{i=1}^N$ as $f = \sum_{j=1}^m \Pi_i g_i^{\alpha_i^j}$ where $N, m, \alpha_i^j \in \mathbb{N}$. We are ready to state our lemma:

Lemma 4.3.1 (Kernel integral on manifold)**.** *Under Assumption (A1), let $h(x) = e^{-x}$, there is $\epsilon_0 > 0$ a constant depending on $\mathcal{M}$, such that $0 < \epsilon < \epsilon_0$, for any $f \in C^s(\mathcal{M}) = C^{\lfloor s \rfloor, \beta}(\mathcal{M})$ where $s > 0$,*

$$(2\epsilon)^{-d/2}\pi^{-d/2} \int_{\mathcal{M}} h(\frac{||x-y||^2}{2\epsilon}) f(y) dV(y) = f(x) + \sum_{i=1}^{\lceil s/2 \rceil - 1} f_i(x)\epsilon^i + O(\epsilon^{s/2}), \quad (4.22)$$

where f_i is only polynomially determined by

(i) Functions which are the derivative of f: $\{\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}(0)\}$ where $|\boldsymbol{j}| \leq 2i$,

(ii) Functions which are the local derivative of the extrinsic manifold coordinate at x, $\{\frac{\partial^{|\boldsymbol{j}|}\rho^{-1}}{\partial u^{\boldsymbol{j}}}(0)\}, \{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}(0)\}$ where $|\boldsymbol{j}| \leq 2i$, and

(iii) Numbers which are finite order moments of the kernel: $\{\int_{\mathbb{R}^d} u^{\boldsymbol{j}} h(||u||^2)du\}$ *where* $|\boldsymbol{j}| \leq \lfloor 2s \rfloor$.

For the $O(\epsilon^{s/2})$ *term, denoted as* $r_{f_\epsilon}(x)$*, it satisfies that* $\sup_{x\in\mathcal{M}} |r_{f_\epsilon}(x)| \leq C(f)\epsilon^{s/2}$*, where* $C(f) = c(\mathcal{M})(1 + \sum_{l=0}^{\lfloor s \rfloor} ||D^l f||_\infty + ||D^{\lfloor s \rfloor} f||_{C^{0,\beta}(\mathcal{M})})$*. Here* $c(\mathcal{M})$ *is some constant, determined by the manifold* $\mathcal{M}$*, only depends on the* $\{||D^l \rho^{-1}||_\infty\}, \{||D^l g||_\infty\}$ *where* $l \leq \lfloor s \rfloor$.

Remark (More general h)**.** *Lemma 4.3.1 can be extended to when* h *is generally a differentiable and decay function on* $[0, \infty)$*. Specifics of such a generalization are given in lemma 4.6.1 in Appendix with proof.*

The boundedness of $O(\epsilon^{s/2})$ term uniform for $x \in \mathcal{M}$ will be used in constructing the RKHS approximation of Hölder f in proposition 4.3.2 to show that the approximation error term can be bounded by the s-order Hölder norm of f. The declared bound is due to that the constant in the $O(\epsilon^{s/2})$ term is bounded by polynomials of (i) $\{||D^l f||_\infty\}$ where $l \leq \lfloor s \rfloor$ and $||D^{\lfloor s \rfloor} f||_{C^{0,\beta}}$, (ii)$\{||D^l \rho^{-1}||_\infty\}, \{||D^l g||_\infty\}$ where $l \leq \lfloor s \rfloor$, and (iii) finite order moments of kernel: $\{\int_{\mathbb{R}^d} u^{\boldsymbol{j}} h(||u||^2)du\}$ where $|\boldsymbol{j}| \leq 2s$.

Finally, we can use the lemma 4.3.1 to construct a better approximation of f.

Proposition 4.3.2. *For* $f \in C^s(\mathcal{M}) = C^{\lfloor s \rfloor, \beta}(\mathcal{M}), s \in \mathbb{R}^+$ *and* $\epsilon < \epsilon_0$*, we can find* $F(x) = f(x) + \sum_{i=1}^{\lceil s/2 \rceil - 1} F_i(x)\epsilon^i$ *where* $F_i(x)$ *are polynomials of* $\{\frac{\partial^{|\boldsymbol{j}|} \tilde{f}}{\partial s^{\boldsymbol{j}}}\}, \{\frac{\partial^{|\boldsymbol{j}|} \rho^{-1}}{\partial u^{\boldsymbol{j}}}\}, \{\frac{\partial^{|\boldsymbol{j}|} g}{\partial u^{\boldsymbol{j}}}\}$ *where* $|\boldsymbol{j}| \leq 2i$*, such that*

$$\sup_{x\in\mathcal{M}} |G_\epsilon(F)(x) - f(x)| \leq C_1 \epsilon^{s/2} \tag{4.23}$$

$$||G_\epsilon(F)(x)||^2_{\tilde{\mathbb{H}}_\epsilon} \leq C_2 \epsilon^{-d/2} \tag{4.24}$$

where the constant $C_1 > 0$ *depends only on the volume of* $\mathcal{M}$ *and* $||f||_\infty$*, and* $C_2 > 0$ *depends only on (i)* $\{||D^l f||_\infty\}$ *where* $l \leq \lfloor s \rfloor$ *and* $||D^{\lfloor s \rfloor} f||_{C^{0,\beta}}$*; (ii))*$\{||D^l \rho^{-1}||_\infty\}$*,* $\{||D^l g||_\infty\}$ *where* $l \leq \lfloor s \rfloor$.

4.3.2 Posterior contractions for Hölder function

Having theorem 4.2.1, 4.2.2, and proposition 4.3.2 in hand, We now give optimal contraction rate results for $f^* \in C^s(\mathcal{M})$ as the following corollary:

Corollary 4.3.3. *Assume (A1) and $f^* \in C^s(\mathcal{M})$, with the prior (P) satisfying with (A3). The posterior contraction rate with respect to $||\cdot||_n$ will be at least $\varepsilon_n = Cn^{-s/(2s+d)}(\log n)^{d+1}$, for any $s \in \mathbb{R}^+$ and some positive constant C, that is*

$$\Pi(||f - f^*||_n > \varepsilon_n | \{X_i, Y_i\}_{i=1}^n) \to 0 \quad \textit{in probability as } n \to \infty. \tag{4.25}$$

Moreover, with post-processing (B1), we have

$$\Pi(\max(||f_M - f^*||_n, ||f_M - f^*||_2) > C_1\varepsilon_n | \{X_i, Y_i\}_{i=1}^n) \to 0 \quad \textit{in probability as } n \to \infty \tag{4.26}$$

and with probability tending to one,

$$\max(||\hat{f}_M - f^*||_n, ||\hat{f}_M - f^*||_2) \leq C_2\varepsilon_n \tag{4.27}$$

with C_1, C_2 positive constants.

4.3.3 Empirical Bayes prior on Bandwidth

In this section, we construct the bandwidth prior satisfied with (A3). To avoid the estimation of intrinsic dimension d in prior construction, we instead estimate bandwidth prior $p(t)$ by kernel count statistics and KNN. Since the normalizing constant of prior and likelihood does not matter in MCMC, we just give the prior up to a constant.

(P') Gaussian process with empirical Bayes prior on bandwidth:

$$W^t | t \sim GP(0, h_t), \tag{4.28}$$

$$p(t) \propto t^{-a_0} \exp(\frac{-b_0}{\hat{v}_n(t)}) \textit{ when } t > T_n^2; \textit{otherwise set } p(t) = 0 \tag{4.29}$$

where $\hat{v}_n(t) = \frac{1}{n}\sum_{i=1}^n \hat{R}_i(t)$ for $\hat{R}_i(t) = \sum_{j\neq i} h_t(X_i, X_j)$, and T_n is the average of $k-$nearest neighbor distance on a subset $\{X_i\}_{i\in\mathcal{S}}$ with $k = \log(n)^4$, where $\mathcal{S} \subset \{1, ..., n\}$. $a_0, b_0 > 0$ are some pre-chosen constants.

In order to make prior (P') satisfies with (A3), we further require the assumption that the density $\mathbb{P}_X$ has a continuous second-order derivative and is bounded from below and above on the manifold $\mathcal{M}$:

(A4) $\mathbb{P}_X \in C^2(\mathcal{M})$ and is uniformly bounded both from below and above, that is

$$0 < p_{\min} \leq \mathbb{P}_X \leq p_{\max} < \infty, \forall x \in \mathcal{M} \tag{4.30}$$

for some $p_{\min}, p_{\max} > 0$.

Details of the implementation of the KNN algorithms are left in appendix B. We will show that the prior (P'), with high probability, is satisfied with the condition (A3). As a result, this prior can get the optimal posterior rate adaptively.

Proposition 4.3.4. *With probability tending to 1 as n goes to $+\infty$, (P') satisfies with (A3) with some suitable constants under (A4).*

Corollary 4.3.5. *As a result of corollary 4.3.3 and proposition 4.3.4, under assumption (A1)(A4) and that $f^* \in C^s(\mathcal{M}), \forall s > 0$, with prior (P'), the posterior contraction rate with respect to $||\cdot||_n$ will be at least a multiple of $n^{-s/(2s+d)}(\log n)^{d+1}$. If additionally (B1) holds, then the posterior contraction rate with respect to $||\cdot||_n$ and $||\cdot||_2$ will be at least a multiple of a multiple of $n^{-s/(2s+d)}(\log n)^{d+1}$*

All proofs are postponed to section 4.5.

4.4 Numerical experiments

In this section, we present various empirical results to validate our theoretical findings. For the estimation of f^*, we do the posterior sampling of the bandwidth to get

the Bayes estimator $\hat{f}$. Specifically, we do the Metropolis-Hasting algorithm on the marginal posterior $p(t|\{X_i, Y_i\}_{i=1}^N)$ of bandwidth, and then do the evaluation of $\hat{f}$ via the closed form of GP condition on each sample of bandwidth and take the average. We leave the numerical implementation details in the appendix.

4.4.1 Swiss roll data

The dataset consists of samples lying on a 2D manifold embedded $\mathbb{R}^3$. Specifically, X is generated according to the following procedure: the data is sampled from a two-dimensional Swiss roll, denoted by $T = (T_1, T_2, T_3)$, in a three-dimensional ambient space as

$$T_1 = U\cos(U), T_2 = V, T_3 = Usin(U) \tag{4.31}$$

where the variables U and V follow uniform distributions $Unif(\frac{2\pi}{2}, \frac{9\pi}{2})$ and $Unif(0, 20)$ respectively. The response variable Y is dependent on X through the relationship

$$Y = 3 * ((\frac{U - 7\pi/2}{3\pi/2})^2 + \frac{(V - 7.5)^2\pi}{15}) + N(0, 0.1^2) \tag{4.32}$$

To evaluate the quality of the fit, we utilize the empirical and population error, represented as $||\hat{f} - f^*||_n$ and $||\hat{f} - f^*||_2$, which measures the discrepancy between our estimator $\hat{f}$ and the true function f^* at the design points and the whole space respectively. In the Gaussian Process (GP) approach, we employ the empirical Bayes approach outlined in section 4.3.3. For each replicate, we conduct 3000 iterations with the first 1000 iterations discarded as burn-in. The average empirical error and population error are then computed over 50 replicates, as shown in figure 4.2. Notably, the GP estimator demonstrates a relatively rapid convergence rate in this example, and its performance is close to the performance of GP with the post-select bandwidth when number of training data n goes large. This is consistent with our theory that

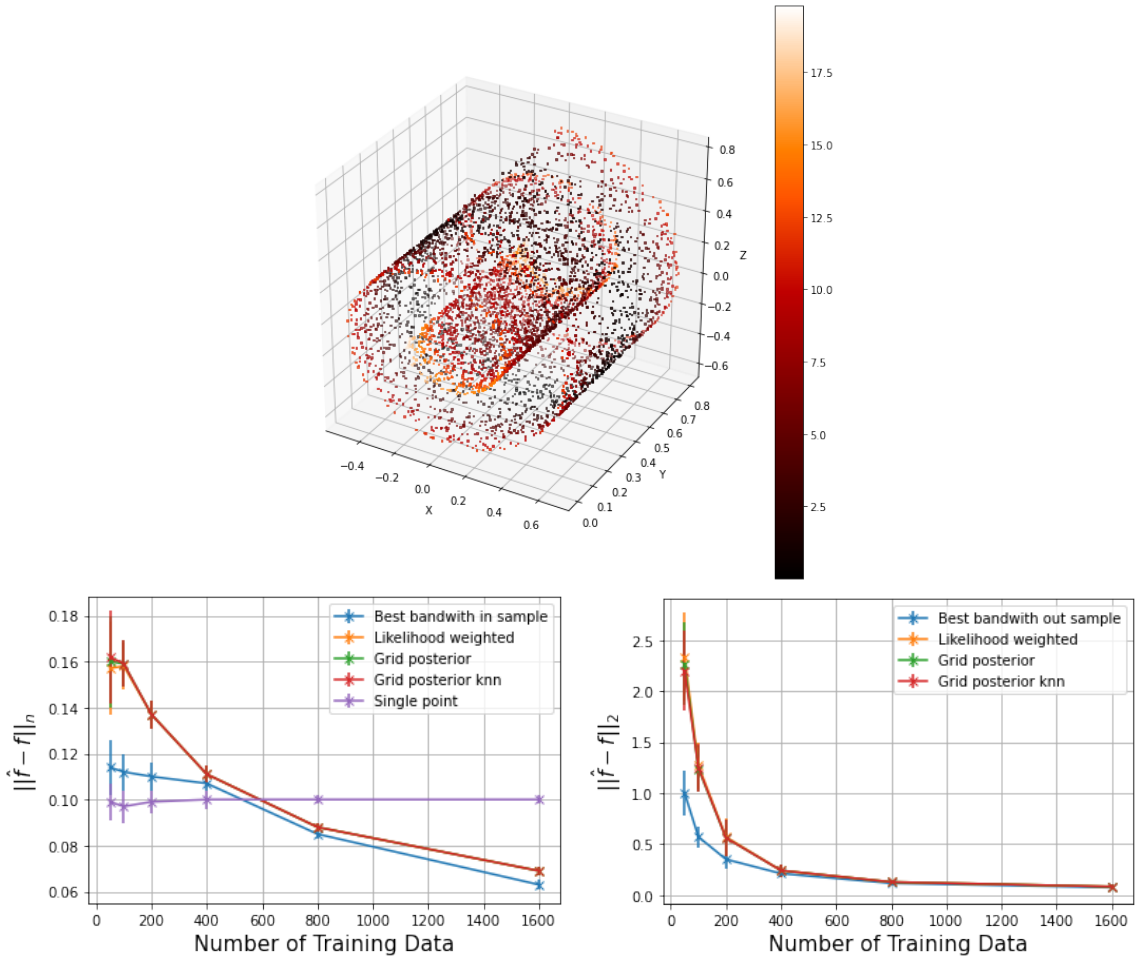

Figure 4.2: The empirical error $||\hat{f} - f||_n$ and population error $||\hat{f} - f||_2$ for GPs with different prior and GP with best post-select bandwidth.

we can use KNN to estimate the bandwidth of prior and get good performance of estimating f^*.

4.4.2 Low dimensional $\mathcal{X}$ of mixed local dimension

We construct a dataset consisting of samples on two subsets, one is the Swiss Roll which is a 2D manifold, and the other is a curve which is a 1D manifold. Both subsets are embedded in $\mathbb{R}^3$, see Figure 4.3 as an illustration. Specifically, X is generated according to the following procedure: With 1/2 probability, we randomly

pick the points from the Swiss roll which is defined in last section with the same $Y's$ dependency on X; with 1/2 probability, we randomly pick the points from the 1d curve $S = (S_1, S_2, S_3)$:

$$S_1 = \frac{7\pi}{2}\cos(\pi U)\cos(4\pi U), S_2 = \frac{7\pi}{2} + \frac{7\pi}{2} cos(\pi U)\sin(4\pi U), S_3 = \frac{7\pi}{2}\sin(\pi U) \quad (4.33)$$

where the variables U follows $Unif(-1, 1)$, and the response variable Y is dependent on X on the curve through the relationship

$$Y = 3 * ((\frac{\sqrt{S_1^2 + S_3^2} - 7\pi/2}{3\pi/2})^2 + \frac{(S_2 - 7.5)^2\pi}{15}) + N(0, 0.1^2) \quad (4.34)$$

To evaluate the quality of the fit, we again use the empirical and population error, represented as $||\hat{f} - f^*||_n$ and $||\hat{f} - f^*||_2$. We employ the empirical Bayes approach outlined in section 4.3.3. For each replicate, we conduct 3000 iterations with the first 1000 iterations discarded as burn-in. The average empirical error and population error are then computed over 50 replicates, as shown in figure 4.3. Consistent with section 4.4.1, figure 4.3 shows that the GP estimator has a relatively rapid convergence rate in this example, and its performance is close to the performance of GP with the post-select bandwidth when number of training data n goes large. This is consistent with our theory that we can estimate f^* even when $\mathcal{X}$ is not manifold but a general subset.

4.4.3 Lucky Cat data

The lucky cat data from [NNM+96] exhibits an intrinsic dimensionality of one, which corresponds to the rotation angle θ. We possess knowledge of the true value of θ, allowing us to create the ground truth function $f(\theta) = \cos(\theta)$, which represents a continuous function on the unit circle. To simulate the responses, we introduce independent Gaussian noises $w_i \sim N(0, 0.1^2)$ to the true values, resulting in $Y_i =$

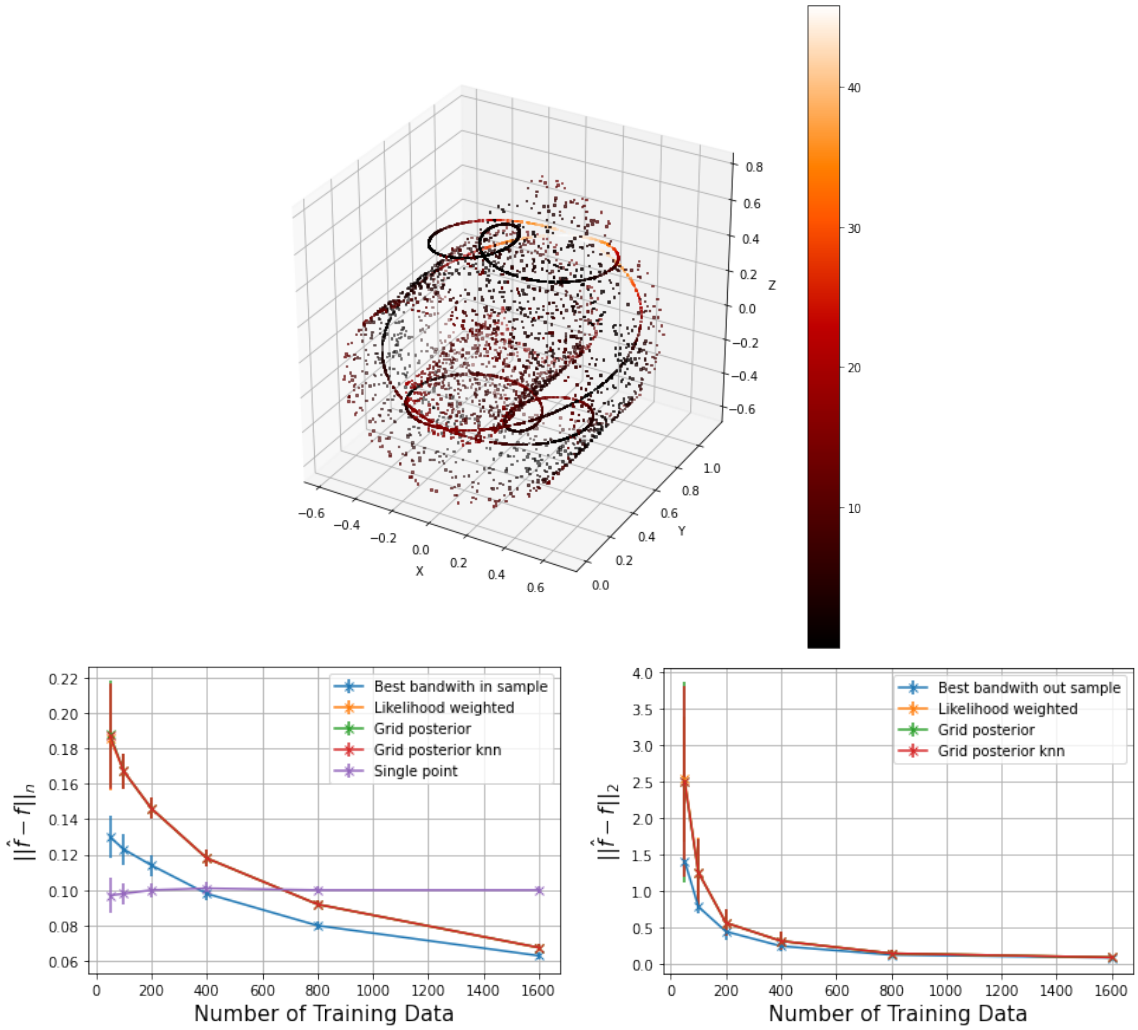

Figure 4.3: $||\hat{f} - f||_n$ and $||\hat{f} - f||_2$ for GPs with different prior and GP with best post-select bandwidth.

$f^*(\theta) + w_i$. In this model, the total sample size is $N = 72$, and the predictors X_i reside in $\mathbb{R}^D$ with $D = 16,384$.

To evaluate the impact of the sample size n on the fitting performance, we randomly partition $n = 18, 36$, and 64 samples into a training set, treating the remaining samples as a testing set. The training set is employed for model fitting, while the testing set is used to assess the estimation accuracy. This process is repeated m = 100 times for each training size n, and population error $||\hat{f} - f^*||_2$ is calculated on the testing set and taken average.

We apply Gaussian Process (GP) with empirical Bayes prior on bandwidth as well as GP with post select bandwidth. To assess prediction performance, we compare our GPs with Lasso [Tib96] and elastic net (EN) [ZH05] under the same settings. These competing models were selected due to their wide usage in high-dimensional regression settings and their notable performance, particularly in sparse model scenarios. Additionally, for each simulation, we run 3,000 iterations with the first 1000 iterations designated as burn-in.

The results, presented in figure 4.4, demonstrate that under each training size n, the GP models are consistent with each other and outperforms the other methods. Furthermore, as n increases, the prediction error of the GP model decays much faster compared to the EN and Lasso methods.

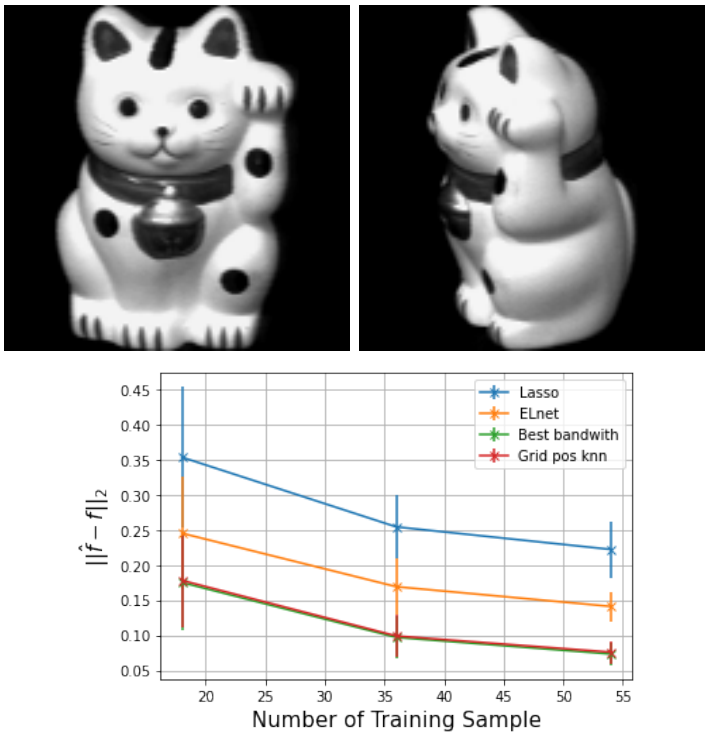

Figure 4.4: $||\hat{f} - f||_2$ for Gps with different prior and Gp with best post-select bandwidth.

4.5 Proofs

4.5.1 Proofs in sec 4.2

The proofs of theorem 4.2.1 and 4.2.2 are similar to the proofs in [VdVVZ09] and [YD16] Before proceeding the proof, we first establish some results similar to the previous papers adaptive to our settings in the appendix.

Proofs for Them 4.2.1. First, we need to prove that the below equations will be satisfied for ε_n a multiple of $n^{\frac{-s}{2s+d}}\log(n)^{k_1}$ and $\bar{\varepsilon}_n$ a multiple of $\varepsilon_n(\log(n))^{k_2}$ with $k_1 = \frac{1+D}{2+d/s}$ and $k_2 = \frac{1+D}{2}$

$$P(||W^t - f^*||_\infty \leq \varepsilon_n) \geq e^{-n\varepsilon_n^2}$$

$$P(W^t \notin B_n) \leq e^{-4n\varepsilon_n^2}$$

$$\log N(\bar{\varepsilon}_n, B_n, ||\cdot||_\infty) \leq n\bar{\varepsilon}_n^2$$

with some Borel measurable subsets B_n of $C(\mathcal{X})$ and for n sufficiently large. This implies that, by [vdVvZ08], the posterior contraction rate with respect to $||\cdot||_n$ will be at least $\varepsilon_n \vee \bar{\varepsilon}_n \leq n^{-s/(2s+d)}(\log(n))^{D+1}$ and this will finish the proof.

To prove the above equations, we define the centered and decentered concentration function of the Gaussian process W^{t^*} for each fixed bandwidth t^* by

$$\phi_0^{t^*}(\varepsilon') = -\log P(||W^{t^*}||_\infty \leq \varepsilon')$$

$$\phi_{f^*}^{t^*}(\varepsilon') = \inf_{q \in \tilde{\mathbb{H}}_{t^*}: ||q - f^*||_\infty \leq \varepsilon'} ||q||^2_{\tilde{\mathbb{H}}_{t^*}} - \log P(||W^{t^*}||_\infty \leq \varepsilon')$$

where the sup norm $||\cdot||_\infty$ is on $\mathcal{X}$. By definition, we know $P(||W^{t^*}||_\infty \leq \varepsilon') = \exp(-\phi_0^{t^*}(\varepsilon'))$, and

$$P(||W^{t^*} - f^*||_\infty \leq 2\varepsilon') \geq e^{-\phi_{f^*}^{t^*}(\varepsilon')}$$

by [KLL94]. By conditions (A2) and lemma 4.6.7, for $t^* \leq \min(\epsilon_0, (\varepsilon_0/C)^{2/s}$,

1) and $\varepsilon_0 > \varepsilon' > C(t^*)^{s/2}$,

$$\phi_{f^*}^{t^*}(\varepsilon') \leq D(t^*)^{-d/2} + C_4(t^*)^{-d/2}(\log(\frac{1}{\sqrt{t^*}\varepsilon'}))^{D+1} \leq K_1(t^*)^{-d/2}(\log(\frac{1}{\sqrt{t^*}\varepsilon'}))^{D+1}$$

By condition (A3), There exists $c_1, c_2, c_3 a_1, a_2, K_1, K_2, C_1, C_2 > 0$, such that

$$\forall t \in [c_1 n^{\frac{-2}{2s+\varrho}} \log^{\frac{1+D}{2s+\varrho}}(n), c_2 n^{\frac{-2}{2s+\varrho}} \log^{\frac{1+D}{2s+\varrho}}(n)], p(t) \geq C_1 t^{-a_1} \exp(-\frac{K_1}{t^{\varrho/2}}), \tag{4.35}$$

$$\forall t \in (0, c_3 n^{\frac{-2}{2s+\varrho}}], p(t) \leq C_2 t^{-a_2} \exp(-\frac{K_2}{t^{\varrho/2}}) \tag{4.36}$$

Therefore, we have bounds of probability

$$\begin{aligned} P(||W^{t^*} - f^*||_\infty \leq 2\varepsilon') &\geq P(||W^{t^*} - f^*||_\infty \leq 2\varepsilon', t^* \in [(C\varepsilon')^{2/s}, c_4(C\varepsilon')^{2/s}]) \\ &\geq \int_{(C\varepsilon')^{2/s}}^{c_3(C\varepsilon')^{2/s}} e^{-\phi_{f^*}^{t^*}(\varepsilon')} p(t^*) dt^* \\ &\geq C_3 e^{-K_3(1/\varepsilon')^{\varrho/s}(\log(1/\varepsilon'))^{D+1}} (C/\varepsilon')^{2a_1/s} (C\varepsilon')^{2/s} \end{aligned}$$

for ε' that satisfies $[(C\varepsilon')^{2/s}, c_4(C\varepsilon')^{2/s}] \subset [c_1 n^{\frac{-2}{2s+\varrho}} \log^{\frac{1+D}{2s+\varrho}}(n), c_2 n^{\frac{-2}{2s+\varrho}} \log^{\frac{1+D}{2s+\varrho}}(n)]$. Therefore

$$P(||W^t - f^*||_\infty \leq \varepsilon_n) \geq \exp(-n\varepsilon_n^2)$$

for ε_n a large multiple of $n^{-s/(2s+\varrho)} \log(n)^{k_1}$ with $k_1 = (1+D)/(2+\varrho/s)$ when n is large enough. Similar to proof of theorem 3.1 in [VdVVZ09] and theorem 2.1 in [YD16], we have

$$B_{N,r,\delta,\varepsilon'} = (N\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}_{r^{-2}}^1 + \varepsilon'\mathbb{B}_1) \cup (\bigcup_{t^* > \delta^{-2}} (N\tilde{\mathbb{H}}_{t^*}^1) + \varepsilon'\mathbb{B}_1))$$

with $\mathbb{B}_1$ is the unit ball of $C(\mathcal{X})$, and it contains the set $N\tilde{\mathbb{H}}_{t^*}^1 + \varepsilon'\mathbb{B}_1$ for any $t^* \in [r^{-2}, \delta^{-2}]$. And we claim, if $r^{-2} < c_3 n^{\frac{-2}{2s+\varrho}} (\log(n))^{\frac{-4(1+D)}{(2+\varrho/s)\varrho}}$, $e^{-\phi_0^{r^{-2}}(\varepsilon')} < 1/4$ and $N \geq 4\sqrt{\phi_0^{r^{-2}}(\varepsilon')}$, then

$$P(W^t \notin B_{N,r,\delta,\varepsilon'}) \leq \frac{2C_5 r^{2(a_2-\varrho+1)} e^{-K_4 r^\varrho}}{K_4 \varrho} + e^{-N^2/8}.$$

This is because $P(W^t \notin B_{N,r,\delta,\varepsilon'}) \leq P(t^* < r^{-2}) + \int_{r^{-2}}^{\infty} P(W^{t^*} \notin B_{N,r,\delta,\varepsilon'})p(t^*)dt^*$. For the first term, we have $P(t^* < r^{-2}) \leq \int_0^{r^{-2}} C_2(t^*)^{-a_2} \exp(-K_2(t^*)^{-\varrho/2})dt^* \leq \frac{2C_4 r^{2(a_2-\varrho+1)}e^{-K_4 r^{\varrho}}}{K_4\varrho}$ if $r^{-2} < c_3 n^{\frac{-2}{2s+\varrho}}(\log(n))^{\frac{-4(1+D)}{(2+\varrho/s)\varrho}}$ (lemma 4.9 [VdVVZ09] and condition (B1)). Now we deal with the second term. For any fixed $t^* > r^{-2}$, we have $P(W^{t^*} \notin B_{N,r,\delta,\varepsilon'}) \leq P(W^{t^*} \notin N\tilde{\mathbb{H}}^1_{t^*} + \varepsilon'\mathbb{B}_1) \leq 1 - \Phi(\Phi^{-1}(e^{-\phi_0^{t^*}(\varepsilon')}) + N) \leq 1 - \Phi(\Phi^{-1}(e^{-\phi_0^{r^{-2}}(\varepsilon')}) + N)$ (last second step by Borell inequality) as $e^{-\phi_0^{t^*}(\varepsilon')} = P(\sup|W^{t^*}| \leq \varepsilon')$ is increasing with t^*. So when $N \geq -2\Phi^{-1}(e^{-\phi_0^{r^{-2}}(\varepsilon')})$, the right hand side is bounded by $1 - \Phi(N/2) \leq e^{-M^2/8}$. The later condition is satisfied when $e^{-\phi_0^{r^{-2}}(\varepsilon')} < 1/4$ and $N \geq 4\sqrt{\phi_0^{r^{-2}}(\varepsilon')}$ by lemma 4.10 [VdVVZ09].

We know the conditions are satisfied if

$$N^2 \geq 16C_4 r^{\varrho}(\log(r/\varepsilon'))^{1+D}, \quad r > \max(1, \frac{1}{\sqrt{\epsilon_0}}, \sqrt{c_3}n^{\frac{1}{2s+\varrho}}(\log(n))^{\frac{2(1+D)}{(2+\varrho/s)\varrho}}), \quad \varepsilon' < \varepsilon_0'$$

for some fixed $\epsilon_0, \varepsilon_0' > 0$ by lemma 4.6.7 and condition (A1). Therefore,

$$P(W^t \notin B_{N,r,\delta,\varepsilon_n}) \leq \exp(-C_0 n\varepsilon_n^2)$$

for r and N satisfying

$$r^{\varrho} = \frac{2C_0}{K_4}n\varepsilon_n^2, \quad N^2 = \max(8C_0, 16C_4)n\varepsilon_n^2(\log(r/\varepsilon_n))^{1+D}.$$

Denote r_n, N_n the solutions of the above equations. By lemma 4.6.5, for $N\sqrt{r/\delta} > 2\varepsilon'$ and $r > \max(1, \frac{1}{\sqrt{\epsilon_0}}, \sqrt{c_3}n^{\frac{1}{2s+\varrho}}(\log(n))^{\frac{2(1+D)}{(2+\varrho/s)\varrho}})$,

$$\log N(2\varepsilon', N\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}^1_{r^{-2}} + \varepsilon'\mathbb{B}_1, ||\cdot||_\infty) \leq \log N(\varepsilon', N\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}^1_{r^{-2}}, ||\cdot||_\infty)$$
$$\leq Kr^d(\log(\frac{N\sqrt{r/\delta}}{\varepsilon'}))^{1+d}$$

By lemma 4.6.4, every element of $N\tilde{\mathbb{H}}^1_{t^*}$ for $t^* > \delta^{-2}$ is uniformly at most $\delta\sqrt{D}\tau N$ distant from a constant function for a constant in the interval $[-N, N]$. Therefore for $\varepsilon' > \delta\sqrt{D}\tau N$,

$$N(3\varepsilon', \bigcup_{t^*>\delta^{-2}} (N\tilde{\mathbb{H}}^1_{t^*}) + \varepsilon'\mathbb{B}_1), ||\cdot||_\infty) \leq N(\varepsilon', [-N, N], |\cdot|) \leq \frac{2N}{\varepsilon'}.$$

With $\delta = \varepsilon'/(2\sqrt{D}\tau N)$, combining the above displays, for $B = B_{N,r,\delta,\varepsilon'}$ with

$$N \geq \varepsilon', N^{3/2}\sqrt{2\tau r}D^{1/4} \geq 2(\varepsilon')^{3/2}, r > \max(1, \frac{1}{\sqrt{\epsilon_0}}, \sqrt{c_3}n^{\frac{1}{2s+\varrho}}(\log(n))^{\frac{2(1+D)}{(2+\varrho/s)\varrho}})$$

which is satisfied when $r = r_n$ and $N = N_n$ with big enough n, we have

$$\log N(3\varepsilon', B, ||\cdot||_\infty) \leq Kr^d(\log(\frac{N^{3/2}\sqrt{2\tau r}D^{1/4}}{(\varepsilon')^{3/2}}))^{1+d} + \log(\frac{2N}{\varepsilon'}).$$

Now for $B_n = B_{N_n,r_n,\delta_n,\varepsilon_n}$, we have

$$\log N(\bar{\varepsilon}_n, B_n, ||\cdot||_\infty) \leq n\bar{\varepsilon}_n^2$$

for $\bar{\varepsilon}_n$ a large multiple of $\varepsilon_n(\log(n))^{k_2}$ with $K_2 = (1 + D)/2$. This finishes the proof for general $\mathcal{X}$.

For the case when $\mathcal{X}$ is a manifold, we can use the manifold part of lemma 4.6.7 and use lemma 4.6.6 instead of 4.6.5. Then the logarithm term can be improved to $\log^{d+1}(n)$. And this finishes the proof. □

Proof of theorem 4.2.2 is based on theorem 4.2.1 and the comparison between $||\cdot||_n$ and $||\cdot||_{L_2}$.

Proof for theorem 4.2.2. To prove theorem 4.2.2, we use the lemma 6.1 and 6.2 in [YD16] to boost the result of theorem 4.2.1, and the proof then is exactly the same as their theorem 2.3. □

Proof for proposition 4.2.3. First, we recollect the three conditions in the poof of theorem 4.2.1 and 4.2.2. Under condition (A3'), the first condition becomes

$$\begin{aligned} P(||W^{t^*} - f^*||_\infty \le 2\varepsilon') &\ge P(||W^{t^*} - f^*||_\infty \le 2\varepsilon', t^* \in [(C\varepsilon')^{2/s}, c_4(C\varepsilon')^{2/s}]) \\ &\ge \int_{(C\varepsilon')^{2/s}}^{c_4(C\varepsilon')^{2/s}} e^{-\phi_{f^*}^{t^*}(\varepsilon')} p(t^*) dt^* \\ &\ge C_3 e^{-K_3(1/\varepsilon')^{(\varrho\vee\varrho_+)/s}(\log(1/\varepsilon'))^{1+D}} (\frac{C}{\varepsilon'})^{a_1/s} (C\varepsilon')^{2/s} \end{aligned}$$

The second condition becomes

$$P(W^t \notin B_{N,r,\delta,\varepsilon'}) \le \frac{2C_5 r^{2(a_2-\varrho_-+1)} e^{-K_4 r^{\varrho_-}}}{K_4 \varrho_-} + e^{-N^2/8}$$

with $N^2 \ge 16C_4 r^\varrho (\log(r/\varepsilon'))^{1+D}$, $r > \max(1, \frac{1}{\sqrt{\epsilon_0}}, \sqrt{c_3} n^{\frac{1}{2s+\varrho_-}} (\log(n))^{\frac{2(1+D)}{(2+\varrho_-/s)\varrho_-}})$, $\epsilon' < \epsilon_0'$ where ϵ_0' is the constant in lemma 4.6.7. The third condition on entropy number is given by

$$\log N(3\varepsilon', B, ||\cdot||_\infty) \le K r^\varrho (\log(\frac{N^{3/2}\sqrt{2\tau r} D^{1/4}}{(\varepsilon')^{3/2}}))^{1+D} + \log(\frac{2N}{\varepsilon'})$$

1. $\varrho_+ \ge \varrho$. Now take $\varepsilon_n = n^{-s/(2s+\varrho_+)}(\log(n))^{k_1}$ with $k_1 = (1+D)/(2+\varrho_+/s)$, and

$$r^{\varrho_-} = \frac{2C_0}{K_4} n\varepsilon_n^2, \quad N^2 = \max(8C_0, 16C_4) n\varepsilon_n^2 (\log(r/\varepsilon_n))^{1+D}.$$

   Then we can get $\bar{\varepsilon}_n$ a multiple of $n^{-\frac{\varrho_-(2s+\varrho_+)-\varrho_+\varrho}{2(2s+\varrho_+)\varrho_-}}(\log(n))^{k_2}$ with $k_2 = \frac{\varrho(1+D)}{2\varrho_-+\varrho_+\varrho_-/s} + \frac{1+D}{2}$. And the validity of $\bar{\varepsilon}_n$ gives the conditions $\varrho_- > \frac{\varrho_+\varrho}{2s+\varrho_+}$.

2. $\varrho > \varrho_+$. Now take $\varepsilon_n = n^{-s/(2s+\varrho)}(\log(n))^{k_3}$ with $k_3 = (1+D)/(2+\varrho/s)$, and

$$r^{\varrho_-} = \frac{2C_0}{K_4} n\varepsilon_n^2, \quad N^2 = \max(8C_0, 16C_4) n\varepsilon_n^2 (\log(r/\varepsilon_n))^{1+D}.$$

Then we can get $\bar{\varepsilon}_n$ a multiple of $n^{-\frac{\varrho_-(2s+\varrho)-\varrho^2}{2(2s+\varrho)\varrho_-}}(\log(n))^{k_4}$ with $k_4 = \frac{\varrho(1+D)}{2\varrho_-+\varrho\varrho_-/s} + \frac{1+D}{2}$. And the validity of $\bar{\varepsilon}_n$ gives the conditions $\varrho_- > \frac{\varrho^2}{2s+\varrho}$. So the final rate will be $\bar{\varepsilon}_n \wedge \varepsilon_n = n^{-\frac{\varrho_-(2s+\varrho)-\varrho^2}{2(2s+\varrho)\varrho_-}}(\log(n))^{k_4}$.

To summarize, the final rate will be $\bar{\varepsilon}_n \wedge \varepsilon_n =$

1. $\rho > \rho_-$: $\bar{\varepsilon}_n = n^{-\frac{\varrho_-(2s+\varrho_+\vee\varrho)-(\varrho_+\vee\varrho)\varrho}{2(2s+\varrho_+\vee\varrho)\varrho_-}}(\log(n))^{k_5} = n^{-(\frac{1}{2}-\frac{\varrho_+\vee\varrho}{\rho_-}\frac{\varrho}{2(2s+\varrho_+\vee\varrho)})}(\log(n))^{k_5}$
   $= n^{-r(\varrho,\varrho_\pm,s)}(\log(n))^{k_5}$ where $k_5 = \frac{\varrho(1+D)}{2\varrho_-+(\varrho_+\vee\varrho)\varrho_-/s} + \frac{1+D}{2}$

2. $\rho \leq \rho_-$: $\varepsilon_n = n^{-s/(2s+\varrho_+)}(\log(n))^{k_6}$ where $k_1 = (1+D)/(2+\varrho_+/s)$

and the condition $\varrho_- > \frac{\varrho_+\vee\varrho}{2s+(\varrho_+\vee\varrho)}$ is needed to let the rate valid. This leads to the natural assumption $\varrho_+ > \varrho$ and change all the $\varrho_+ \vee \varrho$ to ϱ_+. And this finishes the proof. □

4.5.2 Proofs in section 4.3

The proof strategy of lemma 4.3.1 are similar to lemma 8 in [CL06]. The key difference here is that we extend their expansions into higher orders without keeping explicit coefficients.. To finish the proof, we need a lemma on generic homogeneous polynomials and extensions of lemma 6 and 7 in [CL06], which we leave in the Appendix.

Proof for lemma 4.3.1. First we prove the theorem when $s \in \mathbb{N}$. Following the proof of lemma A.5 in [CW22], we first write $G_\epsilon f$, with truncation and for $2\epsilon < \epsilon_0$. ϵ_0 is defined as in lemma A.5 in [CW22] such that the conditions of their lemma A.3

is satisfied. Now we get

$$\begin{aligned} G_\epsilon f &= (2\epsilon)^{-d/2}\pi^{-d/2}\int_{y\in\mathcal{M}:||y-x||\le\sqrt{a_0 2\epsilon\log(\frac{1}{2\epsilon})}} h(\frac{||x-y||^2}{2\epsilon})f(y)dV(y) \\ &+ (2\epsilon)^{-d/2}\pi^{-d/2}\int_{y\in\mathcal{M}:||y-x||>\sqrt{a_0 2\epsilon\log(\frac{1}{2\epsilon})}} h(\frac{||x-y||^2}{2\epsilon})f(y)dV(y) \\ &:= a_1 + a_2 \end{aligned} \tag{4.37}$$

where $a_0 = \frac{1}{0.81}\frac{d+s+10}{a} = \frac{1}{0.9^2}a'$ and $a = 1$. Then the second term $a_2 \le C\epsilon^{s+10}$ by similar computation of [CW22]. It is enough to finish the proof by bounding a_1. To get the estimation of a_1, we do the Taylor expansion on each term of the integral, similar to [CL06]. By the mapping $\tilde{f} : (s_1, ..., s_d) \to f(y(s_1, ..., s_d))$, we get:

$$f(y) = f(x) + \sum_{|\boldsymbol{j}|=1}^{\lceil s\rceil-1}\frac{1}{\boldsymbol{j}!}\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}s^{\boldsymbol{j}} + O(||\sum_{i=1}^{d} s_i^2||^{s/2}). \tag{4.38}$$

Combined with lemma 4.6.9, we further get:

$$f(y) = \tilde{f}(0) + \sum_{i=1}^{\lceil s\rceil-1} P_{x,i}^f(u) + O(||u||^s) \tag{4.39}$$

where coefficients of the terms in homogeneous polynomials $P_{x,j}^f(u)$ just depend on $\{\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}(0)\}$,$\{\frac{\partial^{|\boldsymbol{j}|}\rho^{-1}}{\partial u^{\boldsymbol{j}}}(0)\}, \{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}(0)\}$ where $|\boldsymbol{j}| \le i$, and the residual term just depends on the $\{\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}\}$,$\{\frac{\partial^{|\boldsymbol{j}|}\rho^{-1}}{\partial u^{\boldsymbol{j}}}\}, \{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}\}$ where $|\boldsymbol{j}| \le s$.

Next we do the Taylor expansion of kernel $h(\frac{||x-y||^2}{2\epsilon})$ to $\lceil s/2 \rceil$ order:

$$
\begin{aligned}
h(\frac{||x-y||^2}{2\epsilon}) &= h(\frac{||u||^2}{2\epsilon} + (\frac{||x-y||^2}{2\epsilon} - \frac{||u||^2}{2\epsilon})) \\
&= h(\frac{||u||^2}{2\epsilon}) + \sum_{i=1}^{\lceil s/2 \rceil - 1} \frac{1}{i!} h^{(i)}(\frac{||u||^2}{2\epsilon})(\frac{||x-y||^2}{2\epsilon} - \frac{||u||^2}{2\epsilon})^i \\
&+ O(h^{(\lceil s/2 \rceil)}(\frac{||x-y||^2}{2\epsilon} - \frac{||u||^2}{2\epsilon})^{\lceil s/2 \rceil}) \\
&= h(\frac{||u||^2}{2\epsilon}) + \sum_{i=1}^{\lceil s/2 \rceil - 1} C_i h^{(i)}(\frac{||u||^2}{2\epsilon}) \frac{(\sum_{j=4}^{\lceil s \rceil + 1} Q_{x,j}(u) + O(||u||^{\lceil s \rceil + 2}))^i}{\epsilon^i} \\
&+ O(h^{(\lceil s/2 \rceil)} \frac{||u||^{4\lceil s/2 \rceil}}{\epsilon^{\lceil s/2 \rceil}})
\end{aligned}
\tag{4.40}
$$

Combined with the lemma 4.6.10, now we can write $\pi^{d/2}(2\epsilon)^{d/2}a_1$ as

$$
\begin{aligned}
(2\epsilon)^{d/2}\pi^{d/2}a_1 &= \int_{y \in \mathcal{M}: ||y-x|| \leq \sqrt{a_0 2\epsilon \log(\frac{1}{2\epsilon})}} h(\frac{||x-y||^2}{2\epsilon}) f(y) dV(y) \\
&= \int_{||(u,g(u))|| \leq \sqrt{a_0 2\epsilon \log(\frac{1}{2\epsilon})}} (h(\frac{||u||^2}{2\epsilon}) + \sum_{i=1}^{\lceil s/2 \rceil - 1} C_i h^{(i)}(\frac{||u||^2}{2\epsilon}) * \\
&\frac{(\sum_{j=4}^{\lceil s \rceil + 1} Q_{x,j}(u) + O(||u||^{\lceil s \rceil + 2}))^i}{\epsilon^i} + O(h^{(\lceil s/2 \rceil)} \frac{||u||^{4\lceil s/2 \rceil}}{\epsilon^{\lceil s/2 \rceil}})) * \\
&(\tilde{f}(0) + \sum_{i=1}^{\lceil s \rceil - 1} P_{x,i}^f(u) + O(||u||^s)) * (1 + V_{x,2}(u) + \sum_{i=3}^{\lceil s \rceil - 1} V_{x,i}(u) + O(||u||^{\lceil s \rceil})) du \\
&= \int_{||(u,g(u))|| \leq \sqrt{a_0 2\epsilon \log(\frac{1}{2\epsilon})}} (h(\frac{||u||^2}{2\epsilon}) + \sum_{i=1}^{\lceil s/2 \rceil - 1} C_i h^{(i)}(\frac{||u||^2}{2\epsilon}) * \\
&\frac{(\sum_{j=4}^{\lceil s \rceil + 1} Q_{x,j}(u) + O(||u||^{\lceil s \rceil + 2}))^i}{\epsilon^i}) * \\
&(\tilde{f}(0) + \sum_{i=1}^{\lceil s \rceil - 1} P_{x,i}(u)) + O(h^{(\lceil s/2 \rceil)} \epsilon^{d/2 - \lceil s/2 \rceil} ||u||^{4\lceil s/2 \rceil}) + O(h\epsilon^{d/2} ||u||^s) du
\end{aligned}
\tag{4.41}
$$

where the last line comes from the product of the expansions of f (s order) and dV ($\lceil s \rceil$ order). By lemma 4.6.8, we know coefficients of each term in $P_{x,i}(u)$ is polynomials of $\{\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}(0)\}, \{\frac{\partial^{|\boldsymbol{j}|}\rho^{-1}}{\partial u^{\boldsymbol{j}}}(0)\}, \{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}(0)\}$ where $|\boldsymbol{j}| \leq i$.

Since $2\epsilon < \epsilon_0$, we can recall lemma A.3 in [CW22] and have $0.9||y-x||_{\mathbb{R}^D} = 0.9||(u, g(u))||_{\mathbb{R}^D} \leq ||u(y)||_{\mathbb{R}^d} \leq 1.1||y-x||_{\mathbb{R}^D}$. We use the condition $0.9||y-x||_{\mathbb{R}^D} \leq ||(u, g(u))||_{\mathbb{R}^D}$ and exponential decay of $h/h^{(l)}$ to extend the integral domain to $\mathbb{R}^d$, by adding a small order term $O(\epsilon^{(s+10+d)/2})$:

$$
\begin{aligned}
(2\epsilon)^{d/2}\pi^{d/2}a_1 &= \int_{\mathbb{R}^d} h(\frac{||u||^2}{2\epsilon})\tilde{f}(0)du + \int_{\mathbb{R}^d} h(\frac{||u||^2}{2\epsilon}) \sum_{i=1}^{\lceil s \rceil - 1} P_{x,i}(u)du \\
&+ \sum_{i=1}^{\lceil s/2 \rceil - 1} \int_{\mathbb{R}^d} C_i \frac{(\sum_{j=4}^{\lceil s \rceil + 1} Q_{x,j}(u) + O(||u||^{\lceil s \rceil + 2}))^i}{\epsilon^i} (\tilde{f}(0) + \sum_{k=1}^{\lceil s \rceil - 1} P_{x,k}(u)) h^{(i)}(\frac{||u||^2}{2\epsilon})du \\
&+ \int_{\mathbb{R}^d} O(h^{(\lceil s/2 \rceil)} \epsilon^{d/2 - \lceil s/2 \rceil} ||u||^{4\lceil s/2 \rceil}) + O(h\epsilon^{d/2}||u||^s)du + O(\epsilon^{(s+10+d)/2})
\end{aligned}
\tag{4.42}
$$

So we get

$$
\begin{aligned}
a_1 &= (2\epsilon)^{-d/2}\pi^{-d/2}[\int_{\mathbb{R}^d} h(\frac{||u||^2}{2\epsilon})\tilde{f}(0)du] + (2\epsilon)^{-d/2}\pi^{-d/2}[\int_{\mathbb{R}^d} h(\frac{||u||^2}{2\epsilon}) \sum_{i=1}^{\lceil s \rceil - 1} P_{x,i}(u)du] \\
&+ (2\epsilon)^{-d/2}\pi^{-d/2}[\sum_{i=1}^{\lceil s/2 \rceil - 1} \int_{\mathbb{R}^d} C_i \frac{(\sum_{j=4}^{\lceil s \rceil + 1} Q_{x,j}(u) + O(||u||^{\lceil s \rceil + 2}))^i}{\epsilon^i} (\tilde{f}(0) + \\
&\sum_{k=1}^{\lceil s \rceil - 1} P_{x,k}(u)) h^{(i)}(\frac{||u||^2}{2\epsilon})du] + (2\epsilon)^{-d/2}\pi^{-d/2}[\int_{\mathbb{R}^d} O(h^{(\lceil s/2 \rceil)} \epsilon^{d/2 - \lceil s/2 \rceil} ||u||^{4\lceil s/2 \rceil}) \\
&+ O(h\epsilon^{d/2}||u||^s)du] + O(\epsilon^{(s+10)/2})
\end{aligned}
\tag{4.43}
$$

Now we do the variables changing $u \to \sqrt{2\epsilon}u$, and we get the last integral

$$
(2\epsilon)^{-d/2}\pi^{-d/2} \int_{\mathbb{R}^d} O(h^{(\lceil s/2 \rceil)} \epsilon^{d/2 - \lceil s/2 \rceil} ||u||^{4\lceil s/2 \rceil}) + O(h\epsilon^{d/2}||u||^s)du = O(\epsilon^{s/2}) \tag{4.44}
$$

and its coefficient is satisfied with the dependency in our lemma 4.3.1 by tracking the constant and bounds of remainders for expansion. For the first two terms of a_1, we get

$$(2\epsilon)^{-d/2}\pi^{-d/2}\int_{\mathbb{R}^d} h(\frac{||u||^2}{2\epsilon})\tilde{f}(0)du = \tilde{f}(0)\int_{\mathbb{R}^d}\pi^{-d/2}h(||u||^2)du = \tilde{f}(0) = f(x) \quad (4.45)$$

$$\begin{aligned}(2\epsilon)^{-d/2}\pi^{-d/2}\int_{\mathbb{R}^d} h(\frac{||u||^2}{2\epsilon})\sum_{i=1}^{\lceil s\rceil-1} P_{x,i}(u)du &= \int_{\mathbb{R}^d}\pi^{-d/2}h(||u||^2)\sum_{i=1}^{\lceil s\rceil-1}\epsilon^{i/2}P_{x,i}(u)du \\ &= \sum_{i=1}^{\lceil s\rceil-1} F_i^1\epsilon^i\end{aligned} \quad (4.46)$$

where F_i^1 is satisfied with (1)(2)(3) in our lemma 4.3.1, as the integral of odd order polynomials goes zero.

For each term in the third term of a_1, using lemma 4.6.8 and 4.6.10, the derivative of $h = \exp(-x)$, and cancellation of the integrals of odd order polynomials, and carefully tracking the order, we have

$$\begin{aligned}&(2\epsilon)^{-d/2}\pi^{-d/2}C_i\int_{\mathbb{R}^d}\frac{(\sum_{j=4}^{\lceil s\rceil+1}Q_{x,j}(u)+O(||u||^{\lceil s\rceil+2}))^i}{\epsilon^i}(\tilde{f}(0) \\ &+\sum_{k=1}^{\lceil s\rceil-1}P_{x,k}(u))h^{(i)}(\frac{||u||^2}{2\epsilon})du \\ &= C_i\int_{\mathbb{R}^d}\frac{(\sum_{j=4}^{\lceil s\rceil+1}Q_{x,j}(u)*\epsilon^{j/2}+O(||u||^{\lceil s\rceil+2}\epsilon^{\lceil s/2\rceil+1}))^i}{\epsilon^i}(\tilde{f}(0) \\ &+\sum_{k=1}^{\lceil s\rceil-1}P_{x,k}(u)\epsilon^{k/2})h^{(i)}(||u||^2)du \\ &= \sum_{k=1}^{\lceil s\rceil-1}F_k^{2,i}\epsilon^k + O(\epsilon^{\lceil s/2\rceil})\end{aligned} \quad (4.47)$$

where $F_k^{2,i}$ and the residual term are satisfied with (1)(2)(3) in the lemma 4.3.1.

Combining above three computations, with lemma 4.6.8, we finally get

$$(2\epsilon)^{-d/2}\pi^{-d/2}\int_{\mathcal{M}} h(\frac{||x-y||^2}{2\epsilon})f(y)dV(y) = f(x) + \sum_{i=1}^{\lceil s/2\rceil-1} f_i(x)\epsilon^i + O(\epsilon^{s/2}) \quad (4.48)$$

where f_i and the residual term $O(\epsilon^{s/2})$ is satisfied with (1)(2)(3) in the lemma.

For the general case when $s \notin \mathbb{N}$, the proof is exactly the same as the above proof, except that we use Taylor expansion for Hölder function rather than the usual Taylor expansion to get the equation 4.38. And this ends the proof. □

Proof for proposition 4.3.2. By lemma 4.3.1 with $s > 0$, we have $G_\epsilon f(x) = f(x) + \sum_{i=1}^{\lceil s/2\rceil-1} f_i(x)\epsilon^i + O(\epsilon^{s/2})$. To cancel the first order term of $G_\epsilon F(x) - f(x)$, we can let $F_1 = -f_1$. Then we know F_1 is polynomials of $\{\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}\}, \{\frac{\partial^{|\boldsymbol{j}|}\rho^{-1}}{\partial u^{\boldsymbol{j}}}\}, \{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}\}$ where $|\boldsymbol{j}| \leq 2$, and $F_1 \in C^{s-2}(\mathcal{M})$. Again we can use lemma 4.3.1 to get $\epsilon G_\epsilon F_1 = \epsilon F_1 + \epsilon\sum_{i=1}^{\lceil (s-2)/2\rceil-1} F_{1,i}\epsilon^i + O(\epsilon^{s/2})$, and we can let $F_2 = -F_1 - F_{1,1}$ to cancel the ϵ^2 term. In general, for $i \leq \lceil s/2\rceil - 1$, given $F_0 = f, F_1, ..., F_{i-1}$, we can sequentially define

$$F_i = -(\sum_{j=0}^{i-1} F_{j,i-j}) \quad (4.49)$$

and $F_i \in C^{s-2i}(\mathcal{M})$, which guarantees the induction can continue and we have the expansion

$$\epsilon^i G_\epsilon F_i = \epsilon^i F_i + \epsilon^i \sum_{j=1}^{\lceil s/2\rceil-1-i} F_{j,i}\epsilon^j + O(\epsilon^{s/2}). \quad (4.50)$$

Finally, we have

$$\sup_{x\in\mathcal{M}} |G_\epsilon F(x) - f(x)| \leq C_1\epsilon^{s/2} \quad (4.51)$$

by computation and cancellation of all lower order terms, where C_1 only depends on f and $\mathcal{M}$ by lemma 4.3.1.

Now we turn to prove $||G_\epsilon F(x)||^2_{\tilde{\mathbb{H}}_\epsilon} \leq C_2\epsilon^{-d/2}$. Similar to lemma 5.2 of [YD16], as $\langle h_\epsilon(x,\cdot), h_\epsilon(y,\cdot)\rangle_{\tilde{\mathbb{H}}_\epsilon} = h_\epsilon(x,y)$, we have

$$\begin{aligned} ||G_\epsilon(F)||^2_{\tilde{\mathbb{H}}_\epsilon} &= \frac{1}{(2\pi\epsilon)^d}\int_{\mathcal{M}}\int_{\mathcal{M}} h_\epsilon(x,y)F(x)F(y)dV(x)dV(y) \\ &\leq ||F||^2_\infty \frac{1}{(2\pi\epsilon)^{d/2}}\int_{\mathcal{M}} dV(x)\frac{1}{(2\pi\epsilon)^{d/2}}\int_{\mathcal{M}} h_\epsilon(x,y)dV(y) \end{aligned} \tag{4.52}$$

Applying Lemma 4.3.1 to $f = 1$, we get

$$|\frac{1}{(2\pi\epsilon)^{d/2}}\int_{\mathcal{M}} h_\epsilon(x,y) * 1 dV(y) - 1| \leq C\epsilon \leq C \tag{4.53}$$

Therefore,

$$\begin{aligned} ||G_\epsilon(F)||^2_{\tilde{\mathbb{H}}_\epsilon} &\leq (1+C)||F||^2_\infty \epsilon^{-d/2}\, Vol(\mathcal{M}) \\ &\leq (1+C)(||f||_\infty + ||\Delta_{\mathcal{M}} f||)\epsilon^{-d/2}\, Vol(\mathcal{M}) \\ &\leq C_2\epsilon^{-d/2} \end{aligned} \tag{4.54}$$

And this finishes the proof. □

Proof of proposition 4.3.4. According to the condition (A3) and prior (P'), it is sufficient for us to prove the following stronger results: the two events

1. $\forall t \in [c_1 n^{\frac{-2}{\varrho}} \log^{a_1}(n), c_2 \log^{-1}(n)]$, $C_1 t^{\varrho/2} \leq \hat{v}(t) \leq C_2 t^{\varrho/2}$ for some positive constants a_1, c_1, c_2, C_1, C_2

2. $C_3 n^{\frac{-2}{\varrho}} \log^{a_2}(n) \leq T_n^2 \leq C_4 n^{\frac{-2}{\varrho}} \log^{a_2}(n)$ with some constants $C_3, C_4 > 0$ and $a_2 > a_1 > 0$

hold together with a high probability. To do this, we give two technical lemmas 4.6.11, 4.6.12 characterize the orders and concentration of $\hat{R}_n(t)$ and T_n with high probability in the given regions. And this finishes our proof. □

4.6 Appendix for proofs

4.6.1 Extension of lemma 4.3.1 for general kernel

In this section, we extend 4.3.1 for a general kernel with some regularity assumption. To extend lemma 4.3.1 for a general kernel $k_0 : [0,\infty) \to [0,\infty)$, we require the decaying and moment conditions:

(C1) Smoothness: $k_0 \in C^{\lceil s/2 \rceil}([0,\infty))$.

(C2) Decaying condition: $\exists a, a_l > 0, s.t., |k_0^{(l)}(\xi)| \leq a_l \exp(-a\xi)$ for all $\xi > 0$ and $l = 0, ..., \lceil s/2 \rceil$.

(C3) Normalization: $m_0[k_0] = \int k_0(||u||^2)du = 1$.

We have the following variant of lemma 4.3.1 for general kernel k_0:

Lemma 4.6.1 (Extension of lemma 4.3.1)**.** *Under Assumption (A1) and (C1)-(C3), there is $\epsilon_0 > 0$ a constant depending on $\mathcal{M}$, such that $0 < \epsilon < \epsilon_0$, for any $f \in C^s(\mathcal{M}) = C^{\lfloor s \rfloor, \beta}(\mathcal{M})$ where $s > 0$,$s \notin \mathbb{Z}^+$, and $\beta = s - m$,*

$$\epsilon^{-d/2} \int_{\mathcal{M}} k_0(\frac{||x-y||^2}{\epsilon}) f(y) dV(y) = f(x) + \sum_{i=1}^{\lceil s/2 \rceil - 1} f_i(x)\epsilon^i + O(\epsilon^{s/2})$$

where f_i is only polynomially determined by

(i) Functions which are the derivative of f: $\{\frac{\partial^{|\boldsymbol{j}|}\tilde{f}}{\partial s^{\boldsymbol{j}}}(0)\}$ where $|\boldsymbol{j}| \leq 2i$,

(ii) Functions which are the local derivative of the extrinsic manifold coordinate at x, $\{\frac{\partial^{|\boldsymbol{j}|}\rho^{-1}}{\partial u^{\boldsymbol{j}}}(0)\}, \{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}(0)\}$ where $|\boldsymbol{j}| \leq 2i$, and

(iii) Numbers which are finite order moments of the kernel and its derivative: $m_{\boldsymbol{j}}[k_0^{(l)}] = \int_{\mathbb{R}^d} u^{\boldsymbol{j}} k_0^{(l)}(||u||^2)du, |\boldsymbol{j}| < \infty, 0 \leq l \leq \lceil s/2 \rceil$

For the $O(\epsilon^{s/2})$ term, denoted as $r_{f_\epsilon}(x)$, it satisfies that $\sup_{x\in\mathcal{M}}|r_{f_\epsilon}(x)| \leq C(f)\epsilon^{s/2}$, where $C(f) = c(\mathcal{M})(1+\sum_{l=0}^{\lfloor s\rfloor}||D^l f||_\infty + ||D^{\lfloor s\rfloor}f||_{C^{0,\beta}(\mathcal{M})})$. Here $c(\mathcal{M})$ is some constant, determined by the manifold $\mathcal{M}$, only depends on the $\{||D^l\rho^{-1}||_\infty\}, \{||D^l g||_\infty\}$ where $l \leq \lfloor s\rfloor$.

Proof. Lemma 4.6.1 can be proved in the exact same way as Lemma 4.3.1, and the only difference is that the derivative of the kernel $k_0^{(l)}$ shall be kept in the expansion. □

4.6.2 Technical lemmas for section 4.2

In this section we review the lemmas on RKHSs we need in section 3. The proofs are very similar to their analogue in [vdVvZ08] and [YD16].

Lemma 4.6.2 (Analogue of lemma 5.1 in [YD16])**.** *Consider the squared exponential covariance kernel h_ϵ and its associated RKHSs $\mathbb{H}_\epsilon$ on $[0,1]^D$ and $\tilde{\mathbb{H}}_\epsilon$ on $\mathcal{X}$. For any $f \in \tilde{\mathbb{H}}_\epsilon$, there exists $g \in \mathbb{H}_\epsilon$ such that $g|_{\mathcal{X}} = f$ and $||g||_{\mathbb{H}_\epsilon} = ||f||_{\tilde{\mathbb{H}}_\epsilon}$. Moreover, for any other $g' \in \mathbb{H}_\epsilon$ with $g'|_{\mathcal{X}} = f$, it holds that $||g'||_{\mathbb{H}_\epsilon} \geq ||f||_{\tilde{\mathbb{H}}_\epsilon}$. This implies $||f||_{\tilde{\mathbb{H}}_\epsilon} = \int_{g\in\mathbb{H}_\epsilon, g|_{\mathcal{X}}=f} ||g||_{\mathbb{H}_\epsilon}$*

Lemma 4.6.3 (Analogue of lemma 8.1 in [YD16] and lemma 4.7 in [VdVVZ09])**.** *For the squared exponential covariance kernel h_ϵ and its associated RKHS, $\tilde{\mathbb{H}}_\epsilon$, on $\mathcal{X}$, we denote the unit ball in $\tilde{\mathbb{H}}_\epsilon$ as $\tilde{\mathbb{H}}_\epsilon^1$. If $\epsilon_1 \geq \epsilon_2$, we have the relation*

$$\epsilon_1^{-\frac{1}{4}}\tilde{\mathbb{H}}_{\epsilon_1}^1 \subset \epsilon_2^{-\frac{1}{4}}\tilde{\mathbb{H}}_{\epsilon_2}^1$$

Lemma 4.6.4 (Analogue of lemma 8.2 in [YD16] and lemma 4.8 in [VdVVZ09])**.** *Assume $q(x) \in \tilde{\mathbb{H}}_\epsilon^1$ satisfies $|q(x)| \leq 1$ and $|q(x) - q(x')| \leq \epsilon^{-\frac{1}{2}}||x - x'||\tau$ for any $x, x' \in \mathcal{M}$, where $\tau^2 = \int ||\lambda||^2 d\mu(\lambda)$.*

Lemma 4.6.5 (Covering bounds on RKHS via covering on $\mathcal{X}$, corollary A.8 in [HS21]). *Assume $\mathcal{X} \subset [0,1]^D$ satisfied with condition (A1).For any $0 < \epsilon < \epsilon_0$ and $\varepsilon' > 0$, we have*

$$\log(N(\varepsilon', \tilde{\mathbb{H}}^1_\epsilon, ||\cdot||_\infty)) \leq N(\sqrt{\epsilon}, \mathcal{X}, ||\cdot||_\infty) \cdot \log((N(\varepsilon', \mathbb{H}^1_1, ||\cdot||_\infty)) \leq K\epsilon^{-\rho/2}(\log(\frac{1}{\varepsilon'}))^{D+1}$$

A better covering bounds can be gotten when $\mathcal{X}$ is a manifold.

Lemma 4.6.6 (Lemma 5.4 in [YD16], covering bounds of RKHS on Manifold). *There exists a constant K depending only on d, D and $\mathcal{M}$, such that for any $\varepsilon' < 1/2$ and $\epsilon < min(\epsilon_1 = \frac{4\delta_0^2 d}{\delta^2}, 1)$ where δ_0 is defined in lemma 7.7 of [YD16] and δ in eq 4.3, it holds that*

$$\log(N(\varepsilon', \tilde{\mathbb{H}}^1_t, ||\cdot||_\infty)) \leq Kt^{-d/2}(\log(\frac{1}{\varepsilon'}))^{d+1}$$

Similar to the proof of lemma 4.6 in [vdVvZ08], the covering bounds on RKHS can imply an upper bound on $-\log(||W^t||_\infty \leq \varepsilon')$, as shown in the next lemma:

Lemma 4.6.7 (small ball probability of Gaussian measure). *Consider the kernel h_t and its associated Gaussian process W^ton $\mathcal{X}$ satisfying with condition (A1). Then for some $\epsilon_0 > 0$, there exists constants C and ε_0' that only depend on ϵ_0, μ, ρ, D and $\mathcal{X}$, such that for all $t \leq \min(\epsilon_0, 1)$ and $\varepsilon' < \varepsilon_0'$,*

$$-\log(\sup_{x \in \mathcal{M}} ||W^t||_\infty \leq \varepsilon') \leq Ct^{-d/2}(\log(\frac{1}{\sqrt{t}\varepsilon'}))^{D+1}$$

. When $\mathcal{X}$ is a $d-$dimension manifold, the logarithm term on the right-hand side can be further improved to $(\log(\frac{1}{\sqrt{t}\varepsilon'}))^{d+1}$ with ϵ_1 substituting ϵ_0

4.6.3 Technical lemmas for section 4.3

Lemma 4.6.8 (Closeness of a class of generic homogeneous polynomials). *Consider the polynomials $A = \{P(u) | P(u) = \sum_{i=0}^k P_{x,i}(u), \forall k \in \mathbb{N}\}$ where the coefficients of*

each term in the generic homogeneous polynomials $P_{x,i}(u)$ are polynomials of derivatives of another function $\{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}\}(0)$ up to i order ($|\boldsymbol{j}| \leq i$). Then, A is closed under product and summation, which means, A is a ring.

Proof of lemma 4.6.8. Let $R(u) = \sum_{i=0}^{s} P_{x,i}^{R}(u)$ and $S(u) = \sum_{i=0}^{r} P_{x,i}^{R}(u)$ belong to A. It is enough to check $R + S \in A$ and $RS \in A$. It is enough to keep track of the coefficients of each generic homogeneous polynomial in $R + S$ and RS. We notice $P_{x,i}^{R+S}(u) = P_{x,i}^{R}(u) + P_{x,i}^{S}(u)$ and $P_{x,i}^{RS}(u) = \sum_{j=0}^{i} P_{x,j}^{R}(u) * P_{x,i-j}^{S}(u)$. By keeping track of the order and the ring property of polynomials, the coefficients of both of generic homogeneous polynomials are polynomials of $\{\frac{\partial^{\boldsymbol{j}}g}{\partial u^{\boldsymbol{j}}}(0)\}, |\boldsymbol{j}| \leq i$. Therefore, we finish the proof with $R + S \in A$ and $RS \in A$. □

Lemma 4.6.9 (Extension of lemma 6 in [CL06])**.** *If $y \in \mathcal{M}$ is in a Euclidean ball of radius $\epsilon^{1/2}$ around x, then, for ϵ sufficiently small, then*

$$s_i = u_i + \sum_{j=3}^{\lceil s \rceil - 1} P_{x,j}^{i}(u) + O(||u||^{\lceil s \rceil}) \tag{4.55}$$

where $P_{x,j}^{i}(u)$ are generic homogeneous polynomials of degree i coefficients of which are polynomials of local coordinate derivative $\{\frac{\partial^{|\boldsymbol{k}|}\rho^{-1}}{\partial u^{\boldsymbol{k}}}(0)\}, \{\frac{\partial^{|\boldsymbol{k}|}g}{\partial u^{\boldsymbol{k}}}(0)\}$ where $|\boldsymbol{k}| \leq j$. The residual terms $O(||u||^{\lceil s \rceil})$ just (polynomially) depend on the $\{\frac{\partial^{|\boldsymbol{k}|}\rho^{-1}}{\partial u^{\boldsymbol{k}}}\}, \{\frac{\partial^{|\boldsymbol{k}|}g}{\partial u^{\boldsymbol{k}}}\}$ where $|\boldsymbol{k}| \leq \lceil s \rceil$.

Proof of lemma 4.6.9. Directly do the expansion and we get

$$s_i = \rho_i^{-1}(y) - \rho_i^{-1}(x) = \rho_i^{-1}((u, g(u))) = \sum_{|\boldsymbol{j}|=1}^{\lceil s \rceil - 1} \frac{\partial^{|\boldsymbol{j}|}(\rho_i^{-1}(u, g(u)))}{\partial u^{\boldsymbol{j}}}(0) u^{\boldsymbol{j}} + O(||u||^{\lceil s \rceil})$$

$$= \sum_{j=1}^{\lceil s \rceil - 1} P_{x,j}^{i}(u) + O(||u||^{\lceil s \rceil})$$

Now we only need to prove $P^i_{x,1}(u) = u_i$ and $P^i_{x,2}(u) = 0$. We claim that $u_i = s_i + O((\sum_{i=1}^d s_i^2)^{3/2})$. If the claim is true, $P^i_{x,1}(u) = u_i$ and $P^i_{x,2}(u) = 0$ follow by substitute $u_1, ..., u_d$ to the expansion and compare the terms with the same order.

Next we prove the claim. Consider the geodesic γ connecting x and y parameterized by arclength with $\gamma(0) = x$ and $\gamma(\tau) = y$ where $\tau = \sqrt{\sum_{i=1}^d s_i^2}$. By Taylor expansion

$$y = \gamma(\tau) = \gamma(0) + \tau\gamma^{'}(0) + \frac{\tau^2}{2}\gamma^{''}(0) + O(\tau^3).$$

Notice that $u_i = \langle \gamma(\tau) - \gamma(0), e_i \rangle$, $\tau\gamma'(0) = (s_1, ...s_d)$, and $\gamma^{''}(0)$ is orthogonal to the tangent plane at x. Then, we have $u_i = \langle \tau\gamma^{'}(0) + \frac{\tau^2}{2}\gamma^{''}(0) + O(\tau^3), e_i \rangle = s_i + O(\tau^3)$. So our claim is true.

And we finish the proof. □

Lemma 4.6.10 (Extension of lemma 7 in [CL06]). *If $y \in \mathcal{M}$ is in a Euclidean ball of radius $\epsilon^{1/2}$ around x, then, for ϵ sufficiently small, we have*

$$||y - x||^2 = ||u||^2 + Q_{x,4}(u) + \sum_{i=5}^{\lceil s \rceil - 1} Q_{x,i}(u) + O(||u||^{\lceil s \rceil}) \quad \textit{(metric comparison)} \tag{4.56}$$

$$det(\frac{dy}{du}) = 1 + V_{x,2}(u) + \sum_{i=3}^{\lceil s \rceil - 1} V_{x,i}(u) + O(||u||^{\lceil s \rceil}) \qquad \textit{(volume comparison)} \tag{4.57}$$

where $Q_{x,i}(u)$ are generic homogeneous polynomials of degree i, and coefficients of all items in it are polynomials of $\{\frac{\partial^{|\boldsymbol{j}|} g(0)}{\partial u^{\boldsymbol{j}}}\}$, $|\boldsymbol{j}| \leq i - 2$ and $V_{x,i}(u)$ are generic homogeneous polynomials of degree i, and coefficients of all items in it are polynomials of $\{\frac{\partial^{|\boldsymbol{j}|} g(0)}{\partial u^{\boldsymbol{j}}}\}, |\boldsymbol{j}| \leq i$. Finally, both of the residual terms $O(||u||^{2s})$ just (polynomially) depend on the $\{\frac{\partial^{|\boldsymbol{j}|} g}{\partial u^{\boldsymbol{j}}}\}, |\boldsymbol{j}| \leq \lceil s \rceil$.

Proof of lemma 4.6.10. First, we have

$$
\begin{aligned}
||y-x||^2 &= ||u||^2 + \sum_{i=d+1}^{D} g_i^2(u) \\
&= ||u||^2 + \sum_{i=d+1}^{D} (g_i(0) + \sum_{|\boldsymbol{j}|=1}^{\lceil s \rceil - 1} \frac{\partial^{|\boldsymbol{j}|} g(0)}{\partial u^{\boldsymbol{j}}} u^{\boldsymbol{j}} + O(||u||^{\lceil s \rceil}))^2 \\
&(g_i(0) = 0, \frac{\partial g(0)}{\partial u_i} = 0) \\
&= ||u||^2 + \sum_{i=d+1}^{D} (\sum_{|\boldsymbol{j}|=2}^{\lceil s \rceil - 1} \frac{\partial^{|\boldsymbol{j}|} g(0)}{\partial u^{\boldsymbol{j}}} u^{\boldsymbol{j}} + O(||u||^{\lceil s \rceil}))^2 \\
&= ||u||^2 + \sum_{i=4}^{\lceil s \rceil - 1} Q_{x,i}(u) + O(||u||^{\lceil s \rceil})
\end{aligned}
$$

where coefficients of $Q_{x,i}(u)$ are polynomials of $\{\frac{\partial^{|\boldsymbol{j}|} g(0)}{\partial u^{\boldsymbol{j}}}\}, |\boldsymbol{j}| \leq i-2$ and the residual term $O(||u||^{\lceil s \rceil})$ just depend on the $\{\frac{\partial^{|\boldsymbol{j}|} g}{\partial u^{\boldsymbol{j}}}\}, |\boldsymbol{j}| \leq \lceil s \rceil$, by tracking the dependency on the $\{\frac{\partial^{|\boldsymbol{j}|} g}{\partial u^{\boldsymbol{j}}}\}$. On the other hand, we know $J = [\frac{dy}{du}] = [I_d, \frac{\partial g}{\partial u}]^T$.

So we have

$$|\det(\frac{dy}{du})|^2 = \det(J^T J) = \det(I_d + (\frac{\partial g}{\partial u})^T \frac{\partial g}{\partial u})$$

$$= \det([(\delta_{pq} + \sum_{k=1}^{D-d} \frac{\partial g_k}{\partial u_p}\frac{\partial g_k}{\partial u_q}]_{pq})$$

$$= \det([\langle v_p.v_q\rangle + \langle \frac{\partial g}{\partial u_p}, \frac{\partial g}{\partial u_q}\rangle]_{pq}) \quad (\textit{determinant of Gram Matirx}) \det([\langle \frac{\partial y}{\partial u_p}, \frac{\partial y}{\partial u_q})\rangle]_{pq})$$

$$= \det([\delta_{pq} + \langle \sum_{|\boldsymbol{j}|=0}^{\lceil s\rceil-1} \frac{\partial^{|\boldsymbol{j}|}\partial g}{\partial u_p u^{\boldsymbol{j}}}(0)u^{\boldsymbol{j}} + O(||u||^{\lceil s\rceil}), \sum_{|\boldsymbol{j}|=0}^{\lceil s\rceil-1} \frac{\partial^{|\boldsymbol{j}|}\partial g}{\partial u_q u^{\boldsymbol{j}}}(0)u^{\boldsymbol{j}} + O(||u||^{\lceil s\rceil})\rangle]_{pq})$$

$$(\{\frac{\partial g}{\partial u_i} = 0\})$$

$$= \det([\delta_{pq} + \langle \sum_{|\boldsymbol{j}|=1}^{\lceil s\rceil-1} \frac{\partial^{|\boldsymbol{j}|}\partial g}{\partial u_p u^{\boldsymbol{j}}}(0)u^{\boldsymbol{j}} + O(||u||^{\lceil s\rceil}), \sum_{|\boldsymbol{j}|=1}^{\lceil s\rceil-1} \frac{\partial^{|\boldsymbol{j}|}\partial g}{\partial u_q u^{\boldsymbol{j}}}(0)u^{\boldsymbol{j}} + O(||u||^{\lceil s\rceil})\rangle]_{pq})$$

$$= \det([\delta_{pq} + \sum_{i=2}^{\lceil s\rceil-1} P_{x,i}^{p,q}(u) + O(||u||^{\lceil s\rceil}) \quad \textit{defenition of determinant}$$

$$= \sum_{\sigma\in S_d} sgn(\sigma)\Pi_{p=1}^{d}(\delta_{p\sigma(p)} + \sum_{i=2}^{\lceil s\rceil-1} P_{x,i}^{p,\sigma(p)}(u) + O(||u||^{\lceil s\rceil}))$$

where S_d is d-order permutation group, and $P_{x,i}^{p,\sigma(p)}(u)$ is generic homogeneous polynomials of degree i. The coefficient of each term in $P_{x,i}^{p,\sigma(p)}(u)$ is polynomials of $\{\frac{\partial^{|\boldsymbol{j}|}g(0)}{\partial u^{\boldsymbol{j}}}\}, |\boldsymbol{j}| \le i$, and the residual term $O(||u||^{\lceil s\rceil})$ just depend on the $\{\frac{\partial^{|\boldsymbol{j}|}g}{\partial u^{\boldsymbol{j}}}\}, |\boldsymbol{j}| \le \lceil s\rceil$ by tracking dependency carefully. Then, by direct expansion and lemma 4.6.8, we get

$$|\det(\frac{dy}{du})|^2 = 1 + P_{x,2}(u) + \sum_{i=3}^{\lceil s\rceil-1} P_{x,i}(u) + O(||u||^{\lceil s\rceil})$$

where coefficient dependency of each term in $P_{x,i}(u)$ and residual are the same as before. Finally, by expansion of $\sqrt{1+x} = 1 + \sum_{i=1}^{k} C_{1/2}^{i} x^i + O(x^{k+1})$ and lemma

4.6.8, we get

$$|\det(\frac{dy}{du})| = \sqrt{1 + \sum_{i=2}^{\lceil s\rceil -1} P_{x,i}(u) + O(||u||^{\lceil s\rceil})}$$

$$= 1 + \sum_{j=1}^{\lfloor \lceil s\rceil /2\rfloor} C^i_{1/2}(\sum_{i=2}^{\lceil s\rceil -1} P_{x,i}(u) + O(||u||^{\lceil s\rceil}))^j + O((\sum_{i=2}^{\lceil s\rceil -1} P_{x,i}(u) + O(||u||^{\lceil s\rceil}))^{\lfloor \lceil s\rceil /2\rfloor +1})$$

$$= 1 + V_{x,2}(u) + \sum_{i=3}^{\lceil s\rceil -1} V_{x,i}(u) + O(||u||^{\lceil s\rceil})$$

where $V_{x,i}(u)$ are generic homogeneous polynomials of degree i,and coefficient of each term in $V_{x,i}(u)$ is polynomials of $\{\frac{\partial^{|j|} g(0)}{\partial u^j}\}, |\boldsymbol{j}| \leq i$. The coefficient dependency of the residual term $O(||u||^{\lceil s\rceil})$ are the same as before. So we finish the proof. □

Lemma 4.6.11 (Uniform Concentration of $\hat{R}_n(t)$)**.** *With high probability,* $\hat{R}_n(t) = O(t^{\frac{-2}{d}})$ *for all* $t \in [c_1 n^{\frac{-2}{d}} \log^{3/d}(n), c_2 \log^{-1}(n)]$.

Proof. The proof is essentially the same to the proofs of proposition 2.2 and theorem 2.3 in [CW22], where they give the concentration of the kernel count statistics. Following the same way, we get the concentration for the $\hat{R}_n(t)$ for each t, and the uniform bounds comes from the union bounds and continuity argument. □

Lemma 4.6.12 (Concentration of T_n)**.** *With high probability,* $T_n = O(n^{\frac{-2}{d}} \log^{4/d}(n))$.

Proof. Similarly, the proof is essentially the same to the proofs of proposition 2.2 and theorem 2.3 in [CW22], . □

4.7 Appendix for numerical results

4.7.1 Algorithms

The marginal log-likelihood of t is

$$\log(L(Y|X,t)) = \log(P(Y|X)) = \log(\int p(f|X,t)p(Y|X,f,t)df)$$
$$= -\frac{1}{2}Y^T(h_t(X,X)+\sigma^2 I)^{-1}Y - \frac{1}{2}\log(|h_t(X,X)+\sigma^2 I|) - \frac{n}{2}\log(2\pi) \tag{4.58}$$

And then the marginal posterior of t is

$$p(t|X,Y) \propto L(Y|X,t)p(t) \tag{4.59}$$

As in MCMC the constant in prior and likelihood does not matter, we give the prior up to a constant in algorithm 2. Given t, we have the posterior mean on testing data X^{te} as

$$E(f(X^{te})|Y,X,t) = Y^T(h_t(X,X)+\sigma^2 I)^{-1}h_t(X^{te},X) \tag{4.60}$$

We can use the above equations to get the posterior sample of $f(X^{te})$. We finally give the numerical implementation in algorithm 1.

Algorithm 8: KNN algorithm for (log-)prior construction

Input: hyperparameters a_0, b_0, s_0, data $X = \{X_i\}_{i=1}^n$,

1: Set $k = \log(n)^{1.1}$. Pre-compute $T_n = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} T_n^i$ where $T_n^i = \inf_{r>0}\{r > 0, s.t \sum_{j=1,j\neq i}^N 1_{|X_i - X_j|} \geq k\}$, and $\mathcal{S}$ a random subset of $\{1,...,n\}$ with size s_0 with

2: For given value $\{t\}$, compute $\hat{v}_n(t) = \frac{1}{n}\sum_{i=1}^n \hat{R}_i(t)$ for $\hat{R}_i(t) = \sum_{j\neq i} h_t(X_i, X_j)$.

3: set $\log(p(t)) = -a_0\log(t) - b_0\frac{1}{\hat{v}_n(t)}$ if $t > T_n^2$, else $\log(p(t)) = -\infty$

4.8 Conclusion

This essay presents improved results regarding the rates of posterior contraction in Gaussian process (GP) regression, taking into account the existence of low-dimensional structure in the data as described by the covering number. We establish the conditions under which the posterior contraction rates can adapt to any intrinsic dimension

Algorithm 9: M-H Algorithm for Bayesian Estimator for f on $\{X_j^{te}\}_{j=1}^{n_{te}}$

Input: hyperparameters $\sigma^2, a_0, b_0,$, data $X = \{X_i\}_{i=1}^n, Y = \{Y_i\}_{i=1}^n, X^{te} = \{X_j^{te}\}_{j=1}^{n_{te}}$, number of iterations B.

Output: $\{f^{[b]}(X_j^{te})\}_{j=1}^{n_{te}}$

1: Set initial parameters $\Theta_{[0]} = \{t^{[0]}\}$.
2: **for** $b = 1, \cdots, B$ **do**
3: Sample $t^{[b]}|\{X_i, Y_i\}_{i=1}^n$ from the posterior via Metropolis Hasting from marginal posterior equation 4.59 with prior in algorithm 1. (Step 1)
4: Sample the posterior mean of $\{f^{[b]}(X_j^{te})\}_{j=1}^{n_{te}}|\{X_i, Y_i\}_{i=1}^n, t^{[b]}$ from the conditional posterior mean via equation 4.60 . (Step 2)
5: **end for**
6: Take the average and get the $\{\hat{f}(X_j^{te})\}_{j=1}^{n_{te}}\} = \{\frac{1}{B}\sum_{b=1}^B f^{[b]}(X_j^{te})\}_{j=1}^{n_{te}}$

ϱ, based on the covering number of the domain $\mathcal{X}$. We prove the optimal posterior contraction rate of $O(n^{-s/(2s+\varrho)})$, up to a logarithmic constant, assuming an approximation order of the reproducing kernel Hilbert space (RKHS) on $\mathcal{X}$. When $\mathcal{X}$ is a ϱ-dimensional compact smooth manifold, we prove the RKHS approximation for s-order Hölder functions on the manifold for any positive s, using a novel analysis of kernel approximation on the manifold. This leads to achieving the optimal adaptive posterior contraction rate. To eliminate the need for prior knowledge of the intrinsic dimension, we introduce a novel empirical Bayes prior of bandwidth, which utilizes kernel count statistics and the K-nearest neighbor (KNN) algorithm. We validate our theoretical findings through various numerical experiments.

Despite these significant advancements in theory and algorithm, there remain several promising areas for future research. One trajectory includes extending the approximation theorem for RKHS in general spaces. The process of naturally extending the continuous function beyond manifold to a general set remains unclear, and the relationship between the convergence rate and the intrinsic property of the underlying function is worth exploring. Further, there's room to extend our findings to more flexible kernels, such as the Mat'ern kernel, and to evaluate the theoretical

and practical behavior of these kernels, along with the resulting smoothness and covariance behavior for various structures. Finally, studying the posterior contraction rate of other advanced Gaussian processes on restricted domains and comparing the practical and theoretical properties with the present method, poses an interesting avenue for further exploration.

Chapter 5

Conclusion

This dissertation studies Bayesian inference, specifically in the context of structurally diverse data, by developing cutting-edge methodologies and robust theoretical frameworks. The research work is well-structured into three distinctive sections, each addressing unique data aspects: Dynamical systems, continuous-time Markov Chains, and data with low-dimensional structures.

In the first chapter of the dissertation, we propose a new hierarchical shrinkage GP (HierGP), which incorporates such structure via cumulative shrinkage priors within a GP framework. We show that the HierGP implicitly embeds the well-known principles of effect sparsity, heredity and hierarchy for analysis of experiments [HW92], which allows our model to identify structured sparse features from the response surface with limited data. We propose efficient posterior sampling algorithms for model training and prediction and prove desirable consistency properties for the HierGP. Finally, we demonstrate the improved performance of HierGP over existing models, in a suite of numerical experiments and an application to dynamical system recovery.

In the second chapter of the dissertation, we present a new Bayesian method, *Bayesian Approximation Spectral Inference (BASI)*, for estimating the low-rank representation of the generator of a Continuous-Time Markov Chain (CTMC) from discretely observed data. This development is motivated by the need for practical inference methods from continuous-time Markov chains in fields ranging from biology and physics to finance. Our method relies on a novel probabilistic matrix factorization approach in approximating the likelihood under a low-rank spectral representation. This avoids the computational challenges of evaluating the implicit likelihood

function of generators in previous approaches. We provide theoretical justification for BASI by showing asymptotic properties of the posterior distribution including a Bernstein-von Mises (BvM) type theorem. For numerical implementations, we describe a Markov chain Monte Carlo (MCMC) algorithm relying on Gibbs updating. Our methods are demonstrated through numerical experiments and applications.

In the final part of the dissertation, we present enhanced results on the posterior contraction rates for Gaussian process (GP) regression, considering the presence of low-dimensional structure in the data described by the covering number. We establish the conditions under which the posterior contraction rates become adaptive to any intrinsic dimension ϱ in terms of the covering number of the domain $\mathcal{X}$, and prove the optimal posterior contraction rate $O(n^{-s/(2s+\varrho)})$, up to a logarithmic constant, assuming an approximation order of the reproducing kernel Hilbert space (RKHS) on $\mathcal{X}$. When $\mathcal{X}$ is a ϱ-dimensional compact smooth manifold, we prove the RKHS approximation to s-order Hölder functions on manifold for any positive s by a novel analysis of kernel approximation on manifold, leading to the optimal adaptive posterior contraction rate. We also employ a new empirical Bayes prior to bandwidth using the kernel count statistics and K-nearest neighbor (KNN) algorithm, which eliminates the need for prior knowledge of the intrinsic dimension. We validate our theoretical findings through various numerical experiments.

Overall, these three essays contribute to Bayesian statistics for data with diverse and complex structures. The dissertation integrates theoretical contributions with practical numerical experiments, providing empirical support to the innovative propositions.

Bibliography

[ADL13] Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013.

[AE03] S Assoudou and B Essebbar. A bayesian model for markov chains via jeffrey's prior. *Communications in Statistics - Theory and Methods*, 2003.

[Ale15] Alen Alexanderian. A brief note on the Karhunen-Loeve expansion. *arXiv preprint arXiv:1509.07526*, 2015.

[ÅSWH20] Mattias Åkesson, Prashant Singh, Fredrik Wrede, and Andreas Hellander. Convolutional neural networks as summary statistics for approximate bayesian computation. *arXiv preprint arXiv:2001.11760*, 2020.

[Aza14] Romain Azais. A recursive nonparametric estimator for the transition kernel of a piecewise-deterministic markov process. *ESAIM: Probability and Statistics*, 18:726–749, 2014.

[BAVX21] Fan Bu, Allison E Aiello, Alexander Volfovsky, and Jason Xu. Likelihood-based inference for partially observed stochastic epidemics with individual heterogeneity. *arXiv preprint arXiv:2112.07892*, 2021.

[BCC14] Charles Bordenave, Pietro Caputo, and Djalil Chafai. Spectrum of markov generators on sparse random graphs. *Communications on Pure and Applied Mathematics*, 67(4):621–669, 2014.

[BCG21] Sayantan Banerjee, Ismaël Castillo, and Subhashis Ghosal. Bayesian inference in high-dimensional models. *arXiv:2101.04491*, 2021.

[BCSK21] Gert-Jan Both, Subham Choudhury, Pierre Sens, and Remy Kusters. DeepMoD: Deep learning for model discovery in noisy data. *Journal of Computational Physics*, 428:109985, 2021.

[BD11] Anirban Bhattacharya and David B Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

[BDPW19] Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427, 2019.

[Bea19] Mark A Beaumont. Approximate bayesian computation. *Annual review of statistics and its application*, 6:379–403, 2019.

[BFT13] Sergio Bacallado, Stefano Favaro, and Lorenzo Trippa. Bayesian non-

parametric analysis of reversible markov chains. *The Annals of Statistics*, pages 870–896, 2013.

[BGL17] Austin R Benson, David F Gleich, and Lek-Heng Lim. The spacey random walk: A stochastic process for higher-order data. *SIAM Review*, 59(2):321–345, 2017.

[BHL93] Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575, 1993.

[Bil61] Patrick Billingsley. Statistical methods in markov chains. *The annals of mathematical statistics*, pages 12–40, 1961.

[BL07a] Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pages 177–186, 2007.

[BL07b] Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104(24):9943–9948, 2007.

[BLST19] Frank Ball, Catherine Larédo, David Sirl, and Viet Chi Tran. *Stochastic epidemic models with inference*, volume 2255. Springer Nature, 2019.

[Boc22] Natalia Bochkina. Bernstein-von mises theorem and misspecified models: a review. *arXiv preprint arXiv:2204.13614*, 2022.

[BPK16] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[BPPD15] Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

[BS05] Mogens Bladt and Michael Sørensen. Statistical inference for discretely observed markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):395–410, 2005.

[BTBK14] Steven L Brunton, Jonathan H Tu, Ido Bright, and J Nathan Kutz. Compressive sensing and low-rank libraries for classification of bifurcation regimes in nonlinear dynamical systems. *SIAM Journal on Applied Dynamical Systems*, 13(4):1716–1732, 2014.

[Buc14] Edgar Buckingham. On physically similar systems: Illustrations of the use of dimensional equations. *Physical Review*, 4(4):345, 1914.

[CBL20] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[CKP14] Ismaël Castillo, Gérard Kerkyacharian, and Dominique Picard. Thomas bayes' walk on manifolds. *Probability Theory and Related Fields*, 158(3-4):665–710, 2014.

[CL06] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[CL22] Xiuyuan Cheng and Boris Landa. Bi-stochastically normalized graph laplacian: convergence to manifold laplacian and robustness to outlier noise. *arXiv preprint arXiv:2206.11386*, 2022.

[CMJZ21] Jialei Chen, Simon Mak, V Roshan Joseph, and Chuck Zhang. Function-on-function kriging, with applications to three-dimensional printing of aortic tissues. *Technometrics*, 63(3):384–395, 2021.

[Con14] Benjamin Connault. A weakly dependent bernstein–von mises theorem. Technical report, Working Paper, 2014.

[Con16] Benjamin Connault. Hidden rust models. *Unpublished manuscript, Univ. Pa., Philadelphia*, 2016.

[CPS09] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR, 2009.

[CPS10] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

[CS07] Taeryon Choi and Mark J Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007.

[CS10] Haiyan Cheng and Adrian Sandu. Collocation least-squares polynomial chaos method. In *Proceedings of the 2010 Spring Simulation Multiconference*, pages 1–6, 2010.

[CSHVdV15] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

[CVE06] DT Crommelin and Eric Vanden-Eijnden. Fitting timeseries by continuous-time markov chains: A quadratic programming approach. *Journal of Computational Physics*, 217(2):782–805, 2006.

[CW22] Xiuyuan Cheng and Hau-Tieng Wu. Convergence of graph laplacian with knn self-tuned kernels. *Information and Inference: A Journal of the IMA*, 11(3):889–957, 2022.

[CZA+20] Kathleen Champion, Peng Zheng, Aleksandr Y Aravkin, Steven L Brunton, and J Nathan Kutz. A unified sparse optimization framework to learn parsimonious physics-informed models from data. *IEEE Access*, 8:169259–169271, 2020.

[DCS21] Joel Dyer, Patrick Cannon, and Sebastian M Schmon. Approximate bayesian computation with path signatures. *arXiv preprint arXiv:2106.12555*, 2021.

[DK22] Charles B Delahunt and J Nathan Kutz. A toolkit for data-driven discovery of governing equations in high-noise regimes. *IEEE Access*, 10:31210–31234, 2022.

[DKS13] Josef Dick, Frances Y Kuo, and Ian H Sloan. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

[DMW19] Liang Ding, Simon Mak, and C F J Wu. BdryGP: A new Gaussian process model for incorporating boundary information. *arXiv preprint arXiv:1908.08868*, 2019.

[DN15] Bryan C Daniels and Ilya Nemenman. Automated adaptive inference of phenomenological dynamical models. *Nature Communications*, 6(1):1–8, 2015.

[DNR11] David K Duvenaud, Hannes Nickisch, and Carl Rasmussen. Additive Gaussian processes. *Advances in Neural Information Processing Systems*, 24, 2011.

[DR06] Persi Diaconis and Silke WW Rolles. Bayesian analysis for reversible markov chains. *The Annals of Statistics*, 34(3):1270–1292, 2006.

[DWW20] David B Dunson, Hau-Tieng Wu, and Nan Wu. Graph based Gaussian processes on restricted domains. *arXiv preprint arXiv:2010.07242*,

2020.

[DWW22] David B Dunson, Hau-Tieng Wu, and Nan Wu. Graph based gaussian processes on restricted domains. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):414–439, 2022.

[DYND20] Leo L Duan, Alexander L Young, Akihiko Nishimura, and David B Dunson. Bayesian constraint relaxation. *Biometrika*, 107(1):191–204, 2020.

[EDB18] Alexander Erreygers and Jasper De Bock. Computing inferences for large-scale continuous-time markov chains by combining lumping with imprecision. In *International Conference Series on Soft Methods in Probability and Statistics*, pages 78–86. Springer, 2018.

[FD21] Federico Ferrari and David B Dunson. Bayesian factor analysis for inference on interactions. *Journal of the American Statistical Association*, 116(535):1521–1532, 2021.

[Fil91] James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process. *The annals of applied probability*, pages 62–87, 1991.

[GBCC15] Shirin Golchi, Derek R Bingham, Hugh Chipman, and David A Campbell. Monotone emulation of computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):370–392, 2015.

[GCSR95] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

[GE22] Amin Ghadami and Bogdan I Epureanu. Data-driven prediction in dynamical systems: Recent developments. *Philosophical Transactions of the Royal Society A*, 380(2229):20210213, 2022.

[GGdaetB19] Robert B Gramacy, Maintainer Robert B Gramacy, and Monotone data augmentation extends this Bayesian. Package 'monomvn'. *R package version*, pages 1–9, 2019.

[GGK+23] Peter Georg, Lars Grasedyck, Maren Klever, Rudolf Schill, Rainer Spang, and Tilo Wettig. Low-rank tensor methods for markov chains with applications to tumor progression models. *Journal of Mathematical Biology*, 86(1):7, 2023.

[GGVDV00] Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

[GH13] John Guckenheimer and Philip Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, volume 42. Springer Science & Business Media, 2013.

[Gia19] Dimitrios Giannakis. Data-driven spectral decomposition and forecasting of ergodic dynamical systems. *Applied and Computational Harmonic Analysis*, 47(2):338–396, 2019.

[GNM19] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, pages 2404–2414. PMLR, 2019.

[GS91] Roger G Ghanem and Pol D Spanos. Stochastic finite element method: Response statistics. In *Stochastic Finite Elements: A Spectral Approach*, pages 101–119. Springer, 1991.

[GvdV07] Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192 – 223, 2007.

[GVdV17] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[HBDH18] Ofir Harari, Derek Bingham, Angela Dean, and Dave Higdon. Computer experiments: prediction accuracy, sample size and model complexity revisited. *Statistica Sinica*, pages 899–919, 2018.

[HBL20] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–4248. PMLR, 2020.

[HBSK22] Seth M Hirsh, David A Barajas-Solano, and J Nathan Kutz. Sparsifying priors for Bayesian uncertainty quantification in model discovery. *Royal Society Open Science*, 9(2):211823, 2022.

[HHL19] P Richard Hahn, Jingyu He, and Hedibert F Lopes. Efficient sampling for gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics*, 28(1):142–154, 2019.

[HJL90] Reinhard Höpfner, Jean Jacod, and Lucia Ladelli. Local asymptotic normality and mixed normality for markov statistical models. *Probability theory and related fields*, 86(1):105–129, 1990.

[HKWBC14] Monir Hajiaghayi, Bonnie Kirkpatrick, Liangliang Wang, and Alexan-

dre Bouchard-Côté. Efficient continuous-time markov chain estimation. In *International Conference on Machine Learning*, pages 638–646. PMLR, 2014.

[HS21] Thomas Hamm and Ingo Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *The Annals of Statistics*, 49(6):3153–3180, 2021.

[HSH+21] Yuchen He, Namjoon Suh, Xiaoming Huo, Sungha Kang, and Yajun Mei. Asymptotic theory of l_1-regularized pde identification from a single noisy trajectory. *arXiv:2103.07045*, 2021.

[HW92] Michael Hamada and C F Jeff Wu. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24(3):130–137, 1992.

[IR05] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[JBP+02] S. V. Jones, S. Brown, T. Parker, V. Stubblefield, E. Kössler, C. Simmons, and J. Wittner. Anomalous indeterminate amorphous transitory gradations observed in certain specific nongeneralizable locii attendant to unique fleeting unpredictable phenomena. *Journal of Nonspecific Research*, 88:665–703, 2002.

[JG20] Seonghyun Jeong and Subhashis Ghosal. Unified Bayesian theory of sparse linear regression with nuisance parameters. *arXiv preprint arXiv:2008.10230*, 2020.

[JHD20] Michael Jauch, Peter D Hoff, and David B Dunson. Random orthogonal matrices and the Cayley transform. *Bernoulli*, 26(2):1560–1586, 2020.

[JHD21] Michael Jauch, Peter D Hoff, and David B Dunson. Monte carlo simulation on the Stiefel manifold via polar expansion. *Journal of Computational and Graphical Statistics*, 30(3):622–631, 2021.

[JM15] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[JMS+21] Yi Ji, Simon Mak, Derek Soeder, JF Paquet, and Steffen A Bass. A graphical Gaussian process model for multi-fidelity emulation of expensive computer codes. *arXiv preprint arXiv:2108.00306*, 2021.

[JOB20] James Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Scal-

able approximate MCMC algorithms for the horseshoe prior. *Journal of Machine Learning Research*, 21(73):1–61, 2020.

[KBH+11] Cari G Kaufman, Derek Bingham, Salman Habib, Katrin Heitmann, and Joshua A Frieman. Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *The Annals of Applied Statistics*, 5(4):2470–2492, 2011.

[KG13] Samory Kpotufe and Vikas Garg. Adaptivity to local smoothness and dimension in kernel regression. *Advances in neural information processing systems*, 26, 2013.

[KL14] Georgios Karagiannis and Guang Lin. Selection of polynomial chaos bases via bayesian model uncertainty methods with applications to sparse approximation of pdes with stochastic inputs. *Journal of Computational Physics*, 259:114–134, 2014.

[KLL94] James Kuelbs, Wenbo V Li, and Werner Linde. The gaussian measure of shifted balls. *Probability Theory and Related Fields*, 98(2):143–162, 1994.

[KP95] Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

[Kpo11] Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24, 2011.

[Kre85] U Krengel. Ergodic theorems, de greuter. *Berlin, New York*, 1985.

[LBH22] Xiaoyu Lu, Alexis Boukouvalas, and James Hensman. Additive Gaussian processes revisited. In *International Conference on Machine Learning*, pages 14358–14383. PMLR, 2022.

[LDD20] Sirio Legramanti, Daniele Durante, and David B Dunson. Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, 107(3):745–752, 2020.

[LMS21] Nora Lüthen, Stefano Marelli, and Bruno Sudret. Sparse polynomial chaos expansions: Literature survey and benchmark. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):593–649, 2021.

[Lor63] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130–141, 1963.

[LOSC20] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaus-

sian process meets big data: A review of scalable gps. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11):4405–4423, 2020.

[LOT+11] Yiqi Luo, Kiona Ogle, Colin Tucker, Shenfeng Fei, Chao Gao, Shannon LaDeau, James S Clark, and David S Schimel. Ecological forecasting and data assimilation in a data-rich era. *Ecological Applications*, 21(5):1429–1442, 2011.

[LWZ18] Xudong Li, Mengdi Wang, and Anru Zhang. Estimation of markov chain via rank-constrained likelihood. In *International Conference on Machine Learning*, pages 3033–3042. PMLR, 2018.

[MDJS07] Philipp Metzner, Evelyn Dittmer, Tobias Jahnke, and Ch Schütte. Generator estimation of markov jump processes. *Journal of Computational Physics*, 227(1):353–375, 2007.

[MFK08] Andrew J Majda, Christian Franzke, and Boualem Khouider. An applied mathematics perspective on stochastic modelling for climate. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1875):2427–2453, 2008.

[MGM12] Iain Murray, Zoubin Ghahramani, and David MacKay. Mcmc for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*, 2012.

[MHD22] James Matuk, Amy H Herring, and David B Dunson. Bayesian modeling of nearly mutually orthogonal processes. *arXiv preprint arXiv:2205.12361*, 2022.

[MN89] Peter McCullagh and John A Nelder. *Generalized Linear Models*, volume 37. Chapman & Hall, 1989.

[MN20] Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *arXiv preprint arXiv:2006.10605*, 2020.

[MPF20] Linda Möstel, Marius Pfeuffer, and Matthias Fischer. Statistical inference for markov chains with applications to credit risk. *Computational Statistics*, 35(4), 2020.

[MSW+18] Simon Mak, Chih-Li Sung, Xingjian Wang, Shiang-Ting Yeh, Yu-Hung Chang, V Roshan Joseph, Vigor Yang, and C F Jeff Wu. An efficient surrogate model for emulation and physics extraction of large eddy simulations. *Journal of the American Statistical Association*, 113(524):1443–1456, 2018.

[Mud19] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*, 190:310–322, 2019.

[MW19] Simon Mak and C F Jeff Wu. cmenet: A new method for bi-level variable selection of conditional main effects. *Journal of the American Statistical Association*, 114(526):844–856, 2019.

[Naw28] B. Nawahi. Acoustic production in liquid filled ceramic environments. *J. Chem. Phys.*, 108:9893–9904, 1928.

[NDCW23] Mu Niu, Zhenwen Dai, Pokman Cheung, and Yizhu Wang. Intrinsic gaussian process on unknown manifolds with probabilistic metrics. *arXiv preprint arXiv:2301.06533*, 2023.

[Nel77] JA Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society: Series A (General)*, 140(1):48–63, 1977.

[NNM+96] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). *technical report*, 1996.

[NWS22] Joshua S North, Christopher K Wikle, and Erin M Schliep. A review of data-driven discovery for dynamic systems. *arXiv preprint arXiv:2210.10663*, 2022.

[Owe97] Art B Owen. Scrambled net variance for integrals of smooth functions. *The Annals of Statistics*, 25(4):1541–1562, 1997.

[Pau15] Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 2015.

[PB21] Thomas P Prescott and Ruth E Baker. Multifidelity approximate bayesian computation with sequential monte carlo parameter sampling. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):788–817, 2021.

[PDLN18] Leah F Price, Christopher C Drovandi, Anthony Lee, and David J Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

[PPRS12] Omiros Papaspiliopoulos, Yvo Pokern, Gareth O Roberts, and Andrew M Stuart. Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, 99(3):511–531, 2012.

[PSKL+21] Henri Pesonen, Umberto Simola, Alvaro Köhn-Luque, Henri Vuollekoski, Xiaoran Lai, Arnoldo Frigessi, Samuel Kaski, David T

Frazier, Worapree Maneesoonthorn, Gael M Martin, et al. Abc of the future. *arXiv preprint arXiv:2112.12841*, 2021.

[PSM19] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

[Ras03] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[RBPK17] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

[RJ04] William J Reed and Murray Jorgensen. The double Pareto-lognormal distribution—a new parametric model for size distributions. *Communications in Statistics-Theory and Methods*, 33(8):1733–1753, 2004.

[RPK17] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.

[RPK19] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[RZMC11] Mary A Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of chemical physics*, 134(12):03B624, 2011.

[San05] F. Sanford. Oxidized ferrous materials. *Journal of Junk*, 5(4):324–345, 2005.

[SB13] Sonja Surjanovic and Derek Bingham. Virtual library of simulation experiments: test functions and datasets. *Simon Fraser University, Burnaby, BC, Canada, accessed May*, 13:2015, 2013.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[SCK20] Sebastian M Schmon, Patrick W Cannon, and Jeremias Knoblauch. Generalized posteriors in approximate bayesian computation. *arXiv preprint arXiv:2011.08644*, 2020.

[Sen06] Eugene Seneta. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

[SFK08] András Sobester, Alexander Forrester, and Andy Keane. *Engineering Design via Surrogate Modeling: A Practical Guide*. John Wiley & Sons, 2008.

[SFK12] Fabian Scheipl, Ludwig Fahrmeir, and Thomas Kneib. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012.

[SH70] M. Smith and I. A. Hall. *Handbook of Interstellar Travel*. Dover, New York, 7th edition, 1970.

[She21] Chris Sherlock. Direct statistical inference for finite markov jump processes via the matrix exponential. *Computational Statistics*, 36(4):2863–2887, 2021.

[SL09] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[SL17] Qifan Song and Faming Liang. Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv:1712.08964*, 2017.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25, 2012.

[SN06] Clayton Scott and Robert D Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE transactions on information theory*, 52(4):1335–1353, 2006.

[Ste99] Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 1999.

[SVS11] Terrance Savitsky, Marina Vannucci, and Naijun Sha. Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statistical Science*, 26(1):130, 2011.

[SWNW03] Thomas J Santner, Brian J Williams, William I Notz, and Brain J Williams. *The Design and Analysis of Computer Experiments*, volume 1. Springer, 2003.

[TCBD+20] Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J Gonçalves, David S Greenberg,

and Jakob H Macke. Sbi–a toolkit for simulation-based inference. *arXiv preprint arXiv:2007.09114*, 2020.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[TVA+19] Dustin Tran, Keyon Vafa, Kumar Agrawal, Laurent Dinh, and Ben Poole. Discrete flows: Invertible generative models of discrete data. *Advances in Neural Information Processing Systems*, 32:14719–14728, 2019.

[TW17] Giang Tran and Rachel Ward. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, 15(3):1108–1129, 2017.

[VdPSC+19] Stephanie Van der Pas, James Scott, Antik Chakraborty, Anirban Bhattacharya, and Maintainer Stephanie van der Pas. Package 'horseshoe'. *package*, 2019.

[VdV00] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[vdVvZ08] Aad W van der Vaart and J Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

[VdVVZ09] Aad W Van der Vaart and J Harry Van Zanten. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 2009.

[Wan08] Limin Wang. *Karhunen-Loeve Expansions and Their Applications*. London School of Economics and Political Science, 2008.

[WB16] Xiaojing Wang and James O Berger. Estimating shape constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25, 2016.

[WDP+14] Matthew W Wheeler, David B Dunson, Sudha P Pandalai, Brent A Baker, and Amy H Herring. Mechanistic hierarchical Gaussian processes. *Journal of the American Statistical Association*, 109(507):894–904, 2014.

[Wen08] John F Wendt. *Computational Fluid Dynamics: An Introduction*. Springer Science & Business Media, 2008.

[WH09] C F Jeff Wu and Michael S Hamada. *Experiments: Planning, Analysis, and Optimization*. John Wiley & Sons, 2009.

[WPBS21] David J Warne, Thomas P Prescott, Ruth E Baker, and Matthew J Simpson. Multifidelity multilevel monte carlo to accelerate approximate bayesian parameter inference for partially observed stochastic processes. *arXiv preprint arXiv:2110.14082*, 2021.

[WWHW19] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

[Xiu10] Dongbin Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.

[YD16] Yun Yang and David B. Dunson. Bayesian manifold regression. *The Annals of Statistics*, 44(2):876 – 905, 2016.

[Yvi15] Romain Yvinec. *Piecewise deterministic Markov processes, applications in biology*. PhD thesis, Université Pierre et Marie Curie-Paris 6 (UPMC), 2015.

[YWS+18] Shiang-Ting Yeh, Xingjian Wang, Chih-Li Sung, Simon Mak, Yu-Hung Chang, Liwei Zhang, C F Jeff Wu, and Vigor Yang. Common proper orthogonal decomposition-based spatiotemporal emulator for design exploration. *AIAA journal*, 56(6):2429–2442, 2018.

[YZ05] George Yin and Qing Zhang. *Discrete-time Markov chains: two-time-scale methods and applications*, volume 55. Springer Science & Business Media, 2005.

[YZ08] Gui-Bo Ye and Ding-Xuan Zhou. Learning and approximation by gaussians on riemannian manifolds. *Advances in Computational Mathematics*, 29(3), 2008.

[YZ09] Gui-Bo Ye and Ding-Xuan Zhou. Svm learning and lp approximation by gaussians on riemannian manifolds. *Analysis and Applications*, 7(03):309–339, 2009.

[ZH05] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[ZLWZ22] Ziwei Zhu, Xudong Li, Mengdi Wang, and Anru Zhang. Learning markov models via low-rank optimization. *Operations Research*, 70(4):2384–2398, 2022.

[ZW19] Anru Zhang and Mengdi Wang. Spectral state compression of markov processes. *IEEE transactions on information theory*, 66(5):3202–3231, 2019.

Biography

Tang Tao was born in 1996 in Chengdu, Sichuan, China. He graduated from Chengdu No.7 High School in 2014 and earned dual bachelor's degrees in Biology and Mathematics from Peking University in 2018. He is completing his Ph.D. in Mathematics at Duke University in 2023.

His research interests lie in the applications of probabilistic and statistical methods in biology and physical science, particularly in Uncertainty Quantification, Bayesian Methods, and Sampling Methods.