Nonparametric Bayes for Big Data

by

Yun Yang

Department of Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Surya T. Tokdar

_____
Li Ma

_____
Barbara Engelhardt

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Department of Statistical Science
in the Graduate School of Duke University
2014

Abstract

Nonparametric Bayes for Big Data

by

Yun Yang

Department of Department of Statistical Science
Duke University

Date: _____

Approved:

_____
David B. Dunson, Supervisor

_____
Surya T. Tokdar

_____
Li Ma

_____
Barbara Engelhardt

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Department of Statistical Science
in the Graduate School of Duke University
2014

Abstract

Classical asymptotic theory deals with models in which the sample size n goes to infinity with the number of parameters p being fixed. However, rapid advancement of technology has empowered today's scientists to collect a huge number of explanatory variables to predict a response. Many modern applications in science and engineering belong to the "big data" regime in which both p and n may be very large. A variety of genomic applications even have p substantially greater than n. With the advent of MCMC, Bayesian approaches exploded in popularity. Bayesian inference often allows easier interpretability than frequentist inference. Therefore, it becomes important to understand and evaluate Bayesian procedures for "big data" from a frequentist perspective. In this dissertation, we address a number of questions related to solving large-scale statistical problems via Bayesian nonparametric methods.

It is well-known that classical estimators can be inconsistent in the high-dimensional regime without any constraints on the model. Therefore, imposing additional low-dimensional structures on the high-dimensional ambient space becomes inevitable. In the first two chapters of the thesis, we study the prediction performance of high-dimensional nonparametric regression from a minimax point of view. We consider two different low-dimensional constraints: 1. the response depends only on a small subset of the covariates; 2. the covariates lie on a low dimensional manifold in the original high dimensional ambient space. We also provide Bayesian nonparametric methods based on Gaussian process priors that are shown to be adaptive to

unknown smoothness or low-dimensional manifold structure by attaining minimax convergence rates up to log factors. In chapter 3, we consider high-dimensional classification problems where all data are of categorical nature. We build a parsimonious model based on Bayesian tensor factorization for classification while doing inferences on the important predictors.

It is generally believed that ensemble approaches, which combine multiple algorithms or models, can outperform any single algorithm at machine learning tasks, such as prediction. In chapter 5, we propose Bayesian convex and linear aggregation approaches motivated by regression applications. We show that the proposed approach is minimax optimal when the true data-generating model is a convex or linear combination of models in the list. Moreover, the method can adapt to sparsity structure in which certain models should receive zero weights, and the method is tuning parameter free unlike competitors. More generally, under an M-open view when the truth falls outside the space of all convex/linear combinations, our theory suggests that the posterior measure tends to concentrate on the best approximation of the truth at the minimax rate.

Chapter 6 is devoted to sequential Markov chain Monte Carlo algorithms for Bayesian on-line learning of big data. The last chapter attempts to justify the use of posterior distribution to conduct statistical inferences for semiparametric estimation problems (the semiparametric Bernstein von-Mises theorem) from a frequentist perspective.

To my family

Contents

List of Tables

List of Figures

Acknowledgements

To begin with, I would like to express the deepest gratitude to my advisor Dr. David B. Dunson for his continued support and encouragement. During my Ph.D. study, he gives me the complete freedom of choosing research topics. He listens to any ideas I might have, even as a starting graduate student, and provides pertinent and insightful feedback. His great enthusiasm for research and tremendous energy as a scholar never fail to surprise and influence me.

I thank my committee members, Dr. Surya Tokdar, Dr. Li Ma and Dr. Barbara Engelhardt for their time and support. Dr. Li Ma serves as my first year advisor and gives me luminous guidance that is very helpful for a fresh graduate student. I would like to further thank Dr. Surya Tokdar, my secondary advisor, who teaches me not only Bayesian asymptotic theory but also how to become a real scholar. Thanks also to Dr. Mike West, Dr. Alan Gelfand, Dr. Merlise Clyde, Dr. Scott Schmidler, Dr. Robert Wolpert for their wonderful lectures. It was also a privilege to work with Dr. Guang Cheng at Purdue Unniversity — I have learned a great deal from him not just in technical things, but also other aspects of research such as technical writing.

Thanks go to my fellow Ph.D. colleagues and I would cherish many fond memories with them. To Brian, Doug, James, Monica, Shaan, Shih-Han, Jacopo, Zoey, Maria, Tommy, Nicole, Tsuyoshi, Anirban, Kai and many others: Thank you!

Thanks to Fangpo for her guidance and help in my first two years.

Thanks to Xiangyu as being such a wonderful roommate and table tennis play-

mate.

Thanks to Kang and Zhao as being my best friends outside our department.

Finally, I dedicate this work to my family: my parents, without whom none of this would have been possible. Their unwavering support from the other side of the globe has been the biggest strength in my academic quest.

1

Introduction

1.1 Motivation

Classical asymptotic theory deals with models in which the sample size $n \to \infty$ with the number of parameters p being fixed. However, rapid advancement of technology has empowered today's scientists to collect a huge number of explanatory variables to predict a response. Many modern applications in science and engineering belong to the "big data" regime in which both p and n may be very large. In finance, market data comprises high-frequency measurements of hundreds or thousands of financial instruments over time, leading to many statistical challenges (Fan et al., 2011). A variety of genomic applications fall into the high-dimensional statistics paradigm in which p may even be substantially larger than n. For example, in genome-wide association studies, hundreds of thousands of single-nucleotide polymorphisms are potentially relevant genetic markers for studying human diseases.

It is of fundamental importance to study under what assumptions a particular statistical problem is tractable. For example, it is well known that classical estimators become inconsistent in the regime $p \gg n$ without any additional constraints on

the model. Therefore, a variety of studies try to impose some low-dimensional structures on the high-dimensional ambient space, and quantify performance of different estimators. For example, in high-dimensional linear regression literature, people tend to assume the sparsity condition, under which the response only depends on $d = o(n)$ important predictors among a list of p predictors. In matrix completion problems, the true matrix to be estimated is assumed to be of a low-rank. To judge whether a statistical problem is well-defined, one way is to study its minimax property, which quantifies the best worst case performance that an estimator can achieve. Minimax risks are often related to the size of the model space in terms of metric entropies, which compete with the statistical power of discriminating the truth from others in the model space. The statistical power usually depends on the available information characterized by sample size n. Therefore, high-dimensional statistical problems are solvable if and only if the size of the model space is compatible with the statistical power based on a sample of size n. This explains the reason for seeking various reasonable low-dimensional constraints to restrict high-dimensional problems.

The frequentist literature illustrates the success of applying optimization methods for large scale problems. Many well-known estimators are constructed via penalized M-estimation, where a regularizer penalizes the deviation of the parameter from the low-dimensional structure. An optimal choice for the regularization parameter, which determines the amount of penalization, typically involves some prior knowledge on the true data generating model, such as the number of important predictors in high-dimensional linear regression or the true rank in matrix completion problems. In practice, regularization parameters are often determined via cross-validation. However, a main disadvantage of cross-validation is that every time only a subset of observations are used to fit the model.

In recent years, there has been an emergence of interest in conducting statistical inference for large scale data based on Bayesian procedures. Unlike the optimization

focus of usual frequentist methods, typical Bayesian estimators rely on integration. Performance of Bayesian estimators can be assessed from a frequentist perspective by viewing the observations as generated from a true underlying distribution. Under this perspective, it turns out that many Bayesian procedures with properly chosen priors for large scale problems can accommodate the potential low-dimensional structure in the data — the estimators can automatically adapt to the unknown sparsity level, smoothness level or manifold structure and achieve the minimax optimal convergence rate (van der Vaart and van Zanten, 2009; Rousseau and Mengersen, 2011; Yang and Dunson, 2013; Castillo and van der Vaart, 2012). Therefore, in contrast to many frequentist competitors, these Bayesian procedures do not require any prior knowledge on the truth and are tuning free.

In many applications, the relationship between a response Y and its explanatory variables $X = (X^1, ..., X^p) \in \mathbb{R}^p$ may be highly nonlinear and include interaction. It is of practical importance to develop sensible models with mild assumptions on the relationship between X and Y. This motivates us to treat the structure of this relationship nonparametrically. One way of constructing a nonparametric model is to allow a growing number of parameters to accommodate the complexity of the data. Examples include mixture models with increasing number of components and nonparametric sieve regression (Geman and Hwang, 1982; Hansen, 2012). Under this perspective, high-dimensional parametric models such as linear models can also be treated as nonparametric. Another class of nonparametric models are models whose parameter space are infinite-dimensional. For example, consider a regression model $Y = f(X) + \epsilon$, $\epsilon \sim N(0, \sigma^2)$. If the only assumption on the regression function f is twice differentiable or monotone constraints, then f cannot be characterized by a finite number of parameters and the resulting regression problem is nonparametric. Intuitively, such flexible modeling assumptions allow us to learn the structure of f on a growing resolution scale as more data are collected.

Although optimization methods are often good at obtaining sensible point estimators, they do not provide a natural way to conduct statistical inferences, such as uncertainty quantification. In practice people tend to apply resampling techniques, such as the bootstrap or subsampling, to approximate the sampling distribution of their estimators. In contrast, fully model-based Bayesian procedures offer a standard way to doing inferences based on posterior distributions. It is then of fundamental importance to justify the validity of Bayesian statistical inferences from a frequentist perspective. This justification is especially important for large scale problems since it provides a guidance on how many observations one needs to collect in order to achieve certain estimation accuracy. A well-known result for regular parametric models is given by the Bernstein von-Mises theorem, which states that the posterior distribution tend to converge in total variation distance to a normal distribution centered at a Bayesian estimator $\tilde{\theta}$ with variance the same as the asymptotic variance of $\tilde{\theta}$. As a result, the coverage of the corresponding Bayesian credible region asymptotically coincides with its nominal level. Whether similar frequentist justification for more complicated Bayesian methods, such as semiparametric/nonparametric or high-dimensional procedures, can be proved is still an open question.

With the above motivations in mind, we start to introduce the topics considered in this thesis.

1.2 Research questions and main contributions

Motivated by large scale data, the primary focus of this thesis is on developing practically efficient Bayesian methodology having strong theoretical guarantees. In this section, we briefly summarize the central research questions addressed in later chapters.

Nonparametric regression in high dimensions

The first research question is concerned with the prediction performance of nonparametric regression in the high-dimensional regime from a minimax point of view. Since parametric models in reality can seldom capture the exact dependence structure, it is important to develop sensible regression models

$$y = f(x) + \epsilon, \quad \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

to predict the response y under mild assumptions on f in the high dimensional setting where the sample size n is smaller than the dimensionality p of the covariate vector $x = (x_1, \ldots, x_p)$.

Good statistical methods for *large p small n* regression should scale well with the predictor dimensions and quickly identify any underlying low dimensional structures to facilitate maximum statistical learning from limited data. They must also allow flexible estimation of the function shape and capture predictor interaction. Motivated by these requirements, three types of modeling assumptions are considered: 1. the regression function f depends on d covariates, and $d \ll \min\{n, p\}$, but is otherwise of an arbitrary form; 2. f still depends on a small subset of the covariates, but has an additive form as $\sum_{s=1}^{k} f_s$, where each additive component f_s depends on a small number $d_s \ll \min\{n, p\}$ of covariates which can be different across s; 3. f can potentially depend on all covariates, but the covariate vector $x = (x_1, \ldots, x_p)$ is assumed to lie on a low dimensional manifold $\mathcal{M}$ in the ambient space $\mathbb{R}^p$.

To assess the performance of high-dimensional nonparametric models, we describe a general framework to show the minimax risks for regression problems under L_2 loss. Our contribution is the construction of a general class of Bayesian sieve estimators, which are shown to attain the minimax lower bound provided by Fano's lemma. By applying this general framework, we study the minimax risks for estimating f under the first two sparse assumptions. The minimax risks are shown to be the

sum of two terms: estimation risks and variable selection risks. The estimation risks are the minimax risks of estimating the regression functions as if we knew which predictors are important and the variable selection risks reflect the variable selection uncertainty. We also show that Bayesian nonparametric regression based on Gaussian process (GP) priors and variable selections can not only achieve minimax optimal rates, but are also adaptive to the unknown smoothness levels and numbers of important predictors.

Under the third assumption of low dimensional manifold, it is clear that probabilistic models for learning the manifold $\mathcal{M}$ face daunting statistical and computational hurdles. Therefore, we take a very different approach in attempting to define a simple and computationally tractable model, which bypasses the need to estimate $\mathcal{M}$ but can exploit the lower dimensional manifold structure when it exists. We prove that a simple GP prior with a random length-scale parameter could lead to the minimax-optimal rate in estimating f, and the rate is adaptive to the manifold and smoothness of the regression function. Moveover, we find a counter-intuitive blessing of dimensionality phenomenon, which suggests that by applying random projections, large p facilitates reducing the independent additive noise in x.

High-dimensional nonparametric classification for categorical data

In the second research problem, we consider high-dimensional problems where all data are of categorical nature. The goal is to build a parsimonious model for classification while doing inferences on the important predictors. With categorical predictors, the conditional probabilities $P(Y = y \,|\, X_1 = x_1, \ldots, X_p = x_p)$ can be cast into a $d_1 \times \cdots \times d_p$ tensor for each class label y, with d_j denoting the number of levels of the jth categorical predictor X_j. We use a carefully-structured Tucker factorization to define a model that can characterize any conditional probability, while facilitating variable selections and capturing of higher-order interactions. To overcome the curse

of dimensionality, we make a near low-rank assumption on the conditional probability tensor, under which the posterior is shown to achieve a convergence rate of order $\sqrt{\log p/n}$ up to a $\log n$ term in high dimensional settings. The low-rank assumption for categorical predictors resembles the sparsity assumption for continuous predictors. The real data examples illustrate that this low-rank assumption yields satisfactory classification performance when our model is compared to the state-of-the-art classifiers. In Cornelis et al. (2013), an application of conditional tensor factorization model to crack detection in ultra high resolution multimodal images of paintings demonstrates its potential in solving real high dimensional problems.

Bayesian aggregation in statistical learning

The third research problem focues on Bayesian ensemble learning procedures via aggregation. In many applications, it is not at all clear how to pick one most suitable method out of a list of possible models or learning algorithms $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. Each model/algorithm has its own set of implicit or explicit assumptions under which that approach will obtain at or near optimal performance. However, in practice verifying which if any of these assumptions hold for a real application is problematic. Hence, it is of substantial practical importance to have an aggregating mechanism that can automatically combine the estimators $\hat{f}_1, \ldots, \hat{f}_M$ obtained from the M different approaches $\mathcal{M}_1, \ldots, \mathcal{M}_M$, with the aggregated estimator potentially better than any single one.

Bayesian methods are appealing in providing a probabilistic approach for combining different models together. For example, Bayesian model averaging (BMA) is a widely used approach in practice. The justification for BMA arises from the viewpoint that one of the listed models in the ensemble is the correct underlying model that generates the data. Then, in many cases, as the sample size increases, the posterior probability on this true model converges to one. If the true model is

not in the list, the model with the minimal KL divergence from the true model will instead be assigned probability that is converging to one.

To formally allow the true model to fall outside the ensemble in the Bayesian framework, we propose to aggregate different models instead of averaging them. We focus on two main aggregation strategies: convex aggregation (CA) and linear aggregation (LA). CA aims at selecting the optimal convex combination of the estimators and LA focuses on choosing the optimal linear combination. Modeling the model-specific weights via symmetric Dirichlet distributions, we show that our Bayesian approach obtains the minimax optimal rate up to a log factor of convex/linear aggregation (Tsybakov, 2003). Even if the true model is not a convex/linear combination of the models in the ensemble, we show that the posterior would concentrate around the best approximation of the truth.

Sequential MCMC for on-line learning

The fourth research topic is Bayesian on-line learning for big data. We propose a sequential Markov chain Monte Carlo (SMCMC) algorithm to sample from a sequence of probability distributions $\{\pi_t : t \geqslant 0\}$, which correspond to posterior distributions at different times in on-line applications. SMCMC proceeds as in usual MCMC but with the stationary distribution updated appropriately each time new data arrive. SMCMC has advantages over sequential Monte Carlo (SMC) in avoiding particle degeneracy issues. We provide theoretical guarantees for the marginal convergence of SMCMC under various settings, including both parametric and nonparametric models. Even in batch situations where a full dataset $\{y_1, \ldots, y_n\}$ has been obtained, we can still consider the sequence of posterior distributions $p(\theta^{(t)}|y_1, \ldots, y_t)$ for $t \leqslant n$. The annealing effect (Chopin, 2002) of adding data sequentially can lead to substantial improvements over usual MCMC methods, which incorporate all the data at once and sample serially.

In the theoretical aspect, we prove the ergodicity of a time-inhomogeneous Markov chain with time varying transition kernel $\{T_t : t \geqslant 0\}$, i.e.

$$||T_t \circ \cdots \circ T_1 \circ \pi_0 - \pi_t||_{TV} \to 0, \quad \text{as } t \to \infty,$$

where π_t is the stationary measure associated with T_t, under the assumption that $\{\pi_t : t \geqslant 0\}$ forms a Cauchy sequence, i.e. $||\pi_t - \pi_{t-1}||_{TV} \to 0$, as $t \to \infty$ with $||\cdot||_{TV}$ the total variation distance. We propose a novel condition on verifying the geometric ergodicity of a time-homogenous Markov chain, which greatly simplifies and is weaker than the commonly used conditions such as the local minorization and drift condition (Rosenthal, 1995). In addition, we generalize the SMCMC algorithm and its ergodicity to the case when the dimension of the parameter space is also growing in time.

Bayesian inference for semi-parametric models

The last research topic consider semiparametric estimation problems, where the statistical model $\mathcal{P} = \{P_\lambda : \lambda = (\theta, \eta)\}$ is indexed by two parameters θ and η, with $\theta \in \mathbb{R}^k$ a finite-dimensional parameter of interest and $\eta \in \mathcal{H}$ an infinite-dimensional nuisance parameter. We justify the use of Bayesian credible intervals for θ by studying its frequentist coverage as the sample size goes to infinity based on the so-called Bernstein-von Mises (BvM) theorem. For frequentists considering using a Bayes procedure for uncertainty quantification, it is highly appealing that credible intervals have valid coverage asymptotically. BvM theorems have been established for the marginal posterior of finite dimensional parameter θ in semiparametric models (Shen, 2001; Bickel and Kleijn, 2012; Castillo and van der Vaart, 2012), which state that under certain conditions,

$$\sup_A \left|\Pi(\theta \in B | X_1, \ldots, X_n) - N_k\big(B; \theta_0 + n^{-1/2}\tilde{\Delta}_n, (nI_{\theta_0,\eta_0})^{-1}\big)\right| \to 0, \tag{1.1}$$

in P_{θ_0,η_0}-probability, where $X_1, \ldots, X_n$ are i.i.d observations,

$$\tilde{\Delta}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}^{-1}_{\theta_0,\eta_0} \tilde{l}_{\theta_0,\eta_0}(X_i),$$

$\tilde{l}_{\theta,\eta}$ is the efficient score function and $\tilde{I}_{\theta,\eta}$ the efficient Fisher information. However, results based merely on the first-order expansion of the marginal posterior of θ as (1.1) are unable to reveal how the estimating efficiency of the nuisance parameter η impacts the estimation of θ. Such a delicate relationship can only be revealed by considering a higher-order expansion.

We consider a fully Bayesian framework by putting a joint prior on (θ, η) that is shown to lead to an adaptive convergence rate in estimating η. Moreover, we consider more general cases where the likelihoods are substituted with quasi-likelihoods, which only require assumptions on the forms of the conditional means of Y given (θ, η) instead of assumptions on the complete information about the conditional distribution $P(Y|\theta, \eta)$. This general setting includes generalized partial linear models as special cases. Interestingly, we observe that if independent priors are assigned to θ and η, then even the first-order convergence rate $n^{-1/2}$ of θ would be deteriorated by a bias term depending on the least favorable direction h of the semiparametric model. To eliminate this bias, we propose a dependent prior on θ and η and show that the right hand side of (1.1) for the resulting posterior becomes $O(\sqrt{n}\rho_n^2 + \sqrt{n}\rho_n\kappa_n)$, where κ_n is the approximating error of h. Moreover, this prior is shown to be adaptive to the smoothness of the nuisance part. Therefore, an adaptive second-order efficiency of estimating θ is achieved.

1.3 Outline

In Chapter 2, we derive the Minimax L_2 risks for high dimensional nonparametric regression under two sparsity assumptions: 1. the true regression surface is a sparse

function that depends only on $d = O(\log n)$ important predictors among a list of p predictors, with $\log p = o(n)$; 2. the true regression surface depends on $O(n)$ predictors but is an additive function where each additive component is sparse but may contain two or more interacting predictors and may have a smoothness level different from other components. Broad range general results are presented to facilitate sharp lower and upper bound calculations on minimax risks in terms of modified packing entropies and covering entropies, and are specialized to spaces of additive functions. For either modeling assumption, a practical extension of the widely used Bayesian Gaussian process regression method is shown to adaptively attain the optimal minimax rate (up to $\log n$ terms) asymptotically as both $n, p \to \infty$ with $\log p = o(n)$.

In Chapter 3, our focus is on developing computationally tractable and theoretically supported Bayesian nonparametric regression methods in the context where the predictors lie on a D-dimensional surface. When the subspace corresponds to a locally-Euclidean Riemannian manifold, we show that a Gaussian process regression approach can be applied that leads to the minimax optimal adaptive rate in estimating the regression function under some conditions. The proposed model bypasses the need to estimate the manifold, and can be implemented using standard algorithms for posterior computation in Gaussian processes. Finite sample performance is illustrated in an example data analysis.

In Chapter 4, we consider a categorical response and high-dimensional categorical predictors. The goal is to build a parsimonious model for classification while doing inferences on the important predictors. By using a carefully-structured Tucker factorization, we define a model that can characterize any conditional probability, while facilitating variable selection and modeling of higher-order interactions. Following a Bayesian approach, we propose a Markov chain Monte Carlo algorithm for posterior computation accommodating uncertainty in the predictors to be included. Under near low rank assumptions, the posterior distribution for the conditional probability

is shown to achieve close to the parametric rate of contraction even in ultra high-dimensional settings. The methods are illustrated using simulation examples and biomedical applications.

In Chapter 5, we propose Bayesian convex and linear aggregation approaches motivated by regression applications. We show that the proposed approach is minimax optimal when the true data-generating model is a convex or linear combination of models in the list. Moreover, the method can adapt to sparsity structure in which certain models should receive zero weights, and the method is tuning parameter free unlike competitors. More generally, under an M-open view when the truth falls outside the space of all convex/linear combinations, our theory suggests that the posterior measure tends to concentrate on the best approximation of the truth at the minimax rate. We illustrate the method through simulation studies and several applications.

In Chapter 6, we propose a class of sequential Markov chain Monte Carlo (SMCMC) algorithms to sample from a sequence of probability distributions, corresponding to posterior distributions at different times in on-line applications. SMCMC proceeds as in usual MCMC but with the stationary distribution updated appropriately each time new data arrive. We provide theoretical guarantees for the marginal convergence of SMCMC under various settings, including parametric and nonparametric models. SMCMC exhibits an encouraging improvement over competitors in a simulation study. We also consider an application to on-line nonparametric regression.

In Chapter 7, we study second order expansion of semiparametric BvM theorems and show that the right hand side in (1.1) is $O_{P_0}(\sqrt{n}\rho_n^2)$, with ρ_n the estimation error of the nonparametric part. This second order term motivates us to consider an adaptive prior for the nonparametric part to achieve second order efficiency. As has been observed in recent work by Castillo (2012) and Rivoirard and Rousseau (2012), an adaptive independent prior for parametric and nonparametric parameters

tends to cause a bias term, called semiparametric bias, that can even break down the first-order consistency. We show that by introducing prior dependence, the semiparametric bias can be eliminated by shifting the center of the prior for the nuisance parameter. As a result, a dependent prior can achieve the adaptation to the second order term under mild conditions. We provide simulations to support our theory.

Technical proofs and details are provided in chapter specific appendices at the end of this thesis.

2

High-dimensional sparse nonparametric regression

2.1 Introduction

Rapid advancement of technology has empowered today's scientists to collect a huge number of explanatory variables to predict a response (Bülmann and van de Geer, 2011). Because the relationship between a response Y and its explanatory variables $X = (X^1, \ldots, X^p) \in \mathbb{R}^p$ may be highly nonlinear and include interaction, there is a practical need to develop sensible regression models

$$Y = f(X) + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

under mild assumptions on f in the high dimensional setting, especially when p is much larger than n, the number of observations on (X, Y) available for estimating the regression function f. Good statistical methods for such so called "large p small n regression" should scale well with the predictor dimension and quickly identify any underlying low dimensional structure to facilitate maximum statistical learning from limited data. They must also allow flexible estimation of the function shape and capture predictor interaction.

Efficient statistical learning in high dimensional settings requires strong model

assumptions to avoid "the curse of dimensionality". One attractive assumption is

M1. f potentially depends on all elements of X, but X itself lies in a low dimensional manifold $\mathcal{M}^d$ in the ambient space $\mathbb{R}^p$.

M1 enables naïve nonparametric methods that algorithmically scale well with p to achieve near optimal performance guarantees (Bickel and Li, 2007; Ye and Zhou, 2008; Yang and Dunson, 2013). However for many high dimensional applications, such as gene expression studies, a low dimensional manifold assumption on X may not be tenable or verifiable. In such cases one often assumes a sparse relationship between Y and X such as

M2. f depends on a small subset of d predictors with $d \leqslant \min\{n, p\}$.

M2 has served as the springboard for many widely used regression methods, including high dimensional linear regression approaches, such as the Lasso (Tibshirani, 1996) and the Dantzig selector (Candes and Tao, 2007), and nonparametric regression methods with variable selection, such as the Rodeo (Lafferty and Wasserman, 2008) and Gaussian process regression (Tokdar, ????). The latter two allow flexible estimation of f and is able to capture interactions among the selected important predictors. However, as will be shown later, when f is allowed to be fully nonparametric, M2 enables good statistical learning only when $d \ll \min\{n, p\}$, i.e. the regression function is extremely sparse.

To rectify this without completely giving up on nonparametric shape flexibility, we introduce a third modeling assumption:

M3. f may depend on $d \asymp \min\{n, p\}$ variables but admits an additive structure $f = \sum_{s=1}^{k} f_s$, where each additive component f_s depends on a small $d_s \ll \min\{n, p\}$ number of predictors.

Clearly, M3 subsumes M2 as a special case and in Theorem 2 we reveal that M2 represents the worst end of the difficulty spectrum of statistical learning under M3 as measured by minimax error rates in estimating f under the L_2 loss. At the other end of the spectrum is the special case of a completely additive structure $f(X) = f_1(X^{i_1}) + \cdots + f_d(X^{i_d})$ for which scalable algorithms have been devised (Hastie and Tibshirani, 1986) and attractive minimax error bounds have been derived albeit under the strong assumption that all component functions f_s have the same smoothness (Koltchinskii and Yuan, 2010; Meier and Buhlmann, 2009; Ravikumar et al., 2009; Raskutti et al., 2012).

Compared to either of these two extremes, M3 provides a much more practically attractive theory of large p small n nonparametric regression. It promises to offer efficient statistical learning even when the relationship between Y and X is not extremely sparse. It also avoids the complete additivity assumption and allows explanatory variables to interact with each other. The ability to model and learn variable interaction is a feature of considerable scientific relevance to modern statistical applications.

The aim of this chapter is twofold: to derive the minimax L_2 error rates of estimating f under M3 and to show existence of practical statistical methods that offer adaptive, near optimal performance across the entire M3 model space. Toward the first goal, we present in Theorem 2 sharp upper and lower bounds on the minimax L_2 estimation error under M3 as a function of n, p, component sizes $d_1, \ldots, d_k$ and smoothness properties of the component functions $f_1, \ldots, f_k$, which are allowed to have different levels of smoothness than one another. Both Theorem 1 and the results of Raskutti et al. (2012) follow as corollaries to this general result.

Toward the second goal, we demonstrate that a conceptually straightforward extension of the widely used Gaussian process regression method (see, e.g., Rasmussen and Williams, 2006, for a review) adaptively achieves the optimal minimax rate

across all subclasses of M3 under suitable large p small n asymptotics where p grows almost exponentially in n. In this paper we restrict only to a theoretical study of this new approach, which we call the additive Gaussian process regression. A full fledged methodological development of the same is underway and will be reported elsewhere.

The rest of this chapter is organized as follows. Section 2.2 introduces the notation and some basic assumptions. Section 2.3 summarizes our main minimax results for high dimensional nonparametric regression under M2 and M3. Section 2.4 provides a general framework for characterizing minimax risks. Section 2.5 details the application of the results in section 2.4 to M2 and M3. Section 2.6 shows the adaptive minimax optimality of Bayesian Gaussian process regression. Technical proofs appear in Appendix A.

2.2 Notations

Let (X_i, Y_i), $i = 1, \ldots, n$ denote the observations on (X, Y). We make a stochastic design assumption that $X_1, \ldots, X_n$ are independent and identically distributed (IID) according to some probability measure Q on $\mathbb{R}^p$ and that $f \in L_2(Q)$, the linear space of real valued functions on $\mathbb{R}^p$ equipped with inner product $\langle f, g \rangle_Q = \int f(x)g(x)Q(dx)$ and norm $\|f\|_Q = \langle f, f \rangle^{1/2}$. We do not need to know or estimate Q for the purpose of estimating f, but it is a natural candidate to judge average loss in prediction at future observations of X drawn from Q, as will be the case under simple exchangeability assumptions. The associated minimax risk $r_n(\Sigma, Q, \sigma)$ of estimating f under a model M is defined as

$$r_n^2(\Sigma, Q, \sigma) = \inf_{\hat{f} \in \mathcal{A}_n} \sup_{f \in \Sigma} E_{f,Q} \, \|\hat{f} - f\|_Q^2$$

where $\Sigma \subset L_2(Q)$ is the function space specified by the model M, $\mathcal{A}_n$ is the space of all measurable functions of data to $L_2(Q)$ and $E_{f,Q}$ denotes expectation under the

model: $X_i \sim Q$, $Y_i|X_i \sim N(f(X_i), \sigma^2)$, independently across $i = 1, \cdots, n$. When no risk of ambiguity is present, we will shorten $r_n(\Sigma, Q, \sigma)$ to simply r_n and call r_n the minimax rate.

We will focus on function spaces characterized by smoothness conditions in addition to sparsity properties specified by models M2 and M3. Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any d dimensional multi-index $a = (a_1, \ldots, a_d) \in \mathbb{N}_0^d$ define $|a| = a_1 + \cdots + a_d$ and let D^a denote the mixed partial derivative operator $\partial^{|a|}/\partial x_1^{a_1} \cdots \partial x_d^{a_d}$. For any real number b let $\lfloor b \rfloor$ denote the largest integer strictly smaller than b. The Hölder class $\Sigma(\alpha, L, d)$ indexed by the triplet (α, L, d), is defined as the set of all d-variate $l = \lfloor \alpha \rfloor$ times differentiable functions f on $[-1, 1]^d$ such that:

$$||f||_{C^\alpha} = \max_{|a|=l} \sup_{x,y \in [-1,1]^d, x \neq y} \frac{|D^a f(x) - D^a f(y)|}{|x - y|^{\alpha - l}} \leqslant L. \tag{2.1}$$

A d-variate function f will be loosely referred to as an α-smooth function if it belongs to $\Sigma(\alpha, L, d)$ for some $L < \infty$.

We encode sparsity in a p-dimensional space through binary inclusion vectors $b \in \{0, 1\}^p$ and for any $x = (x^1, \ldots, x^p) \in \mathbb{R}^p$, let $x^b = (x^j : b_j = 1)$ denote the vector of $|b| = \sum_{j=1}^p b_j$ predictors picked by b. For M2, we will focus on "sparse" function spaces indexed by $\alpha, L > 0$, $d, p \in \mathbb{N}$ defined as:

$$\Sigma_S(\alpha, L, d, p) = \{x \mapsto g(x^b) : g \in \Sigma(\alpha, L, |b|) \text{ with } b \in \{0, 1\}^p \text{ and } |b| \leqslant d\}.$$

Without loss of generality, we assume that each element f in $\Sigma_S(\alpha, L, d, p)$ has zero mean with respect to Q, i.e. $\int f(x) Q(dx) = 0$, since otherwise we can always subtract the mean from f without changing its smoothness.

For M3, we will consider "additive" function spaces indexed by $\alpha, L \in (0, \infty)^k$,

$d \in \mathbb{N}^k$, $k, p, \bar{d} \in \mathbb{N}$ defined as:

$$\begin{aligned}\Sigma_A\big(\alpha, L, d, k, p, \bar{d}\big) =& \{x \mapsto f_1(x^{b_1}) + \cdots + f_k(x^{b_k}) : f_s \in \Sigma(\alpha_s, L_s, |b_s|),\\ & b_s \in \{0,1\}^p, |b_s| \leqslant d_s, b_s \not\subset b_t, b_{1j} + \cdots + b_{kj} \leqslant \bar{d},\\ & \text{for } s, t = 1, \ldots, k; s \neq t; j = 1, \ldots, p\},\end{aligned}$$

i.e., the elements of $\Sigma_A(\alpha, L, d, k, p, \bar{d})$ decompose into k irreducible components with a cap d_s on the interaction order of component s. Also, each predictor is restricted to appear in at most $\bar{d}$ many of the k components. Again, we will assume without loss of generality that each component function f_s is zero mean with respect to Q. Under this assumption, $\langle f_s, f_t \rangle_Q \neq 0$ if and only if f_s and f_t share common important predictors, i.e., $\sum_{j=1}^p b_{sj} b_{tj} \neq 0$. Consequently for each s, there are at most $d_s(\bar{d} - 1)$ indices $t \neq s$ such that $\langle f_s, f_t \rangle_Q \neq 0$, and hence

$$\begin{aligned}||f||_Q^2 &= \sum_{s=1}^k ||f_s||_Q^2 + \sum_{s=1}^k \sum_{t \neq s} \langle f_s, f_t \rangle_Q\\ &\leqslant \sum_{s=1}^k ||f_s||_Q^2 + \frac{1}{2} \sum_{s=1}^k \sum_{t: t \neq s, \langle f_s, f_t \rangle \neq 0} (||f_s||_Q^2 + ||f_t||_Q^2)\\ &\leqslant \sum_{s=1}^k ||f_s||_Q^2 + \sum_{s=1}^k d_s(\bar{d} - 1)||f_s||_Q^2\\ &\leqslant \{1 + d_{\max}(\bar{d} - 1)\} \sum_{s=1}^k ||f_s||_Q^2, \end{aligned} \tag{2.2}$$

where $d_{\max} = \max(d_1, \ldots, d_k)$. This inequality plays a key role in calculating covering entropies of the function spaces $\Sigma_A(\alpha, L, d, k, p, \bar{d})$. These entropy numbers behave well even when p and k are arbitrarily large, as long as $d_{\max}$ and $\bar{d}$ remain small.

The covering number $N(\epsilon, \Sigma, \rho)$ of a function space Σ equipped with a metric ρ is defined as the minimal number of ρ-balls of radius ϵ needed to cover Σ. It is customary to call $\log N(\epsilon, \Sigma, \rho)$ the ϵ covering entropy of Σ under ρ. A related

notion is the packing number of Σ under ρ, which is defined as the maximal number of ϵ separated elements in Σ. For linear space Σ, we introduce a new concept, the modified packing number $C(\epsilon, K, \Sigma, \rho)$ defined as the maximal number of elements of Σ that are ϵ distance apart from each other and each have norm smaller than $K\epsilon$. By $A(x) \sim B(x)$ for two functions $A(x)$ and $B(x)$, we mean $0 < \lim A(x)/B(x) < \infty$, where the limit is either $x \to 0$ or $x \to \infty$ determined by the specific context.

2.3 Minimax results for large-p small-n nonparametric regression

2.3.1 A brief overview of existing results

The minimax risk under M1 is well known (Bickel and Li, 2007; Ye and Zhou, 2008; Yang and Dunson, 2013). Bickel and Li (2007) show that multivariate local polynomial regression can adapt to the lower dimensional structure in the sense that it achieves the minimax rate $n^{-\alpha/(2\alpha+d)}$ when f is known to be α-smooth and $\alpha \leqslant 2$. Yang and Dunson (2013) consider Bayesian nonparametric regression with Gaussian process priors and prove that under M1, Gaussian process priors can achieve the minimax rate $n^{-\alpha/(2\alpha+d)}$ up to some log factor with additional adaptation to an unknown α that does not exceed 2.

However under M2 and M3, precise calculations of r_n and theoretical results on which estimation methods attain the minimax rates are known only under additional simplifying assumptions on the shape of f, or, for inference tasks that are simpler than prediction. In the linear model setup, Raskutti et al. (2011) show that with Σ taken as the set of functions $f(x) = x^T\beta$ with β in an l_0 ball of $\mathbb{R}^p$ and under some regularity conditions on the design matrix,

$$r_n^2 \asymp \frac{d \log(p/d)}{n}$$

up to some multiplicative constant, where d is the number of important predictors. As we will see later, this is the typical minimax risk associated with variable selection

uncertainty. Note that for $q = 0$, the l_q norm precisely encodes the sparsity condition of M_2. Wainwright (2009a) and Wainwright (2009b) consider minimax lower bounds for support recovery. For a review on various types of minimax risks for high dimensional linear models, see Verzelen (2012). Many authors have also obtained near minimax optimal convergence rates of various methods for linear regression under the L_2 loss, such as Bickel et al. (2009), Candes and Tao (2007), Meinshausen and Yu (2009) and Zhang and Huang (2008).

As a non-linear, non-parametric generalization of their results, Raskutti et al. (2012) consider sparse additive models with univariate components, which is a special case of M3 with each $d_s = 1$ and with each f_s being α-smooth for a common $\alpha > 0$. For this model they show

$$r_n^2 \asymp k\delta_n^2 + \frac{k \log p}{n},$$

where k is the component number and $\delta_n = n^{-\frac{\alpha}{2\alpha+1}}$ – the minimax risk of estimating an α-smooth univariate function. The minimax risk in this case can be decomposed into two terms, where the first term is the sum of minimax risks of estimating each component and the second term is the variable selection uncertainty.

As indicated earlier, an entirely different generalization of the linear model is the fully sparse nonparametric regression model of M2. To the best of our knowledge, the only result in this context is Comminges and Dalalyan (2012), who analyze minimax risks of support recovery under the variable selection framework. They show that if $d \log(p/d)/n$ is lower bounded by some positive constant α_0, then for some constant $c > 0$,

$$\inf_{\hat{J}_n} \sup_{f \in \Sigma} P_f(\hat{J}_n \neq J_f) \geqslant c,$$

where $\hat{J}_n$ ranges over all variable selection estimators, i.e., measurable maps of data to the space of all subsets of $\{1, \ldots, p\}$, Σ is the space of all differentiable functions

that depend on only d many predictors and have squared integrable gradients, and $J_f \subset \{1, \ldots, p\}$ is the index set of truly important predictors associated with f. This result is the reason we call the term $d \log(p/d)/n$ the minimax risk associated with variable selection uncertainty. In fact, for large p, the numerator $d \log(p/d)$ in the second term is asymptotically of the same order of the log of $\binom{p}{d}$, the number of ways to select d important predictors from p covariates. Therefore, it is reasonable to expect that any estimation problem related to high dimensional variable selection should include a variable selection uncertainty term $d \log(p/d)/n$.

2.3.2 Results on minimax rates under M2 and M3

In this paper we provide sharp upper and lower bounds to the minimax L_2 prediction risk for both M2 and M3 under the following condition on the predictor distribution Q:

Assumption Q. $Q = Q_0^p$ where Q_0 is a probability measure on $[-1, 1]$ that admits a Lebesgue density q_0 satisfying: $\inf_{u \in [-1,1]} q_0(u) > 0$ and $\sup_{u \in [-1,1]} q_0(u) < \infty$.

The main condition we need is independence among the predictors. They do not necessarily need to be identically distributed, though that additional assumption keeps notations tidier. Also, the independence assumption is needed only for providing a sharp lower bound to the minimax rate, but is not needed either for calculating a sharp upper bound or for deriving the posterior convergence rates of the additive Gaussian process regression method.

Theorem 1 (Minimax risk for M2)**.** *Under Assumption Q*

$$r_n^2(\Sigma_S(\alpha, L, d, p), Q, \sigma) \asymp \left(\frac{n}{\sigma^2}\right)^{-\frac{2\alpha}{2\alpha+d}} + \frac{\sigma^2 d \log(p/d)}{n}. \tag{2.3}$$

As we can see, the minimax risk in Theorem 1 consists of two terms. The first term corresponds to the minimax risk for estimating a d-variate function f_0 with the

knowledge of which d covariates are the important predictors. To make this term meaningful, d should be smaller compared to $\log n$. The second term is incurred by variable selection uncertainty, which is consistent with the results of Comminges and Dalalyan (2012).

Theorem 2 (Minimax risk for M3)**.** *Under Assumption Q,*

$$r_n^2(\Sigma_A(\alpha, L, d, k, p, \tilde{d})) \asymp c(d_{\max}, \tilde{d}) \sum_{s=1}^{k} \left\{ \left(\frac{n}{\sigma^2}\right)^{-\frac{2\alpha_s}{2\alpha_s+d_s}} + \frac{\sigma^2 d_s \log(p/d_s)}{n} \right\} \tag{2.4}$$

where $c(d_{\max}, \tilde{d})$ is a number between $1/B$ and $\sqrt{B}$ with $B = 1 + d_{\max}(\bar{d} - 1)$.

Toward proving these results, we first provide several fundamental results on how to calculate such sharp bounds over a general nonparametric function space Σ. Lower bounds are derived by using well known information-theoretic arguments (Yang and Barron, 1999). For upper bounds, we establish existence of Bayesian estimators with desired risks. Our construction borrows from Bayesian posterior convergence theory (Ghosal et al., 2000). We specialize these results to the cases of M2 and M3. It is more difficult to calculate minimax risk bounds for M3 than for the univariate additive case of Raskutti et al. (2012) where different components are assumed to be from the same function space. In the univariate case, zero mean components depending on different predictors are always orthogonal under the inner product $\langle \cdot, \cdot \rangle_Q$. However, in the general additive case, different components can share the same predictors and break down the orthogonality.

2.4 General theorems on characterizing minimax risks

2.4.1 Upper bounds for minimax risks

Theorem 3. *If $(\epsilon_n : n = 1, 2, \ldots)$ satisfies $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$ and $n\epsilon_n^2 \geqslant \sigma^2 \log N(\epsilon_n, \Sigma, ||\cdot||_Q)$, then there exists a prior Π_n over Σ such that for any $f_0 \in \Sigma$,*

$$E_{f_0,Q}\Pi_n\big(f : ||f - f_0||_Q > M\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n\big) \to 0, \tag{2.5}$$

for some fixed $M > 0$. Furthermore, if $\hat{f}$ is defined as the maximizer of $g \mapsto \Pi_n\big(f : ||f - g||_Q \leqslant M\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n\big)$ then

$$P_{f_0,Q}\big(||\hat{f} - f_0||_Q \leqslant 2M\epsilon_n\big) \to 1, \quad \text{as } n \to \infty.$$

In Theorem 3, we use the subscript n to indicate the dependence of the sample size on the constructed prior Π_n. The quantity ϵ_n in this statement can be understood as the posterior convergence rate, which means that the posterior probability measure assigns almost all its mass to a sequence of $||\cdot||_Q$-balls in Σ whose radii shrink towards f_0 at rate ϵ_n.

Although Theorem 3 ensures the convergence of $||\hat{f} - f_0||_Q$ to zero in probability, it does not characterize the decay rate of the posterior probability of $\{f \in \Sigma : ||\hat{f} - f_0||_Q \leqslant M\epsilon_n\}$. This decay rate of the tail probability is important for estimating the $L_2(Q)$ risk $E||\hat{f} - f_0||_Q^2$. To control this tail probability, we need to constrain the complexity of Σ in terms of the uniform covering entropy, which is defined for any $\epsilon > 0$ by $\sup_R \log N(\epsilon, \Sigma, ||\cdot||_R)$, with R ranging over all probability distributions (or all discrete probability distributions) on the support of Σ.

Theorem 4 (Upper bounds for minimax risks II)**.** *If $(\epsilon_n : n = 1, 2, \ldots)$ satisfies $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$ and $n\epsilon_n^2 \geqslant \sigma^2 \sup_R \log N(\epsilon, \Sigma, ||\cdot||_R)$, then there exists a prior Π_n over Σ such that for any $f_0 \in \Sigma$,*

$$E_{f_0,Q}\big\{\Pi_n\big(f : ||f - f_0||_n > M\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n\big)\big\} = \exp(-Cn\epsilon_n^2), \tag{2.6}$$

for some fixed numbers M and C. Furthermore, if $\hat{f}$ is defined as either the posterior expectation of f or the maximizer of $g \mapsto \Pi_n\big(f : ||f - g||_n \leqslant M\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n\big)$ then

$$P_{f_0,Q}\big(||\hat{f} - f_0||_n \geqslant 2M\epsilon_n\big) \leqslant \exp(-Cn\epsilon_n^2).$$

Moreover, if Σ is uniformly bounded, then for some $D > 0$,

$$\max\big\{E_{f_0,Q}\big(||\hat{f} - f_0||_Q^2\big),\ E_{f_0,Q}\big(||\hat{f} - f_0||_n^2\big)\big\} \leqslant D\epsilon_n^2.$$

The assumption on the uniform covering entropy is not a strong one and is commonly used in many statistical problems involving function spaces, such as Koltchiskii and Pancheko (2005). In particular, the uniform covering entropies of the function spaces under M2 and M3 are finite for any $\epsilon > 0$ and have the same order as $\log N(\epsilon, \Sigma, ||\cdot||_Q)$. Therefore, Theorem 4 implies exponentially decay rate of the posterior probabilities of $\{f : ||f - f_0||_Q > M\epsilon_n\}$ for Σ_S and Σ_A.

2.4.2 Review of lower bounds for minimax risks

Theorem 5 (Lower bounds for minimax risks)**.** *Let ϵ_n to be a positive sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \leqslant (2K^2)^{-1}\sigma^2 \log C(2\epsilon_n, K, \Sigma, ||\cdot||_Q)$ for some $K > 0$, then*

$$\inf_{\hat{f} \in \mathcal{A}_n} \sup_{f \in \Sigma} P_{f,Q}\big\{||\hat{f} - f||_Q \geqslant \epsilon_n\big\} \geqslant \frac{1}{2}.$$

Therefore, the minimax risk under the $L_2(Q)$ loss satisfies $r_n^2(\Sigma, Q, \sigma) \geqslant \frac{1}{2}\epsilon_n^2$.

At a first sight, Theorem 3 and Theorem 5 seem to contradict each other since in the regular parametric models where Bernstein von-Mises theorem holds and $\epsilon_n \sim n^{-1/2}$, the posterior distribution of $\sqrt{n}||f - f_0||_Q$ is approximately normal and $\Pi_n(||f - f_0||_Q \geqslant M\epsilon) \asymp \exp(-CM^2) \nrightarrow 0$ for some $C > 0$ and any M. In fact, Theorem 3 only apply for nonparametric cases where the condition $n\epsilon_n^2 \to \infty$ rules out the parametric cases. Therefore, the results imply that when the minimax rate is slower than the parametric rate $n^{-1/2}$, there is a phase transition in the sense that for some critical value M_0, we have

$$\lim_{n\to\infty} \inf_{\hat{f}} \sup_{f \in \Sigma} P_{f,Q}\big\{||\hat{f} - f||_Q \geqslant M\epsilon_n\big\} \begin{cases} > 0, & M < M_0; \\ = 0, & M > M_0. \end{cases}$$

However, since our primary interest is in the asymptotic order of the minimax rate $r_n(\Sigma, Q, \sigma)$, we will not attempt to determine the exact multiplicative constant in it.

By Theorem 3 and Theorem 5, if we can obtain a tight lower bound $\log \hat{C}(\epsilon)$ to $\log C(\epsilon, K, \Sigma, ||\cdot||_Q)$ for some $K > 0$ and a tight upper bound $\log \hat{N}(\epsilon)$ to $\log N(\epsilon, \Sigma, ||\cdot$

$||_Q)$, such that $\log \hat{C}(\epsilon) \sim \log \hat{N}(\epsilon)$ as $\epsilon \to 0$, then $r_n(\Sigma, Q, \sigma)$ will be determined up to a multiplicative constant as the solution of the equation $\log \hat{N}(\epsilon) \sim n\epsilon^2$. With such an ϵ, the corresponding prior Π_n in Theorem 3 can be considered as asymptotically least favorable from a decision-theoretical point of view.

2.4.3 Auxiliary results for function spaces with additive structures

Consider a general framework where the additive function space takes the form $\mathcal{F} = \bigoplus_{s=1}^k \mathcal{F}_s = \{f = \sum_{s=1}^k f_s : f_s \in \mathcal{F}_s, s = 1, \ldots, k\}$ for k function spaces $\mathcal{F}_1, \ldots, \mathcal{F}_k$. In the sequel, K is a fixed constant and $\log C(\epsilon, K, \mathcal{F}, ||\cdot||_Q)$, $\log C(\epsilon, K, \mathcal{F}_s, ||\cdot||_Q)$, $\log N(\epsilon, \mathcal{F}, ||\cdot||_Q)$ and $\log N(\epsilon, \mathcal{F}, ||\cdot||_Q)$ will be abbreviated as $\log C(\epsilon, K)$, $\log C_s(\epsilon, K)$, $\log N(\epsilon)$ and $\log N_s(\epsilon)$.

Next, we study the minimax risks associated with $\mathcal{F}$. We make two assumptions:

F1. $||f||_Q^2 \leqslant B \sum_{s=1}^k ||f_s||_Q^2$, $\quad \forall f = \sum_{s=1}^k f_s \in \mathcal{F}$ for some constant $B > 0$;

F2. For any $\epsilon_1, \ldots, \epsilon_k > 0$, there exist mutually orthogonal modified ϵ_s-packing sets $\mathcal{E}_s(\epsilon_s)$ of size $C_s(\epsilon_s, K)$ for $s = 1, \ldots, k$, i.e. $\forall s \neq t$, $f_s \in \mathcal{E}_s(\epsilon_s)$ and $f_t \in \mathcal{E}_t(\epsilon_t)$, $\langle f_s, f_t \rangle_Q = 0$.

Under the near orthogonal condition F1, $||f - g||_Q^2$ can be bounded by a multiple of $\sum_{s=1}^k ||f_s - g_s||_Q^2$ for any two functions $f = \sum_{s=1}^k f_s$ and $g = \sum_{s=1}^k g_s$ in $\mathcal{F}$. This property plays a key role in obtaining an upper bound to the covering entropy of $\mathcal{F}$. F2 is important for constructing a sufficiently large packing set for $\mathcal{F}$. Σ_A is close to $\bigoplus_{s=1}^k \Sigma_S(\alpha_s, L_s, d_s)$ up to a negligible subset caused by the non-inclusive constraints on the additive components. Therefore, the results in this subsection on $\mathcal{F}$ can be easily transferred to Σ_A.

The following theorem provides lower and upper bounds to $\log C(\epsilon/2)$ and $\log N(K_1\epsilon)$ in terms of $\{\log C_s(\epsilon)\}$ and $\{\log N_s(\epsilon)\}$ under F1 and F2.

Theorem 6 (Entropies for additive spaces)**.** *Under assumption F1 and F2, for any* $\epsilon > 0$,

$$\log C\left(\frac{\epsilon}{2}, \sqrt{B}K\right) \geqslant K_0 \sum_{s=1}^{k} \log C_s(\bar{\alpha}_s \epsilon, K),$$

where $K_0 > 0$ *is some universal constant and* $(\bar{\alpha}_1, \ldots, \bar{\alpha}_k) \in \mathbb{R}_+^k$ *are the solution of*

$$\frac{\log C_1(\alpha_1 \epsilon, K)}{\alpha_1^2} = \cdots = \frac{\log C_k(\alpha_k \epsilon, K)}{\alpha_k^2} = \sum_{s=1}^{k} \log C_s(\alpha_s \epsilon, K). \tag{2.7}$$

Moreover, for any nonnegative vector $(\alpha_1, \ldots, \alpha_k)$ *satisfying* $\sum_{s=1}^k \alpha_s^2 = 1$,

$$\log N(\sqrt{B}\epsilon) \leqslant \sum_{s=1}^{k} \log N_s(\alpha_s \epsilon).$$

In particular, the above holds for the $(\bar{\alpha}_1, \ldots, \bar{\alpha}_k)$ *in* (2.7)*.*

If for each $\mathcal{F}_s$, we have a lower bound $\log \hat{C}_s(\epsilon)$ and upper bound $\log \hat{N}_s(\epsilon)$ to $\log C_s(\epsilon)$ and $\log N_s(\epsilon)$ so that for any fixed constant $a_1 > 0, a_2 > 0$, $\log \hat{C}_s(a_1\epsilon) \sim \log \hat{N}_s(a_2\epsilon)$ as $\epsilon \to 0$ then by Theorem 6, we can obtain lower and upper bounds for $\log C(\epsilon)$ and $\log N(\epsilon)$ respectively so that $\log \hat{C}(a_1\epsilon) \sim \log \hat{N}(a_2\epsilon)$ as $\epsilon \to 0$. Combining this observation with Theorem 3 and Theorem 5, we have the following corollary on minimax risks of $\mathcal{F}$.

Corollary 7 (Minimax risks for additive spaces)**.** *Under assumptions F1 and F2, the minimax risk of estimating a function* $f \in \mathcal{F} = \bigoplus_{s=1}^k \mathcal{F}_k$ *is* $\epsilon_n^2 = \sum_{s=1}^k \delta_{n,s}^2$*, where* $\delta_{n,s}$ *is the solution of* $\log \hat{N}_s(\delta_s^2) \sim n\delta_s^2$ *for* $s = 1, \ldots, k$*.*

2.5 Applications of the general results to M2 and M3

In this section, we provides tight lower/upper bounds for the modified packing entropies and covering entropies of Σ_S and Σ_A. Then with the help of Theorem 5 and Theorem 3, we can obtain the minimax risks of Σ_S and Σ_A.

2.5.1 Minimax lower bounds for high dimensional regression

In this subsection, we study modified packing entropies of the relevant sparse regression spaces. With the help of Theorem 5, lower bounds on these quantities provide lower bounds for the minimax risks.

Lemma 8 (Modified packing entropy lower bounds)**.** *Assume assumption Q. Then for $\epsilon > 0$, Σ and $N > 0$ in any of the following cases:*

1. $\Sigma = \Sigma(\alpha, L, d)$ *and* $\log N \geqslant K_1(L/\epsilon)^{d/\alpha}$*;*
2. $\Sigma = \Sigma_S(\alpha, L, d, p)$ *and* $\log N \geqslant K_1(L/\epsilon)^{d/\alpha} + d\log(p/d)$*;*
3. $\Sigma = \Sigma_A\big((\alpha_1, \ldots, \alpha_k),\ (L_1, \ldots, L_k),\ (d_1, \ldots, d_k), p, \bar{d}\big)$ *and* $\log N \geqslant K_1 \sum_{s=1}^k (L_s/(\alpha_s\epsilon))^{\beta_s} + K_1 \sum_{s=1}^k d_s \log(p/d_s)$*, for some* $K_1 > 0$*, where* $\beta_s = d_s/\alpha_s$ *and* $(\alpha_1, \ldots, \alpha_k)$ *solves*

$$\sum_{s=1}^k \alpha_s^2 = 1, \quad \left(\frac{L_1}{\epsilon}\right)^{\beta_1} \frac{1}{\alpha_1^{2+\beta_1}} = \cdots = \left(\frac{L_k}{\epsilon}\right)^{\beta_k} \frac{1}{\alpha_k^{2+\beta_k}}, \tag{2.8}$$

there exist $N+1$ functions $\{f_s\}_{s=0}^N \subset \Sigma$ such that

(i). $f_0 = 0,\ ||f_s||_Q \leqslant K_2\epsilon,\ 1 \leqslant s \leqslant N,$

(ii). $d(f_s, f_t) \geqslant \epsilon,\ 0 \leqslant s < t \leqslant N,$

for some $K_2 > 0$ independent of ϵ and L or $\{L_s\}$. This implies

$$\log C\big(\epsilon, K_2, \Sigma(\alpha, L, d), ||\cdot||_Q\big) \geqslant K_1\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}},$$

$$\log C\big(\epsilon, K_2, \Sigma_S(\alpha, L, d, p), ||\cdot||_Q\big) \geqslant K_1\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}} + d\log\frac{p}{d},$$

$$\log C\big(\epsilon/2, \sqrt{B}K_2, \Sigma_A\big((\alpha_1, \ldots, \alpha_k), (L_1, \ldots, L_k), (d_1, \ldots, d_k), p, \bar{d}\big), ||\cdot||_Q\big)$$

$$\geqslant K_1 \sum_{s=1}^k \left(\frac{L_s}{\alpha_s\epsilon}\right)^{\beta_s} + K_1 \sum_{s=1}^k d_s \log\frac{p}{d_s},$$

for $B = 1 + d_{max}(\bar{d} - 1)$.

The above lemma indicates that the "size" of $\Sigma_S(\alpha, L, d)$ is characterized by $\beta = d/\alpha$, which will be referred to as the complexity index. To appreciate the above modified packing entropy lower bound for the additive function space Σ_A, we consider two special cases. In the first case, all additive components are univariate with the same smoothness α and magnitude L. The same framework is considered in Raskutti et al. (2012). In this case, $\alpha_1 = \cdots = \alpha_k = k^{-1/2}$ and the lower bound for the modified packing entropy becomes $K_1 k(\sqrt{k}L/\epsilon)^{1/\alpha} + K_1 k \log p$. By Theorem 5, this provides a lower bound to the minimax risk as $\epsilon_n^2 \sim k n^{-\frac{2\alpha}{2\alpha+1}} + k \log p/n$, which is the same as the minimax risk obtained in Raskutti et al. (2012) when the univariate additive function spaces are α-smooth Hölder classes.

In the second case, assume k to be fixed and one additive component to be much more complex than the rest, i.e. $\beta_1 = d_1/\alpha_1 \gg \beta_s = d_s/\alpha_s$ for $s = 2, \ldots, k$. In this case, $\alpha_1 \approx 1$ and $(\alpha_s \epsilon)^{-\beta_s} \ll \epsilon^{-\beta_1}$ for $s > 1$. As a result, the lower bound to the modified packing entropy is dominated by the first component as $K_1(L_1/\epsilon)^{\beta_1} + K_1 \sum_{s=1}^k d_s \log(p/d_s)$. As a result, the lower bound for the minimax risk becomes $\epsilon_n^2 \sim n^{-\frac{2}{\beta_1+2}} + \sum_{s=1}^k d_s \log(p/d_s)/n$, in which the first term is dominated by the slowest convergence rate of the additive components, while the second term is still determined by the overall variable selection uncertainty.

2.5.2 Minimax upper bounds for high dimensional regression

In this subsection, we study the covering entropies, which provide upper bounds for the corresponding minimax risks by Theorem 5. In the proofs, the distribution Q is not necessarily the common marginal distribution of the components of X, but can be any distribution on $[-1, 1]^p$.

Birman and Solomjak (1967) provide an upper bound for the covering entropy of $\Sigma(\alpha, L, d)$ under sup norm, which is of the same order as the lower bound for the modified packing entropy obtained in Theorem 8. Since $||\cdot||_Q$ is dominated by $||\cdot||_\infty$, their result also provides an upper bound for the covering entropy of $\Sigma(\alpha, L, d)$ under the $||\cdot||_Q$ norm. Based on this, we can obtain upper bounds for the covering entropy of $\Sigma_S(\alpha, L, d, p)$ and $\Sigma_A\big((\alpha_1, \ldots, \alpha_k), (L_1, \ldots, L_k), (d_1, \ldots, d_k), p, \bar{d}\big)$ as the following lemma shows.

Lemma 9 (Covering entropy upper bounds)**.** *For any $\epsilon > 0$, we have*

$$\log N\big(\epsilon, \Sigma(\alpha, L, d), ||\cdot||_Q\big) \leqslant K\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}},$$

$$\log N\big(\epsilon, \Sigma_S(\alpha, L, d, p), ||\cdot||_Q\big) \leqslant K\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}} + d\log\frac{p}{d},$$

$$\log N\Big(\sqrt{B}\epsilon, \Sigma_A\big((\alpha_1, \ldots, \alpha_k),\ (L_1, \ldots, L_k), (d_1, \ldots, d_k), p, \bar{d}\big), ||\cdot||_Q\Big)$$

$$\leqslant K\sum_{s=1}^{k}\left(\frac{L_s}{\alpha_s\epsilon}\right)^{\beta_s} + \sum_{s=1}^{k} d_s\log\frac{p}{d_s},$$

where K is a positive constant independent of ϵ and L or $\{L_s\}$ and $(\alpha_1, \ldots, \alpha_k)$ solves (2.8)*.*

Similar to Lemma 8, as long as B remains small, the lower bounds for the modified packing entropies and minimax risks are also upper bounds up to multiplicative constants, i.e. these bounds are sharp. In addition, since the upper bounds in Lemma 9 do not depend on Q, they also serve as upper bounds to the uniform covering entropies defined before Theorem 4.

2.6 Adaptive near minimax optimality of Bayesian additive Gaussian process regression

Although the Bayesian estimator constructed in the proof of Theorem 3 attains the minimax rate, it is essentially a mathematical construct and its practical implementation is nearly infeasible. Also, it requires the use of a different prior distribution for different sample sizes, which may not be attractive in practice. In this section, we demonstrate the existence of practical Bayesian methods based on single prior distributions that adapt automatically across various function spaces satisfying M2 and M3.

Gaussian process (GP) priors are widely used in nonparametric regression. Adaptivity and near minimax optimality of Bayesian GP regression methods are known for low dimensional applications (van der Vaart and van Zanten, 2009). We investigate extensions of these methods to sparse high dimensional settings. We show that with appropriate point mass mixture priors for Bayesian variable selection, GP priors are still guaranteed to attain the minimax rates up to some log factors.

2.6.1 GP and its adaptive rate optimality for fixed p

We briefly review the theory developed by van der Vaart and van Zanten (2009) on adaptive posterior contraction rate of Gaussian Process (GP) priors. Consider a GP $W = (W_x; x \in [-1,1]^d)$ on $[-1,1]^d$. The law $GP(m, K)$ of W is completely determined by its mean function $m(x) = EW_x$ and covariance function $K(x, x') = E(W_x - m(x))(W_{x'} - m(x'))$. We consider a zero mean and stationary GP, where the covariance function $K(x, x') = EW_xW_{x'}$ only depends on $x - x'$. The square exponential kernel $\exp(-||x - x'||^2)$ is a common choice for $K(x, x')$. By Bochner's theorem,

$$K(x, x') = \int e^{-i(\lambda, x-x')} d\mu(\lambda),$$

where the finite Borel measure μ on R^d is called the spectral measure of W. van der Vaart and van Zanten (2009) focus on GPs whose spectral measure has exponential tails: for some $\delta > 0$,

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

van der Vaart and van Zanten (2008a) propose a set of conditions that ensure the posterior convergence rate of GP priors for estimating the function $f_0 \in C[-1,1]^d$ in the regression problem $Y|X \sim N(f_0(X), \sigma^2)$ to be at least ϵ_n as:

$$P(||W - f_0||_\infty \leqslant \tilde{\epsilon}_n) \geqslant e^{-n\tilde{\epsilon}_n^2}, \tag{2.9}$$

$$P(W \notin \mathcal{B}_n) \leqslant e^{-4n\tilde{\epsilon}_n^2}, \tag{2.10}$$

$$\log N(\epsilon_n, \mathcal{B}_n, ||\cdot||_\infty) \leqslant n\epsilon_n^2, \tag{2.11}$$

where $(\mathcal{B}_n : n \geqslant 1)$ is a sequence of subsets of $C[-1,1]^d$, called sieves and $(\tilde{\epsilon}_n : n \geqslant 1)$ is a sequence satisfying $\tilde{\epsilon}_n < \epsilon_n$, $\lim_{n\to\infty} n\tilde{\epsilon}_n^2 = \infty$.

van der Vaart and van Zanten (2008a) show that the prior concentration condition (2.9) is intimately connected with the concentration function $\phi_{f_0}(\epsilon)$ since $P(||W - f_0||_\infty \leqslant \epsilon_n) \geqslant e^{-\phi_{f_0}(\epsilon)}$, where the concentration function is defined as the sum of two terms:

$$\phi_{f_0}(\epsilon) = \inf_{h \in \mathbb{H}: ||h - f_0||_\infty \leqslant \epsilon} ||h||_{\mathbb{H}}^2 - \log P(||W||_\infty \leqslant \epsilon),$$

where $(\mathbb{H}, ||\cdot||_{\mathbb{H}})$ is the reproducing kernel Hilbert space (RKHS) associated with the GP W. The first term measures how well f_0 is approximated by the elements in $\mathbb{H}$. The second term, the so-called small ball probability, characterizes the probability mass of W assigned to a ϵ ball around f_0. An upper bound for the small ball probability can be directly obtained by the condition (2.11) (Lemma 4.6 in van der Vaart and van Zanten (2009)).

To achieve adaptation to unknown smoothness, van der Vaart and van Zanten (2009) propose to rescale W by a random length scale parameter A as $W_A = (W_{Ax} :$

$x \in [-1,1]^d$), where A^d follows a gamma distribution $\mathrm{Ga}(a_1, a_2)$ with scale parameter a_1 and rate parameter a_2. For $f_0 \in \Sigma(\alpha, L, d)$, Stone (1982) shows that the minimax rate of estimating f_0 is $L^{d/(2\alpha+d)} n^{-\alpha/(2\alpha+d)}$ (which is also implied by Lemma 8, Lemma 9, Theorem 3 and Theorem 5), where n is the sample size. van der Vaart and van Zanten (2009) prove that by introducing A, the posterior distribution of W_A can achieve the minimax rate up to some logarithm factors. Hereafter, we use either a superscript or a subscript $a(\mathrm{A})$ to indicate the dependence on the (random) length scale. For example, we write the covariance function of W_a by $K_a(x, x')$.

To verify condition (2.9), van der Vaart and van Zanten (2008a) show that for sufficiently large n

$$P(||W_A - f_0||_\infty \leqslant \rho_n) \geqslant e^{-n\rho_n^2}, \tag{2.12}$$

for ρ_n a large multiple of $L^{d/(2\alpha+d)} n^{-\alpha/(2\alpha+d)} (\log n)^{(1+d)/(2+d/\alpha)}$. To satisfy condition (2.10) and (2.11), they construct a sequence of sieves taking some specific forms. The following lemma summarizes their constructions. Since the results in this lemma play a key role in our later proofs, we provide an outline of a proof extracted from van der Vaart and van Zanten (2009) for completeness.

Lemma 10. *For positive constants M, r, ϵ, δ, let*

$$\mathcal{B}_{M,r,\epsilon,\delta} = \left\{(r/\delta)^{d/2} M \mathbb{H}_1^r + \epsilon \mathbb{B}_1\right\} \cup \left\{ \bigcup_{a<\delta} M \mathbb{H}_1^a + \epsilon \mathbb{B}_1 \right\}, \tag{2.13}$$

where $\mathbb{H}_1^r$ is the unit ball of the RKHS $\mathbb{H}^r$ associated with W_r and $\mathbb{B}_1$ is the unit ball of $C[-1,1]^d$ in sup-norm. Suppose that the prior density g for A satisfies

$$B_1 a^p \exp(-D_1 a^d) \leqslant g(a) \leqslant B_2 a^p \exp(-D_2 a^d),$$

for some B_1, B_2, D_1, D_2 and $p > 0$, which is true when A^d follows a $Ga(p+1, D)$ prior. Then there exist some universal positive constants C_0, C_1, C_2, C_3, C_4, $a_0 > 1$, $\epsilon_0 < 1/2$, such that for every $r > a_0, 0 < \epsilon_0, M^2 > C_0 r^d (\log(r/\epsilon))^{1+d}$ and $\delta =$

$\epsilon/(2d^{3/2}M)$, *the following inequalities hold:*

$$P(W_A \notin \mathcal{B}_{M,r,\epsilon,\delta}) \leqslant C_1 r^{p-d+1} e^{-C_2 r^d} + e^{-M^2/8}, \tag{2.14}$$

$$\log N(3\epsilon, \mathcal{B}_{M,r,\epsilon,\delta}, ||\cdot||_\infty) \leqslant C_3 r^d \left(\log \frac{M^{3/2}\sqrt{2d^{3/2}r}}{\epsilon^{3/2}} \right)^{1+d} + 2\log \frac{C_4 M}{\epsilon}. \tag{2.15}$$

As a result, for an arbitrary sequence $(\bar{\epsilon}_n : n \geqslant 1)$ satisfying $\lim_{n\to\infty} \bar{\epsilon}_n = 0$ and $\lim_{n\to\infty} n\bar{\epsilon}_n^2 = \infty$, the sequence of sieves $(\mathcal{B}_n = \mathcal{B}_{M_n,r_n,\epsilon_n,\delta_n} : n \geqslant 1)$ with r_n^d a large multiple of $n\bar{\epsilon}_n^2$, M_n^2 a large multiple of $n\bar{\epsilon}_n^2(\log n)^{1+d}$ and $\delta_n = \bar{\epsilon}_n/(2|b|^{3/2}M_n$ satisfy the following inequalities: for some universal positive constants C_4, C_5, L,

$$\begin{aligned} P(W_A \notin \mathcal{B}_n) &\leqslant e^{-C_4 n\bar{\epsilon}_n^2}, \\ \log N(L\bar{\epsilon}_n, \mathcal{B}_n, ||\cdot||_\infty) &\leqslant C_5 n\bar{\epsilon}_n^2(\log n)^{1+d}. \end{aligned} \tag{2.16}$$

With the special choice of $\bar{\epsilon}_n = \delta_n$, van der Vaart and van Zanten (2009) prove that GP priors with random length scales can achieve posterior contraction rate at least $\epsilon_n = \rho_n(\log n)^{(1+d)/2}$, which is a large multiple of $L^{d/(2\alpha+d)}n^{-\alpha/(2\alpha+d)}(\log n)^\gamma$ with $\gamma = (1+d)/(2+d/\alpha) + (1+d)/2$. We would like to emphasize the flexibility of the choice of $\bar{\epsilon}_n$ in (2.16), since it is crucial in the later construction of sieves in the proof for the adaptive property in terms of variable selection.

2.6.2 GP with high dimensional variable selection

In this subsection, we consider the estimation of f under M2. We extend the GP prior to include variable selection. Let $B \in \{0,1\}^p$ represent a random inclusion vector and b_0 be the inclusion vector corresponding to f_0 that generates the data $Y_i|X_i \sim N(f_0(X_i), \sigma^2)$, $i = 1, \ldots, n$. Use the notation $W_a^b = \left(W_{x^b}^b : x^b \in [0,1]^{|b|}\right)$ to denote the GP with covariance function $K_a^b(x^b, x'^b)$.

We consider the following GP variable selection (GPVS) prior to model the un-

known function, denoted by W:

$$\begin{aligned} &P(B=b)\propto p^{-|b|}(1-p^{-1})^{p-|b|}I(|b|\leqslant d_0),\\ &A^{|B|}|B\sim \mathrm{Ga}(a_1,a_2),\\ &W=W_A^B|A,B\sim GP(0,K_A^B), \end{aligned} \tag{2.17}$$

where d_0 is a prespecified hyperparameter, interpreted as the prior belief on the maximum number of important predictors. The following provides a posterior contraction rate ϵ_n of this prior.

Theorem 11. *Assume $f_0 \in \Sigma(\alpha_0, L_0, d_0)$. If $p \to \infty$ as $n \to \infty$ and $d_0 \geqslant |b_0|$, then the posterior contraction rate ϵ_n of the GPVS prior is at least*

$$L_0^{c_0/(2+c_0)} n^{-1/(2+c_0)} (\log n)^{\beta_1} + \sqrt{\frac{d_0 \log p}{n}} (\log n)^{\beta_2},$$

where $c_0 = |b_0|/\alpha_0$ and $\beta_1 = (1+|b_0|)/(2+c_0) + (1+d_0)/2$ and $\beta_2 = (1+d_0)/2$.

By Theorem 11, the contraction rate is adaptive to the unknown smoothness α_0 and number of important predictors d_0, and almost attains the minimax rate indicated by Theorem 1. The first part in the rate $n^{-\alpha_0/(2\alpha_0+|b_0|)}\ (\log n)^{\beta_1}$ does not involve the dimensionality p of the covariates and corresponds to the minimax rate $n^{-\alpha_0/(2\alpha_0+|b_0|)}$ of estimating a $|d_0|$ variate function up to a logarithmic factor as if we knew the important predictors. However, for this result to hold, we require $d_0 \geqslant |b_0|$. Since we do not know $|b_0|$ in advance, ideally we need to specify d_0 large enough to cover $|b_0|$. We can allow d_0 to slowly grow with n such that the logarithmic factor is still asymptotically smaller compared to n^λ for any $\lambda > 0$. For example, $d_0 = (\log n)^\kappa$, where $0 < \kappa < 1$. In the second part, since we do not know $|b_0|$ but only an upper bound d_0, we have $d_0 \log p/n$ instead of $|b_0| \log p/n$ in the variable selection uncertainty error.

2.6.3 Additive GP with high dimensional variable selection

In this subsection, we consider the regression problem under the assumption M3. Suppose that the true function f_0 has an additive form:

$$f_0(x) = \sum_{h=1}^{k_0} f_{0,h}(x^{b_{0,h}}), \tag{2.18}$$

where $b_{0,h}$ is the inclusion vector for the h-th component. Assume the Hölder smoothness of the $|b_{0,h}|$ variate function $f_{0,h}$ is α_h and its magnitude is L_h. Under such assumptions, $f_0 \in \Sigma_A\big((\alpha_{0,1}, \ldots, \alpha_{0,k_0}), (L_{0,1}, \ldots, L_{0,k_0}), (d_{0,1}, \ldots, d_{0,k_0}), p, \bar{d}\big)$. Since the number k_0 of components is unknown, we introduce a prior for the random component number $K \in \{1, \ldots, K_0\}$, where K_0 is a sufficiently large but fixed number. Conditioning on K, each component can be specified by the GPVS prior (2.17). Denote $b_h(B_h)$ and $a_h(A_h)$ the (random) inclusion vector and (random) length scale for the h-th component. As a result, the additive GP variable selection (AGPVS) prior for the random additive function W has the following hierarchical model: for $\sum_{k=0}^{K_0} p_k = 1$, $p_k > 0, k = 0, 1, \ldots, K_0$,

$$\begin{aligned}
&P(K = k) = p_k, \text{ for } k = 0, 1, \ldots, K_0,\\
&P(B_h = b_h) \propto p^{-|b_h|}(1 - p^{-1})^{p-|b_h|} I(|b_h| \leqslant d_0), \text{ for } h \leqslant K,\\
&A_h^{|B_h|} | B_h \sim \mathrm{Ga}(a_1, a_2), \text{ for } h \leqslant K,\\
&W_h^{B_h} | A_h, B_h \sim GP(0, K_{A_h}^{B_h}), \text{ for } h \leqslant K,\\
&W = \sum_{h=1}^{K} W_h^{B_h}.
\end{aligned}$$

The posterior contraction rate of the AGPVS prior is provided by the following theorem:

Theorem 12. *Assume that* $f_0 \in \Sigma_A\big((\alpha_{0,1}, \ldots, \alpha_{0,k_0}), (L_{0,1}, \ldots, L_{0,k_0}), (d_{0,1}, \ldots, d_{0,k_0}), p, \bar{d}\big)$. *If* $p \to \infty$ *as* $n \to \infty$, $d_0 \geqslant \max_{1 \leqslant h \leqslant k_0} |d_{0_h}|$ *and* $k_0 \leqslant K_0$, *then the posterior*

contraction rate ϵ_n of the AGPVS prior is at least

$$\sqrt{K_0}\Big(\sum_{h=1}^{k_0} L_{0,h}^{c_{0,h}/(2+c_{0,h})} n^{-1/(2+c_{0,h})} (\log n)^{\beta_{1,h}} + \sqrt{\frac{K_0 d_0 \log p}{n}} (\log n)^{\beta_2}\Big),$$

where $c_{0,h} = |b_{0,h}|/\alpha_{0,h}$, $\beta_{1,h} = (1 + |b_{0,h}|)/(2 + c_{0,h}) + (1 + d_0)/2$ *and* $\beta_2 = (1 + d_0)/2$.

In practice, in order to accommodate the unknown number k_0 of components, which is assumed to be fixed, we can allow K_0 to slowly grow with the sample size n in a slow rate and still attain a near optimal rate. For example, if K_0 is of order $O((\log n)^\gamma)$ for some $\gamma > 0$, then the convergence rate only differs from the minimax rate up to a logarithmic factor. Again, since we only know upper bounds d_0 and K_0 for $d_{0,h}$ and k_0 respectively, we have $K_0 d_0 \log p/n$ instead of $\sum_{h=1}^{k_0} |b_{0,h}| \log p/n$ in the variable selection uncertainty error.

3

Nonparametric regression on manifolds

3.1 Introduction

Dimensionality reduction in nonparametric regression is of increasing interest given the routine collection of high-dimensional predictors in many application areas. In particular, our primary focus is on the regression model

$$Y_i = f(X_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{3.1}$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^D$, f is an unknown regression function, and ϵ_i is a residual having variance σ^2. We face problems in estimating f accurately due to the moderate to large number of predictors D. Fortunately, in many applications, the predictors have support that is concentrated near a d-dimensional subspace $\mathcal{M}$. If one can learn the mapping from the ambient space to this subspace, the dimensionality of the regression function can be reduced massively from D to d, so that f can be much more accurately estimated.

There is an increasingly vast literature on the topic of subspace learning, but there remains a lack of approaches that allow flexible non-linear dimensionality reduction, are scalable computationally to moderate to large D, have theoretical guarantees

and provide a realistic characterization of uncertainty. Regarding this last point, we would like to be able to characterize uncertainty in estimating the regression function f, in functionals of f of interest and in predictions. Typical two-stage approaches in which one conducts dimensionality reduction in a first stage, and then plugs the d-dimensional features into a next stage regression may provide a point estimate with good properties but do not characterize uncertainty in this estimate.

With this motivation, we focus on Bayesian nonparametric regression methods that allow $\mathcal{M}$ to be an unknown Riemannian manifold. One natural direction is to choose a prior to allow uncertainty in $\mathcal{M}$, while also placing priors on the mapping from x_i to $\mathcal{M}$, the regression function relating the lower-dimensional features to the response, and the residual variance. Some related attempts have been made in the literature. Tokdar et al. (2010) propose a logistic Gaussian process model, which allows the conditional response density $f(y|x)$ to be unknown and changing flexibly with x, while reducing dimension through projection to a linear subspace. Their approach is elegant and theoretically grounded, but does not scale efficiently as D increases and is limited by the linear subspace assumption. Also making the linear subspace assumption, Reich et al. (2011) proposed a Bayesian finite mixture model for sufficient dimension reduction. Page et al. (2013) instead propose a method for Bayesian nonparametric learning of an affine subspace motivated by classification problems.

There is also a limited literature on Bayesian nonlinear dimensionality reduction. Gaussian process latent variable models (GP-LVMs) (Lawrence, 2003) were introduced as a nonlinear alternative to PCA for visualization of high-dimensional data. Kundu and Dunson (2011) proposed a related approach that defines separate Gaussian process regression models for the response and each predictor, with these models incorporating shared latent variables to induce dependence. The latent variables can be viewed as coordinates on a lower dimensional manifold, but daunting problems

arise in attempting to learn the number of latent variables, the distribution of the latent variables, and the individual mapping functions while maintaining identifiability restrictions. Chen et al. (2010) instead approximate the manifold through patching together hyperplanes. Such mixtures of linear subspace-based methods may require a large number of subspaces to obtain an accurate approximation even when d is small.

It is clear that probabilistic models for learning the manifold face daunting statistical and computational hurdles. In this article, we take a very different approach in attempting to define a simple and computationally tractable model, which bypasses the need to estimate $\mathcal{M}$ but can exploit the lower-dimensional manifold structure when it exists. In particular, our goal is to define an approach that obtains a minimax-optimal adaptive rate in estimating f, with the rate adaptive to the manifold and smoothness of the regression function. Surprisingly, we show that this can be achieved with a simple Gaussian process prior.

Section 3.2 defines the proposed model and gives some basic geometric background along with a heuristic motivation for the model. Section 3.3 contains simulation studies of finite sample performance relative to competitors, and Section 7 discusses the results. Appendix B.1 contains a more thorough background of necessary geometric concepts. Appendix B.2 provides the technical proofs.

3.2 Gaussian processes on manifolds

3.2.1 Background

Gaussian processes (GP) are widely used as prior distributions for unknown functions due to tractable posterior computation and strong theoretical guarantees. For example, in the nonparametric regression (3.1), a GP can be specified as a prior for the unknown function f. In classification, the conditional distribution of the binary

response Y_i is related to the predictor X_i through a known link function h and a regression function f as $Y_i|X_i \sim Ber\big[h\{f(X_i)\}\big]$, where f is again given a GP prior. The following developments will mainly focus on the regression case. The GP with squared exponential covariance is a commonly used prior in the literature. The law of the centered squared exponential GP $\{W_x : x \in \mathcal{X}\}$ is entirely determined by its covariance function,

$$K^a(x, y) = EW_xW_y = \exp(-a^2||x - y||^2), \tag{3.2}$$

where the predictor domain $\mathcal{X}$ is a subset of $\mathbb{R}^D$, $||\cdot||$ is the usual Euclidean norm and a is a length scale parameter. Although we focus on the squared exponential case, our results can be extended to a broader class of covariance functions with exponentially decaying spectral density, including standard choices such as Matérn, with some elaboration. We use $GP(m, K)$ to denote a GP with mean $m : \mathcal{X} \to \mathbb{R}$ and covariance $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Given n independent observations, the minimax rate of estimating a D-variate function that is only known to be Hölder s-smooth is $n^{-s/(2s+D)}$ (Stone, 1982). A function in $\mathbb{R}^D$ is said to be Hölder s-smooth if it has bounded mixed partial derivatives up to order $\lfloor s \rfloor$ for $\lfloor s \rfloor$ the largest integer strictly smaller than s with the partial derivative of order $\lfloor s \rfloor$ being Lipschitz-continuous of order $s - \lfloor s \rfloor$. Surprisingly, van der Vaart and van Zanten (2009) proved that, for Hölder s-smooth functions, a prior specified as

$$W^A|A \sim GP(0, K^A), \quad A^D \sim Ga(a_0, b_0), \tag{3.3}$$

for $Ga(a_0, b_0)$ the Gamma distribution with pdf $p(t) \propto t^{a_0-1}e^{-b_0t}$ leads to the minimax rate $n^{-s/(2s+D)}$ up to a logarithmic factor $(\log n)^\beta$ with $\beta \sim D$ adaptively over all $s > 0$ without knowing s in advance. The superscript in W^A indicates the dependence on A, which can be viewed as a scaling or inverse bandwidth parameter. Although the sample paths from this GP prior are almost surely infinitely differentiable, an

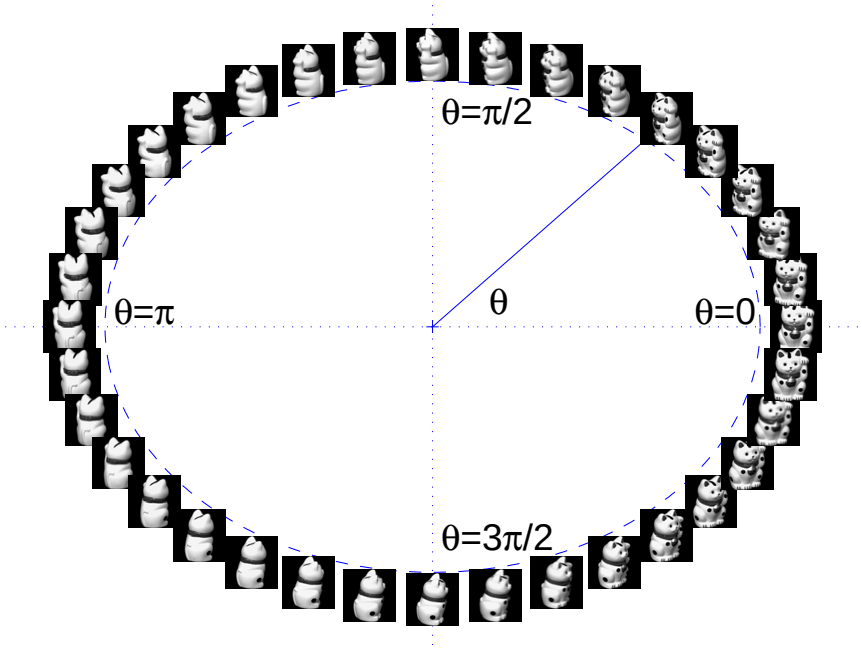


FIGURE 3.1: In this data, 72 size 128×128 images were taken for a "lucky cat" from different angles: one at every 5 degrees of rotation. 36 images are displayed in this figure.

intuitive explanation for such smoothness adaptibility is that less regular or wiggly functions can be well approximated by shrinking the long path of a smooth function by a large factor a.

In many real problems, the predictor X can be represented as a vector in high dimensional Euclidean space $\mathbb{R}^D$, where D is called the ambient dimensionality. Due to the curse of dimensionality, the minimax rate $n^{-s/(2s+D)}$ will deteriorate rapidly as D increases. This will become extremely fatal in the notorious small n large p problem, where D can be much larger than the sample size n. In such high dimensional situations, there is no hope to accurately estimate the regression function f without any assumption on the true model. One common assumption requires that f only depends on a small number $d \ll n$ of components of the vector X that are identified as important. In the GP prior framework, Savitsky et al. (2011) proposed to use "spike and slab" type point mass mixture priors for different scaling parameters

for each component of X to do Bayesian variable selection. Bhattacharya et al. (2012) showed that carefully calibrated implementations of this approach can lead to minimax adaptive rates of posterior concentration. However, variable selection is a very restrictive notion of dimension reduction. Our focus is on a different notion, which is that the predictor lies on a manifold $\mathcal{M}$ of intrinsic dimension d much lower than the ambient space dimension D. This manifold can be considered as a d dimensional hyper surface in $\mathbb{R}^D$. A rigorous definition is described in section 3. A concrete example is shown in Fig.3.1. These data (Nene et al. (1996)) consist of 72 images of a "lucky cat" taken from different angles $5°, 10°, \ldots$. The predictor $X \in \mathbb{R}^{128^2}$ is obtained by vectorizing the 128×128 image. The response Y is a continuous function f of the rotation angle $\theta \in [0, 2\pi]$ satisfying $f(0) = f(2\pi)$, such as *sin* or *cos* functions. Intuitively, the predictor X concentrates on a circle in $D = 128^2$-dim ambient space and thus the intrinsic dimension d of X is equal to one, the dimension of the rotation angle θ.

3.2.2 Our model and rate adaptivity

When $X \in \mathcal{M}$ with $\mathcal{M}$ d-dimensional, a natural question is whether we can achieve the intrinsic rate $n^{-s/(2s+d)}$ for f Hölder s-smooth without estimating $\mathcal{M}$. Surprisingly, the answer is affirmative. Ye and Zhou (2008) showed that a least squares regularized algorithm with an appropriate d dependent regularization parameter can ensure a convergence rate at least $n^{-s/(8s+4d)}(\log n)^{2s/(8s+4d)}$ for functions with Hölder smoothness $s \leqslant 1$. Bickel and Li (2007) proved that local polynomial regression with bandwidth dependent on d can attain the minimax rate $n^{-s/(2s+d)}$ for functions with Hölder smoothness $s \leqslant 2$. However, similar adaptive properties have not been established for a Bayesian procedure. In this paper, we will prove that a GP prior on the regression function with a proper prior for the scaling parameter can lead to the minimax rate for functions with Hölder smoothness $s \leqslant \{2, \gamma - 1\}$, where γ is the

smoothness of the manifold $\mathcal{M}$. In the remainder of this section, we first propose the model, and then provide a heuristic argument explaining the possibility of manifold adaptivity. Formal definitions and descriptions of important geometric concepts can be found in the next section.

Analogous to (3.3), we propose the prior for the regression function f as

$$W^A|A \sim GP(0, K^A), \quad A^d \sim Ga(a_0, b_0), \tag{3.4}$$

where d is the intrinsic dimension of the manifold $\mathcal{M}$ and K^a is defined as in (3.2) with $||\cdot||$ the Euclidean norm of the ambient space $\mathbb{R}^D$. Although the GP in (3.4) is specified through embedding in the $\mathbb{R}^D$ ambient space, we essentially obtain a GP on $\mathcal{M}$ if we view the covariance function K^a as a bivariate function defined on $\mathcal{M} \times \mathcal{M}$. Moreover, this prior has two major differences with usual GPs or GP with Bayesian variable selection:

1. Unlike GP with Bayesian variable selection, all predictors are used in the calculation of the covariance function K^a;
2. The dimension D in the prior for inverse bandwidth A is replaced with the intrinsic dimension d.

Generally, the intrinsic dimension d is unknown and needs to be estimated. Many estimation methods has been proposed (Carter et al., 2010; Camastra and Vinviarelli, 2002; Levina and Bickel, 2004; Little et al., 2009). For example, Levina and Bickel (2004) considered a likelihood based approach and Little et al. (2009) relies on singular value decomposition of local sample covariance matrix. We will use Levina and Bickel (2004) to obtain an estimator $\hat{d}$ and then plug in this estimator into our prior (3.4) to obtain an empirical Bayes approach.

In our model, we only need to estimate the intrinsic dimensionality d rather than the manifold $\mathcal{M}$. Most algorithms for learning $\mathcal{M}$ become computationally

demanding as the ambient space dimensionality D increases, while estimating d is fast even when D is tens of thousands. Moreover, although we use the full data in the calculation of the covariance function, computation is still fast for moderate sample sizes n regardless of the size of D since only pairwise Euclidean distances among D-dimensional predictors are involved whose computational complexity scales linearly in D. This dimensionality scalability provides huge gains over two stage approaches (section 2.3) in high dimensional regression settings even though they can also achieve the optimal posterior convergence rate (Theorem 15).

Intuitively, one would expect that geodesic distance should be used in the square exponential covariance function (3.2). However, there are two main advantages of using Euclidean distance instead of geodesic distance. First, when geodesic distance is used, the covariance function may fail to be positive definite. In contrast, with Euclidean distance in (3.2), K^a is ensured to be positive definite. Second, for a given manifold $\mathcal{M}$, the geodesic distance can be specified in many ways through different Riemannian metrics on $\mathcal{M}$ (section 3.1). According to Lemma 68, all these geodesic distances are equivalent to each other and the Euclidean distance on $\mathbb{R}^D$. Therefore, by using the Euclidean distance, we bypass the need to estimate geodesic distance, but still reflect the geometric structure of the observed predictors in terms of pairwise distances.

We provide heuristic explanations on why the rate can adapt to the predictor manifold through two observations. The first focuses on the possibility of obtaining an intrinsic rate for the regression problem (3.1) per se. Although the ambient space is $\mathbb{R}^D$, the support $\mathcal{M}$ of the predictor X is a d dimension submanifold of $\mathbb{R}^D$. As a result, the GP prior specified in section 2.1 has all probability mass on the functions supported on this support, leading the posterior contraction rate to entirely depend on the evaluations of f on $\mathcal{M}$. More specifically, the posterior contraction rate is

lower bounded by any sequence $\{\epsilon_n : n \geqslant 1\}$ such that

$$\Pi\big(d(f, f_0) > \epsilon_n | X^n\big) \to 0, \quad n \to \infty,$$

where $\Pi(A|X^n)$ is the posterior probability of A and $d^2(f, f_0) = (1/n)\sum_{i=1}^n \big(f(x_i) - f_0(x_i)\big)^2$ under fixed design or $d^2(f, f_0) = \int_{\mathcal{M}} \big(f(x) - f_0(x)\big)^2 G(dx)$ under random design, with G the marginal distribution for predictor X. Hence, $d(f, f_0)$ measures the discrepancy between f and the truth f_0, and only depends on the evaluation of f on $\mathcal{M}$. Therefore, in a prediction perspective, we only need to fit and infer f on $\mathcal{M}$. Intuitively, we can consider a special case when the points on manifold $\mathcal{M}$ have a global smooth representation $x = \phi(t)$, where $t \in \mathbb{R}^d$ is the global latent coordinate of x. Then the regression function

$$f(x) = f\big[\phi(t)\big] \triangleq h(t), \quad t \in \mathbb{R}^d, \tag{3.5}$$

is essentially a d-variate s-smooth function if ϕ is sufficiently smooth. Then estimation of f on $\mathbb{R}^D$ boils down to estimation of h on $\mathbb{R}^d$ and the intrinsic rate would be attainable. For the general case, we can consider parameterizing a compact manifold $\mathcal{M}$ by a finite number of local charts $\{(U_i, \phi_i) : i = 1, \ldots, m\}$ and obtain (3.5) for x in each local neighborhood $U_i \subset \mathcal{M}$. However, since the parametrization μ in (3.5) is unknown or even does not exist, one possible goal is to develop methods that can adapt to low dimensional manifold structure.

This motivates the second observation on the possibility of obtaining the intrinsic rate via the ambient space GP prior specified in (3.3). With this prior, the dependence among $\{f(x_i)\}_{i=1}^n$ is entirely characterized by the covariance matrix $(K^A(x_i, x_j))_{n\times n}$, which depends on the pairwise Euclidean distance e among observed predictors $\{x_i\}_{i=1}^n$. Ideally, a distance $d_{\mathcal{M}}$ used in the covariance matrix should be an intrinsic distance, which measures the distance by traveling from one point to the other without leaving $\mathcal{M}$. More formally, an intrinsic distance is defined as the infimum of the length of all paths between two points. In the special case of (3.5),

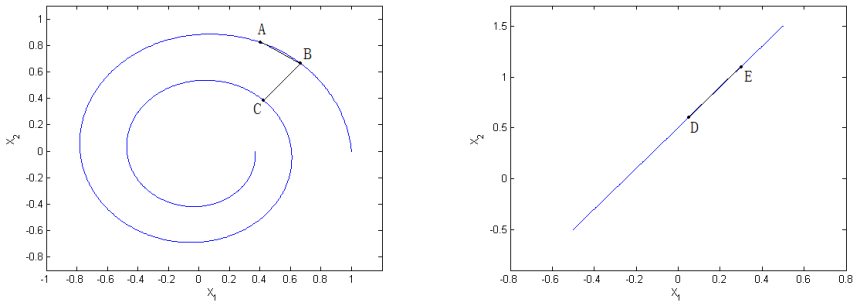


FIGURE 3.2: Examples of one dimensional submanifolds in $\mathbb{R}^2$.

$d_{\mathcal{M}}(x, x')$ would be $e\big(\phi^{-1}(x), \phi^{-1}(x')\big)$ if ϕ is an isometric embedding from $\mathbb{R}^d$ into $\mathbb{R}^D$. Fig. 3.2 also gives two simple examples where $\mathcal{M}$ is a one dimensional submanifold in $\mathbb{R}^2$. Although B and C are close in Euclidean distance, they are far away in terms of intrinsic distance, which is the length of the arc from B to C. Fortunately, Lemma 68 in the next section suggests that for compact submanifolds, this bad phenomenon only occurs for remote points — d and $d_{\mathcal{M}}$ will become comparable as two points move close. Moreover, as two points A and B become closer, using d to approximate the intrinsic distance $d_{\mathcal{M}}$ only introduces higher order error (see Proposition 67) proportional to the curvature of $\mathcal{M}$, which characterizes local distortion. In contrast, in the right plot is a straight segment in $\mathbb{R}^2$. In this case Euclidean distance always matches the intrinsic distance and whether the $\mathcal{M}$ itself is known would make no difference in predicting f since a straight segment is locally flat and has zero curvature.

A typical nonparametric approach estimates $f(x)$ by utilizing data at points near x, such as averaging over samples in a δ_n-ball around x, where the bandwidth δ_n decreases with sample size n. It is expected that as more observations come in, properly shrinking δ_n could suppress both bias and variance, where the former is caused by local averaging and the latter is due to measurement error. This is only

possible when f has certain smoothness such that large local fluctuations are not allowed. Therefore bandwidth tends to decrease at rate $n^{-1/(2s+d)}$ depending on the smoothness level s of f. Since the scaling parameter a in the covariance function K^a serves as an inverse bandwidth which would grow at rate $n^{1/(2s+d)}$, remote points tend to have exponentially decaying impact. As a result, one can imagine that accurate approximation of local intrinsic distance could provide good recovery of f as if we know the manifold and the associated intrinsic metric $d_{\mathcal{M}}$. Note that for manifold $\mathcal{M}$, the notion of "closeness" is characterized by the geodesic distances defined on $\mathcal{M}$. Often geodesic distances on $\mathcal{M}$ are not uniquely determined (section 3.1). Fortunately, Lemma 68 implies that for compact submanifolds, all distance metrics induced by Riemannian metrics on $\mathcal{M}$ are equivalent. Therefore we can choose any valid Riemannian metric as the base metric, which is the one induced by the ambient Euclidean metric in this paper. The following theorem is our main result which formalizes the above observations.

Theorem 13. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of R^D. For any $f_0 \in C^s(\mathcal{M})$ with $s \leqslant \min\{2, \gamma - 1\}$, if we specify the prior as (3.3), then (B.8) will be satisfied for ϵ_n a multiple of $n^{-s/(2s+d)}(\log n)^{\kappa_1}$ and $\bar{\epsilon}_n$ a multiple of $\epsilon_n(\log n)^{\kappa_2}$ with $\kappa_1 = (1+d)/(2+d/s)$ and $\kappa_2 = (1+d)/2$. This implies that the posterior contraction rate will be at least a multiple of $n^{-s/(2s+d)}(\log n)^{d+1}$.*

The ambient space dimension D implicitly influences the rate via a multiplicative constant. This theorem suggests that the Bayesian model (3.4) can adapt to both the low dimensional manifold structure of X and the smoothness $s \leqslant 2$ of the regression function. The reason the near optimal rate can only be allowed for functions with smoothness $s \leqslant 2$ is the order of error in approximating the intrinsic distance $d_{\mathcal{M}}$ by the Euclidean distance d (Proposition 67). Even if the intrinsic dimensionality d is misspecified as d', the following theorem still ensures the rate to be much better

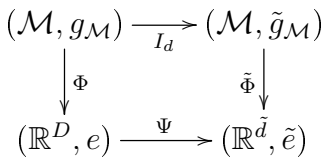


FIGURE 3.3: (Communicative) diagrams explaining the relationship between original ambient space and feature space.

than $n^{-O(1/D)}$ when d' is not too small.

Theorem 14. *Assume the same conditions as in Theorem 13, but with the prior specified as* (3.3) *with* $d' \neq d$ *and* $d' > d^2/(2s+d)$.

1. *If* $d' > d$, *then the posterior contraction rate will be at least a multiple of* $n^{-s/(2s+d')}(\log n)^{\kappa}$, *where* $\kappa = (1+d)/(2+d'/s)$;
2. *If* $\frac{d^2}{2s+d} < d' < d$, *then the posterior contraction rate will be at least a multiple of* $n^{-\frac{(2s+d)d'-d^2}{2(2s+d)d'}}(\log n)^{\kappa}$, *where* $\kappa = (d+d^2)/(2d'+dd'/s) + (1+d)/2$.

3.2.3 Dimensionality reduction and diffeomorphism invariance

Tenenbaum et al. (2000) and Roweis and Saul (2000) initiated the area of manifold learning, which aims to design non-linear dimensionality reduction algorithms to map high dimensional data into a low dimensional feature space under the assumption that data fall on an embedded non-linear manifold within the high dimensional ambient space. A combination of manifold learning and usual nonparametric regression leads to a two-stage approach, in which a dimensionality reduction map from the original ambient space $\mathbb{R}^D$ to a feature space $\mathbb{R}^{\tilde{d}}$ is estimated in the first stage and a nonparametric regression analysis with low dimensional features as predictors is conducted in the second stage. As a byproduct of Theorem 13, we provide a theoretical justification for this two stage approach under some mild conditions.

Fig. 3.3 describes relationships used in formalizing this theory. The original predictor manifold $\mathcal{M}$ sits in the ambient space $\mathbb{R}^D$. A Riemannian metric $g_{\mathcal{M}}$

on $\mathcal{M}$ is induced by the embedding map Φ and the Euclidean metric e on $\mathbb{R}^D$. $\Psi : \mathbb{R}^D \to \mathbb{R}^{\tilde{d}}$ is a dimensionality reduction map such that the restriction $\Psi_{\mathcal{M}}$ of Ψ on the embedding image $\Phi(\mathcal{M}) \simeq \mathcal{M}$ is a diffeomorphism, which requires $\Psi_{\mathcal{M}}$ to be injective and both $\Psi_{\mathcal{M}}$ and its inverse to be smooth. The former requirement would imply $\tilde{d} \geqslant d$. Diffeomorphism is the least and only requirement such that both the intrinsic dimension d of predictor X and smoothness s of regression function f are invariant. Ψ will naturally induce an embedding

$$\tilde{\Phi} = \Psi \circ \Phi : (\mathcal{M}, \tilde{g}_{\mathcal{M}}) \to (\mathbb{R}^{\tilde{d}}, \tilde{e}), \tag{3.6}$$

where the new Riemannian metric $\tilde{g}_{\mathcal{M}}$ is induced by the Euclidean metric $\tilde{e}$ of $\mathbb{R}^{\tilde{d}}$. Finally I_d is an identity map between the same set $\mathcal{M}$ with different Riemannian metrics. Such a map Ψ could also be chosen so that the induced embedding $\tilde{\Phi}$ satisfies some good properties, such as the equivariant embedding in shape analysis (Kent, 1992). Due to the dimensionality reduction, the regression function becomes

$$f(x) = f\big[\Psi_{\mathcal{M}}^{-1}(\tilde{x})\big] \triangleq \tilde{f}(\tilde{x}),$$

where $\tilde{f}$ is a well defined function on the manifold $\mathcal{M}$ represented in $\mathbb{R}^{\tilde{d}}$ and has the same smoothness as f. Therefore, by specifying a GP prior (3.3) directly on $\mathbb{R}^{\tilde{d}}$, we would be able to achieve a posterior contraction rate at least $n^{-s/(2s+d)}(\log n)^{d+1}$. The above heuristic can be formalized into the following theorem.

Theorem 15. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of R^D. Suppose that $\Psi : \mathbb{R}^D \to \mathbb{R}^{\tilde{d}}$ is an ambient space mapping (dimension reduction) such that Ψ restricted on $\Phi(\mathcal{M})$ is a $C^{\gamma'}$-diffeomorphism onto its image. Then by specifying the prior* (3.3) *with $\{\Psi(X_i)\}_{i=1}^n$ as observed predictors and Euclidean norm of $\mathbb{R}^{\tilde{d}}$ as $||\cdot||$ in* (3.2)*, for any $f_0 \in C^s(\mathcal{M})$ with $s \leqslant \min\{2, \gamma - 1, \gamma' - 1\}$, (B.8) will be satisfied for $\epsilon_n = n^{-s/(2s+d)}(\log n)^{\kappa_1}$ and $\bar{\epsilon}_n = \epsilon_n(\log n)^{\kappa_2}$ with $\kappa_1 = (1+d)/(2+d/s)$ and $\kappa_2 = (1+d)/2$. This implies that the posterior contraction rate will be at least $\epsilon_n = n^{-s/(2s+d)}(\log n)^{d+1}$.*

3.2.4 Measurement error in the predictors

In applications, predictor X_i may not exactly lie on the manifold $\mathcal{M}$. We assume that $X_i = X_{i0} + \epsilon_i$, where $X_{i0} \in \mathcal{M}$ falls on the manifold and $\epsilon_i \sim N_D(0, \sigma_X^2 I_D)$ are i.i.d measurement errors. In this case, choosing a linear projection map $\Psi^P \in \mathbb{R}^{\tilde{d} \times D}$ as the dimensionality reduction Ψ in the previous section can provide huge gain in terms of smoothing the data. As long as the elements of Ψ^P do not have large variations, the central limit theorem ensures that the noise part $\Psi^P \epsilon$ has order $O_p(D^{-1/2})$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^{D \times n}$. It is not straightforward to deterministically specify a linear projection Ψ^P having good properties. Hence, we consider randomly generating Ψ^P by sampling the elements i.i.d from a common distribution. The following multiplier central limit theorem (van der Vaart and Wellner, 2000, Lemma 2.9.5) provides support.

Lemma 16. *Let $Z_1, Z_2, \ldots$ be i.i.d. Euclidean random vectors with $EZ_i = 0$ and $E||Z_i||^2 < \infty$ independent of the i.i.d. sequence $\xi_1, \xi_2, \ldots$, with $E\xi_i = 0$ and $E\xi_i^2 = 1$. Then conditionally on $Z_1, Z_2, \ldots$,*

$$\sqrt{m} \sum_{j=1}^{m} \xi_i Z_i \to N(0, cov(Z_1)) \text{ in distribution,}$$

for almost every sequence $Z_1, Z_2, \ldots$.

For a fixed row $\Psi_l^P = (\zeta_{l1}, \ldots, \zeta_{lD})$, its i.i.d components ζ_{lj} can be viewed as ξ_j in the lemma. Denote the rows of the noise matrix ϵ by $\epsilon_{(1)}, \ldots, \epsilon_{(D)}$. Viewing $\epsilon_{(j)}$ as the Z_j, by Lemma 16, we obtain that the new projected lth predictor vector $\Psi_l^P (X_1, \ldots, X_n)^T \in \mathbb{R}^{\tilde{d}}$ has noise $\Psi_l^P \epsilon = \sum_{j=1}^{D} \Psi_{lj} \epsilon_j = O_p(D^{-1/2})$. Therefore, the noise in the original predictors is reduced by random projection. The question is then whether the projected predictors can be included in a GP regression without sacrificing asymptotic performance relative to using X_{i0}. The answer is the affirmative relying on Theorem 15 by the following argument.

Theorem 15 only requires that Ψ^P is a diffeomorphism when restricted on $\mathcal{M}$. Surprisingly, Baraniuk and Wakin (2009) (Theorem 3.1) proved more than this in the sense that for a compact d-dimensional Riemannian submanifold $\mathcal{M}$ of $\mathbb{R}^D$ and a column normalized random projection Ψ^P, if the projected dimension $\tilde{d}$ is larger than $O\big(d\delta^{-2}\log(CD\delta^{-1})\log(\rho^{-1})\big)$, where C is a positive constant depending on $\mathcal{M}$, then with probability at least $1-\rho$, for every pair of points $x, y \in \mathcal{M}$, the following holds

$$(1-\delta)\sqrt{\frac{\tilde{d}}{D}} \leqslant \frac{||\Psi^P x - \Psi^P y||}{||x-y||} \leqslant (1+\delta)\sqrt{\frac{\tilde{d}}{D}},$$

where $||\cdot||$ is the Euclidean norm in $\mathbb{R}^D$ or $\mathbb{R}^{\tilde{d}}$. This theorem implies that Ψ^P preserve the ambient distances up to a scaling $\sqrt{\tilde{d}/D}$ on the manifold by choosing $\delta \ll 1$. In addition, this distance preservation property can also be extended to geodesic distances (Baraniuk and Wakin, 2009, Corollary 3.1). Under the noised case, by normalizing the columns in Ψ^P, the noise $\Psi_l^P \epsilon$ has order $O_p(D^{-1})$, which is of higher order compare to the scaling $O(D^{-1/2})$ in this theorem. Therefore, even if noise exists, a combination of the distance preservation property with the fact that Ψ^P is a linear map implies that with large probability, Ψ^P would be a diffeomorphism when restricted on $\mathcal{M}$. Then Theorem 15 ensures that applying random projections in the first stage and plug in these projected predictors in a second state will not sacrifice anything asymptotically relative to using X_{i0} in the GP.

3.3 Numerical example

We provide a numerical example using the lucky cat data (Fig. 3.1). This data set has intrinsic dimensionality one, which is the dimension of the rotation angle θ. Since we know the true value of θ, we create the truth $f_0(\theta) = \cos\theta$ as a continuous function on the unit circle. The responses are simulated from $Y_i = f_0(\theta_i) + \epsilon_i$ by adding

independent Gaussian noises $\epsilon_i \sim N(0, 0.1^2)$ to the true values. In this model, the total sample size $N = 72$ and the predictors $X_i \in \mathbb{R}^p$ with $D = 16,384$. To assess the impact of the sample size n on the fitting performance, we randomly divide $n = 18$, 36 and 64 samples into training set and treat the rest as testing set. Training set is used to fit a model and testing set to quantify the estimation accuracy. For each training size n, we repeat this procedure for $m = 100$ times and calculate the square root of mean squared prediction error (MSPE) on the testing set,

$$\sum_{l=1}^{m} \frac{1}{N-n} \sum_{i \in T_l} ||\hat{Y}_i - f_0(\theta_i)||^2,$$

where T_l is the lth testing set and $\hat{Y}_i$ is an estimation of $E[Y|X_i] = f_0(\theta_i)$. We apply three GP based algorithms on this data set: 1. vanilla GP specified by (3.4); 2. Two stage GP (2GP) where the D-dimensional predictors were projected into $\mathbb{R}^2$ by using Laplacian eigenmap (Belkin, 2003) in the first stage and then a GP with projected features as predictors was fitted in the second stage; 3. Random projection GP (RPGP) where the new predictors were produced by projecting the original predictors into $\mathbb{R}^{1000}$ with a random projection matrix $\Psi^P = (\Psi_{lj}) \in \mathbb{R}^{1000 \times 16384}$ with $\Psi_{lj} \sim$ i.i.d. $N(0, 1)$. To assess the prediction performance, we also compare our GP prior based models (3.4) with lasso (Tibshirani, 1996) and elastic net (EN) (Zou and Hastie, 2005) under the same settings. We choose these two competing models because they are among the most widely used methods in high dimensional regression settings and perform especially good when the true model is sparse. In the GP models, we set $d = 1$ since the sample size for this dataset is too small for most dimension estimation algorithms to reliably estimate d. In addition, for each simulation, we run $10,000$ iterations with the first $5,000$ as burn-in.

The results are shown in Table. 3.1. As we can see, under each training size n, GP performs the best. Moreover, as n increases, the prediction error of GP decays much faster than EN and Lasso: when $n = 18$, the square root of MSPEs

Table 3.1: Square root of MSPE for the lucky cat data by using three different approaches over 100 random splitting are displayed. The numbers in the parenthesis indicate the standard deviations.

	$n = 18$	$n = 36$	$n = 54$
EN	.416(.152)	.198(.042)	.149(.031)
LASSO	.431(.128)	.232(.061)	.163(.038)
GP	.332(.068)	.128(.036)	**.077(.014)**
2GP	**.181(.051)**	**.124(.038)**	.092(.021)
RPGP	.340(0.071)	.130(.039)	**.077(.015)**

by using EN and lasso are about 125% of that by using GP; however as n increases to 54, this ratio becomes about 200%. Moreover, the standard deviation of square root of MSPEs by using GP are also significantly lower than those by using lasso and EN. Among GP based methods, RPGP has slightly worse performance than GP under small training size, but as n grows to 54, they have comparable MSPEs. It is not surprising that 2GP has better performance than GP when n is small since the dimensionality reduction map Ψ is constructed using the whole dataset (the Laplacian eigenmap code we use cannot do interpolations). Therefore when the training size n become closer to the total data size 72, GP becomes better. In addition, GP is computationally faster than 2GP due to the manifold learning algorithm in the first stage of 2GP.

To compare the performances between GP and RPGP in the case when there are noises in the predictors, we add $N(0, \sigma_X I_D)$ noises into each predictor vector X_i with noise levels $\sigma_X = 0, 10, 20, 40$ and 80, where the range of predictors is $0 \sim 255$. We also change the projected dimension $\tilde{d}$ from 10 to $1,000$. The training size n is fixed at 54. Table. 3.2 displays the results.

As we can see, for small $\tilde{d} = 10$ or 100, applying GP on the original predictors appears to be better than RPGP on the projected predictors under any settings. As $\tilde{d}$ grows to $1,000$, GP and RPGP have similar performances in the noise free setting. However, as noises are added to the predictors, RPGP with $\tilde{d} = 1,000$ outperforms

Table 3.2: Square root of MSPE for the lucky cat data with noised predictors. results over 100 random splitting are displayed. The numbers in the parenthesis indicate the standard deviations.The numbers after RPGP indicates the projected dimension $\tilde{d}$.

σ_X	0	10	20	40	80
GP	**.077(.014)**	.095(.015)	.116(.017)	.180(.020)	**.276(.23)**
RPGP(10)	.275(.065)	.291(.069)	.335(.075)	.452(.085)	.606(.102)
RPGP(100)	.106(.023)	.116(.026)	.143(.033)	.225(.043)	.360(.065)
RPGP(1000)	**.077(.015)**	**.088(.017)**	**.102(.018)**	**.178(.021)**	.289(.033)

GP. However, as the noise increases to the order comparable to the signals, GP becomes close to and finally outperforms RPGP. In addition, the standard deviation of RPGP also grows rapidly as noise increases. This suggests that GP might be more stable than RPGP under small signal-to-noise ratio scenarios.

3.4 Discussion

In this work, we considered a nonparametric Bayesian prior for high dimensional regression when the predictors are assumed to be lying on a low dimensional intrinsic manifold. The proposed prior can be considered as an extension of a Gaussian process prior on Euclidean space to a general submanifold. We show that this GP prior can attain near optimal posterior convergence rate that can adapt to both the smoothness of the true function ($s \leqslant 2$) and the underlying intrinsic manifold $\mathcal{M}$. Our theorem validates the surprising phenomenon suggested by Bickel in his 2004 Rietz lecture (Bickel and Li, 2007) under the GP prior scenario:

> "... the procedures used with the expectation that the ostensible dimension D is correct will, with appropriate adaptation not involving manifold estimation, achieve the optimal rate for manifold dimension d."

Moreover, we also provide theoretical guarantees for two stage GP with dimensionality reduction. We suggest the use of random projection GP as a special two stage GP when noises exist in the predictors.

One possibility of our future work is to investigate whether the smoothness requirement $s \leqslant 2$ could be relaxed. This extension will be dependent on whether Lemma 71 could be improved to $s \geqslant 2$. Currently we construct the approximation function $I_a(f)$ in RKHS through convolving f with the covariance function. It is not clear whether this is the best way to approximate the function f by elements in the RKHS.

A second possibility is to build a coherent model not only estimating the regression function $E[Y|X]$, but simultaneously learning the dimensionality d of the intrinsic manifold $\mathcal{M}$. Our current GP prior (3.4) completely ignores the information contained in the marginal distribution P_X of the predictor X. As an alternative, we can only model part of P_X and therefore utilize some of P_X's information, such as the support or dimensionality of $\mathcal{M}$.

4

Bayesian conditional tensor factorizations for high-dimensional classification

4.1 Introduction

Classification problems involving high-dimensional categorical predictors have become common in a variety of application areas, with the goals being not only to build an accurate classifier but also to identify a sparse subset of important predictors. For example, genetic epidemiology studies commonly focus on relating a categorical disease phenotype to single nucleotide polymorphisms encoding whether an individual has 0, 1 or 2 copies of the minor allele at a large number of loci across the genome. In such applications, it is expected that interactions play an important role, but there is a lack of statistical methods for identifying important predictors that may act through both main effects and interactions from a high-dimensional set of candidates. Our goal is to develop nonparametric Bayesian methods for addressing this gap.

There is a rich literature on methods for prediction and variable selection from high or ultra high-dimensional predictors with a categorical response. The most

common strategy would rely on logistic regression with the linear predictor having the form $x_i'\beta$, with $x_i = (x_{i1}, \ldots, x_{ip})'$ denoting the predictors and $\beta = (\beta_1, \ldots, \beta_p)'$ regression coefficients. In high-dimensional cases in which p is the same order of n or even $p > n$, classical methods such as maximum likelihood break down but there is a rich variety of alternatives ranging from penalized regression to Bayesian variable selection. Popular methods include L_1 penalization (Tibshirani, 1996) and the elastic net (Zou and Hastie, 2005), which combines L_1 and L_2 penalties to accommodate $p \gg n$ cases and allow simultaneous selection of correlated sets of predictors. For efficient L_1 regularization in generalized linear models including logistic regression, Park and Hastie (2007) proposed a solution path method. Genkin et al. (2007) propose a related Bayesian approach for high-dimensional logistic regression under Laplace priors. Wu et al. (2009) applied L_1 penalized logistic regression to genome wide association studies. Potentially, related methods can be applied to identify main effects and epistatic interactions (Yang et al., 2010), but direct inclusion of interactions within a logistic model creates a daunting dimensionality problem limiting attention to low-order interactions and modest numbers of predictors.

These limitations have motivated a rich variety of nonparametric classifiers, including classification and regression trees (CART) (Breiman et al., 1984) and random forests (RFs) (Breiman, 2001). CART partitions the predictor space so that samples within the same partition set have relatively homogeneous outcomes. CART can capture complex interactions and has easy interpretation, but tends to be unstable computationally and lead to low classification accuracy. RFs extend CART by creating a classifier consisting of a collection of trees that are all used to vote for classification. RFs can substantially reduce variance compared to a single tree and result in high classification accuracy, but provide an uninterpretable machine that does not yield insight into the relationship between specific predictors and the outcome. Moreover, through our simulation results in section 6, we found that random

forests did not behave well in high dimensional low signal-to-noise cases.

Our focus is on developing a new framework for nonparametric Bayes classification through tensor factorizations of the conditional probability $P(Y = y \,|\, X_1 = x_1, \ldots, X_p = x_p)$, with $Y \in \{1, \ldots, d_0\}$ a categorical response and $X = (X_1, \ldots, X_p)'$ a vector of p categorical predictors. The conditional probability can be expressed as a $d_1 \times \cdots \times d_p$ tensor for each class label y, with d_j denoting the number of levels of the jth categorical predictor X_j. If $p = 2$ we could use a low rank matrix factorization of the conditional probability, while in the general p case we could consider a low rank tensor factorization. Such factorizations must be non-negative and constrained so that the conditional probabilities add to one for each possible X, and are fully flexible in characterizing the classification function for sufficiently high rank. Dunson and Xing (2009) and Bhattacharya et al. (2012) applied two different tensor decomposition methods to model the joint probability distribution for multivariate categorical data. Although an estimate of the joint pmf can be used to induce an estimate of the conditional probability, there are clear advantages to bypassing the need to estimate the high-dimensional nuisance parameter corresponding to the marginal distribution of X.

We address such issues using a Bayesian approach that places a prior over the parameters in the factorization, and provide strong theoretical support for the approach while developing a tractable algorithm for posterior computation. Some advantages of our approach include (i) fully flexible modeling of the conditional probability allowing any possible interactions while favoring a parsimonious characterization; (ii) variable selection; (iii) a full probabilistic characterization of uncertainty providing measures of uncertainty in variable selection and predictions; and (iv) strong theoretical support in terms of rates at which the full posterior distribution for the conditional probability *contracts* around the truth. Notably, we are able to obtain near a parametric rate even in ultra high-dimensional settings in which the number

of candidate predictors increases exponentially with sample size. Such a result differs from frequentist convergence rates in characterizing concentration of the entire posterior distribution instead of simply a point estimate. Similar contraction rate results in p diverging with n settings are currently only available in simple parametric models, such as the normal means problem (Castillo and van der Vaart, 2012) and generalized linear models (Jiang, 2006). Although our computational algorithms do not yet scale to massive dimensions, we can accommodate $1,000$s of predictors.

4.2 Conditional Tensor Factorizations

In section 2.1, we briefly introduce the tensor factorization techniques and describe their relevance to high-dimensional classification. In section 2.2 and 2.3, we characterize two desirable properties, which only rely on the structure of our proposed model.

4.2.1 Tensor factorization of the conditional probability

Although there is a rich literature on tensor decompositions, little is in statistics. The focus has been on two factorizations that generalize matrix singular value decomposition (SVD). The most popular is parallel factor analysis (PARAFAC) (Harshman, 1970; Harshman and Lundy, 1994; Zhang and Golub, 2001), which expresses a tensor as a sum of r rank one tensors, with the minimal possible r defined as the rank (Fig.4.1). The second approach is Tucker decomposition or higher-order singular value decomposition (HOSVD), which was proposed by Tucker (1966) for three-way data and extended to arbitrary orders by De Lathauwer et al. (2000). HOSVD expresses $d_1 \times \cdots \times d_p$ tensor $A = \{a_{c_1 \cdots c_p}\}$ as

$$a_{c_1 \cdots c_p} = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} g_{h_1 \cdots h_p} \prod_{j=1}^{p} u^{(j)}_{h_j c_j}, \tag{4.1}$$

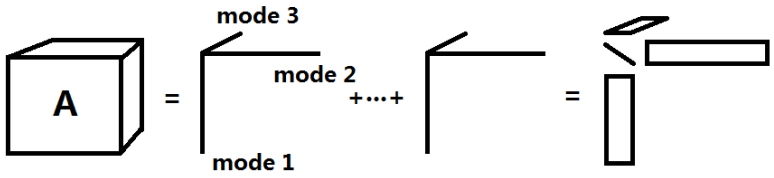


FIGURE 4.1: A diagram describes PARAFAC for 3 dimensional tensor. The lines in the middle correspond to the mode vectors corresponding to each mode of the tensor. The rightmost representation draws analogy to the matrix SVD.

where $k_j(\leqslant d_j)$ is the j-rank for $j = 1, \ldots, p$ and $G = \{g_{h_1 \cdots h_p}\}$ is a core tensor, with constraints on G such as low rank and sparsity imposed to induce better data compression and fewer components compared to PARAFAC (Fig.4.2). This is intuitively suggested by comparing Fig.4.1 and Fig.4.2: PARAFAC can be considered as a special case of HOSVD when the core tensor G is restricted to be diagonal. In HOSVD, the j-rank k_j is the rank of the mode j matrix $A_{(j)}$, defined by rearranging elements of the tensor A into a $d_j \times d_1 \cdots d_{j-1} d_{j+1} \cdots d_p$ matrix such that each row consists of all elements $a_{c_1 \cdots c_p}$ with the same c_j. Although k_j can be close to d_j, low rank approximations of A can lead to high accuracy and provide satisfactory results (Eldén and Savas (2009),Vannieuwenhoven et al. (2012)).

For probability tensors, we need nonnegative versions of such decompositions (Kim and Choi (2007)) and the concept of rank changes accordingly (Cohen and Rothblum, 1993). In the following, we solely consider nonnegative HOSVD, where all quantities in (4.1) are nonnegative. We define $k = (k_1, \ldots, k_p)$ to be a multirank of a nonnegative tensor A if: 1. A has a representation (4.1) with k; 2. k has the minimum possible size, which is defined by $|k| = \prod_{j=1}^{p} k_j$. Note that the rank in this definition might not be unique but representations with different multirank k have the same number of parameters in the core tensors. This suggests that the multirank k reflects the best possible tensor compression level.

The conditional probability $P(Y = y | X_1 = x_1, \ldots, X_p = x_p)$ can be structured as

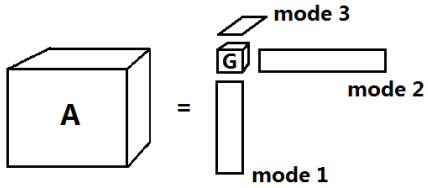


FIGURE 4.2: A diagram describes HOSVD for 3 dimensional tensor. The smaller cube G is the core tensor and the rectangles are the mode matrices $u^{(j)}$'s corresponding to each mode of the tensor.

a $d_0 \times d_1 \times \cdots \times d_p$ dimensional tensor. We call such tensors *conditional probability tensors*. Let $\mathcal{P}_{d_1,\ldots,d_p}(d_0)$ denote the set of all conditional probability tensors, so that $P \in \mathcal{P}_{d_1,\ldots,d_p}(d_0)$ implies

$$P(y|x_1,\ldots,x_p) \geqslant 0 \ \forall y, x_1,\ldots,x_p, \quad \sum_{y=1}^{d_0} P(y|x_1,\ldots,x_p) = 1 \ \forall x_1,\ldots,x_p.$$

To ensure that P is a valid conditional probability, the elements of the tensor must be non-negative with constraints on the first dimension for Y. A primary goal is accommodating high-dimensional covariates, with the overwhelming majority of cells in the table corresponding to unique combinations of Y and X unoccupied. In such settings, it is necessary to encourage borrowing information across cells while favoring sparsity.

Our proposed model for the conditional probability has the form:

$$P(y|x_1,\ldots,x_p) \quad = \quad \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \lambda_{h_1 h_2 \ldots h_p}(y) \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j), \tag{4.2}$$

with the parameters subject to

$$\sum_{c=1}^{d_0} \lambda_{h_1 h_2 \ldots h_p}(c) = 1, \text{ for any possible combination of } (h_1, h_2, \ldots, h_p),$$

$$\sum_{h=1}^{k_j} \pi_h^{(j)}(x_j) = 1, \text{ for any possible pair of } (j, x_j). \tag{4.3}$$

Analogous to HOSVD, we preserve the names core tensor for $\Lambda = \{\lambda_{h_1 \cdots h_p}(y)\}$ and mode matrices for $\pi = \{\pi^{(j)}_{h_j}(x_j)\}$. More specifically, the $d_j \times k_j$ matrix $\pi^{(j)}$ with (u, v)th element $\pi^{(j)}_v(u)$ will refer to the jth mode matrix. Similar to the definition of multirank for nonnegative tensors, we define $k = (k_1, \ldots, k_p)$ to be a multirank of the conditional probability tensor P if: 1. P has a representation (C.12) satisfying the constraints (C.10) with k; 2. k has the minimum possible size $|k|$. In the rest of this article, we always consider the representation (C.12) with a multirank k. Intuitively, $(d_0 - 1)|k|$ is equal to the degrees of freedom of the core tensor Λ, and controls the complexity of the model. By allowing $|k|$ to gradually increase with sample size, one can obtain a sieve estimator. The value of k_j controls the number of parameters used to characterize the impact of the jth predictor. In the special case in which $k_j = 1$, the jth predictor is excluded from the model, so sparsity can be imposed by setting $k_j = 1$ for most j's.

We format the conditional probability $P(y|x_1, \ldots, x_p)$ as a $d_1 \times \cdots \times d_p$ vector

$$Vec\{P(y|-)\} = \big\{P(y|1, \ldots, 1, 1), P(y|1, \ldots, 1, 2), \ldots, P(y|1, \ldots, 1, d_p), \ldots, P(y|1, \ldots, d_{p-1}, d_p), \ldots, P(y|d_1, \ldots, d_{p-1}, d_p)\big\}'$$

and $\lambda_{h_1, \ldots, h_p}(y)$ as a $k_1 \times \cdots \times k_p$ vector

$$Vec\{\Lambda(y)\} = \big\{\lambda_{1,\ldots,1,1}(y), \lambda_{1,\ldots,1,2}(y), \ldots, \lambda_{1,\ldots,1,k_p}(y), \ldots, \lambda_{1,\ldots,k_{p-1},k_p}(y), \ldots, \lambda_{k_1,\ldots,k_p}(y)\big\}'.$$

Let $\pi^{(j)}$ be a $d_j \times k_j$ matrix with $\pi^{(j)}_v(u)$ as the (u, v)th element. It is a stochastic matrix, so rows sum to one, by constraint (C.10). Then representation (C.12) can be written in vector form:

$$Vec\{P(y|-)\} = \big(\pi^{(1)} \otimes \pi^{(2)} \otimes \cdots \otimes \pi^{(p)}\big) Vec\{\Lambda(y)\}, \text{ for } y = 1, \ldots, d_0, \qquad (4.4)$$

where $\otimes$ denotes the Kronecker product. Furthermore, if we let $Mat(P)$ and $Mat(\Lambda)$ be two stochastic matrices with the yth column $Vec\{P(y|-)\}$ and $Vec\{\Lambda(y)\}$ respectively for $y = 1, \ldots, d_0$, then we can write the above d_0 identities together as:

$$Mat(P) = \big(\pi^{(1)} \otimes \pi^{(2)} \otimes \cdots \otimes \pi^{(p)}\big) Mat(\Lambda).$$

The following theorem provides basic support for factorization (C.12)-(C.10) through showing that any conditional probability has this representation. The proof of this theorem, which can be found in the appendix, sheds some light on the meaning of $k_1, \ldots, k_p$ and how it is related to a sparse structure of the tensor.

Theorem 17. *Every $d_0 \times d_1 \times d_2 \times \cdots \times d_p$ conditional probability tensor $P \in \mathcal{P}_{d_1,\ldots,d_p}(d_0)$ can be decomposed as (C.12), with $1 \leqslant k_j \leqslant d_j$ for $j = 1, \ldots, p$. Furthermore, $\lambda_{h_1 h_2 \ldots h_p}(y)$ and $\pi^{(j)}_{h_j}(x_j)$ can be chosen to be nonnegative and satisfy the constraints (C.10).*

We can simplify the representation through introducing p latent class indicators $z_1, \ldots, z_p$ for $X_1, \ldots, X_p$, with Y conditionally independent of $(X_1, \ldots, X_p)$ given $(z_1, \ldots, z_p)$. The model can be written as

$$\begin{aligned} Y_i|z_{i1}, \ldots, z_{ip} &\sim \text{Multinomial}\big(\{1, \ldots, d_0\}, \lambda_{z_{i1},\ldots,z_{ip}}\big), \\ z_{ij}|X_j &\sim \text{Multinomial}\big(\{1, \ldots, k_j\}, \pi^{(j)}_1(X_j), \ldots, \pi^{(j)}_{k_j}(X_j)\big), \end{aligned} \tag{4.5}$$

where $\lambda_{z_{i1},\ldots,z_{ip}} = \big\{\lambda_{z_{i1},\ldots,z_{ip}}(1), \ldots, \lambda_{z_{i1},\ldots,z_{ip}}(d_0)\big\}$. Marginalizing out the latent class indicators, the conditional probability of Y given $X_1, \ldots, X_p$ has the form in (C.12).

4.2.2 Bias-variance trade off

In tensor factorization model (C.12), the multirank k controls the sparsity, characterizing the impact of each predictor X_j through the "effective category count" k_j. For example, if the level of X_1, say $1, 2, 3$, can be divided into 2 classes $\{1\}$ and $\{2, 3\}$

such that $P(Y = y|X_1 = 2, \ldots, X_p = x_p) \equiv P(Y = y|X_1 = 3, \ldots, X_p = x_p)$, then k_1 is equal to 2. The following illustration suggests that to select k, we can use a hard clustering approximation by setting $\pi_{h_j}^{(j)}(x_j)$ to be either zero or one (section 4.2).

We initially provide a heuristic argument to demonstrate the tendency of our model to produce low mean squared error (MSE), which is defined as:

$$\begin{aligned}
\mathrm{MSE}(\tilde{P}) &= \int \sum_{y=1}^{d_0} E\big(\tilde{P}(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)\big)^2 G(dx_1, \ldots, dx_p) \\
&= \int \sum_{y=1}^{d_0} \big(E\tilde{P}(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)\big)^2 G(dx_1, \ldots, dx_p) \\
&\quad + \int \sum_{y=1}^{d_0} Var\tilde{P}(y|x_1, \ldots, x_p) G(dx_1, \ldots, dx_p) \\
&\triangleq \mathrm{Bias}^2(\tilde{P}) + \mathrm{Var}(\tilde{P}), \qquad (4.6)
\end{aligned}$$

where $\tilde{P}$ is an estimator of the truth P_0, G is the joint marginal distribution of the covariates X and the expectation is taken with respect to the joint distribution of (X, Y). Our focus is on obtaining accurate estimates of the conditional probability $P(Y|X)$; accurate estimates will lead to accurate classification while containing information on classification uncertainty, of critical importance in medical decision making among other areas.

For simplicity of exposition, assume the response Y to be binary. Denote by $\mathcal{T}$ the set of all conditional probability tensors parameterized by (C.12). Let $\mathcal{T}_0$ be a subset of $\mathcal{T}$ consisting of models with $\pi_{h_j}^{(j)}(x_j)$ being either zero or one. Then given k and π, $\pi^{(j)}$ uniquely determines a hard clustering of X_j: $X_j = x_j$ belongs to the $h_j(x_j)$th cluster, where $h_j(x_j)$ is the unique h_j such that $\pi_{h_j}^{(j)}(x_j) = 1$. Consider approximating P_0 by this subset $\mathcal{T}_0$. Intuitively, the best MSE attained within $\mathcal{T}_0$ gives an upper bound on the optimal MSE achievable by the whole model class $\mathcal{T}$. To demonstrate the bias-variance trade-off in terms of the selection of the multirank

k, we compare the MSE of the maximum likelihood estimators (MLE) in model space $\mathcal{T}_0$ under different k and the clustering scheme determined by π. Define

$$\epsilon_M = \inf_{P \in \mathcal{T}_0 : |k(P)| \leqslant M} ||P - P_0||,$$

where $|k(P)|$ denotes the size of the multirank of the conditional probability tensor P and

$$||P - P_0|| = \left\{ \int \sum_{y=1}^{2} |P(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)|^2 G(dx_1, \ldots, dx_p) \right\}^{1/2}. \quad (4.7)$$

ϵ_M can be interpreted as the smallest error or bias caused by approximating P_0 using $P \in \mathcal{T}_0$ with size $|k(P)| \leqslant M$, related to compressibility of P_0.

Under degeneracy of the π's, $P(y|x_1, \ldots, x_p) = \lambda_{h_1(x_1)\ldots h_p(x_p)}(y)$, where $h_j(x_j)$ is defined previously as the unique h_j such that $\pi_{h_j}^{(j)}(x_j) = 1$. Given k and π, the MLE of $\lambda_{h_1 \ldots h_p}$ is the sample frequencies of $Y_i = y$ among all observations with covariates $X_i = (X_{i1}, \ldots, X_{ip})$ satisfying $h_j(X_{ij}) = h_j$ for each $j = 1, \ldots, p$:

$$\hat{\lambda}_{h_1 \ldots h_p}(i) = \frac{\sum_{(x_1,\ldots,x_p): h_j(x_j) = h_j} \sum_{i=1}^{n} I(X_{i1} = x_1, \ldots, X_{ip} = x_p, Y_i = i)}{\sum_{(x_1,\ldots,x_p): h_j(x_j) = h_j} \sum_{i=1}^{n} I(X_{i1} = x_1, \ldots, X_{ip} = x_p)}, \ i = 1, 2,$$

where $0/0$ is defined to be 0 for simplicity. Although given k and π an unbiased estimator does not exist due to model misspecification, the following lemma shows that this MLE is still optimal in terms of minimizing the bias. A proof is sketched in the appendix.

Lemma 18. *Given k and π, among all estimators of λ's, the MLE defined above minimizes the Bias$^2(\tilde{P})$ in* (4.6).

This lemma indicates that the ϵ_M has another characterization as

$$\epsilon_M = \min_{(k,\pi): |k| \leqslant M, \pi \text{ degenerate}} \text{Bias}\big(\hat{P}(k, \pi)\big),$$

where $\hat{P}(k,\pi)$ is the MLE of P given (k,π).

Intuitively, under the degeneracy of π, n samples are separated into $|k|$ clusters to estimate the corresponding λ's, and the variance term in (4.6) should be of order $|k|/n$. The following lemma formalizes this and a proof is sketched in the appendix.

Lemma 19. *Given k and π, the $Var(\tilde{P})$ as defined in* (4.6) *for the MLE $\hat{P}$ satisfies*

$$Var\big(\hat{P}(k,\pi)\big) = C|k|/n + O(|k|/n^2), \tag{4.8}$$

where the constant $C \in [a,b]$, where $a,b>0$ only depends on P_0 and G.

Combining Lemma 18 and 19, given k and π, the MSE of MLE $\hat{P}$ satisfies:

$$\mathrm{MSE}\big(\hat{P}(k,\pi)\big) \geqslant \epsilon_{|k|}^2 + C\frac{|k|}{n} + O(|k|/n^2).$$

This reflects the so-called bias-variance trade-off for our model: as $|k|$ increases, the model becomes more complex and thus the bias term decreases; however, the variance term increases as more parameters are introduced. Therefore, there exists an optimal model size $|k|$ that solves $|k| = n\epsilon_{|k|}^2$ minimizing the MSE. This typical trade-off also appears in the Assumption B in section 3.2 where the posterior convergence rate is studied.

4.2.3 Borrowing of information

The previous section discussed the bias-variance trade-off for a subclass of models specified by (C.12), where π's are degenerate at zero and one. In this section, we illustrate another desirable property by allowing π's to be continuous on $[0,1]$: borrowing of information across cells corresponding to each combination of $X_1,\ldots,X_p$. Letting $w_{h_1,\ldots,h_p}(x_1,\ldots,x_p) = \prod_j \pi_{h_j}^{(j)}(x_j)$, model (C.12) is equivalent to

$$P(Y=y|X_1=x_1,\ldots,X_p=x_p) = \sum_{h_1,\ldots,h_p} w_{h_1,\ldots,h_p}(x_1,\ldots,x_p)\lambda_{h_1\ldots h_p}(y), \tag{4.9}$$

and constraints (C.10) imply $\sum_{h_1,\ldots,h_p} w_{h_1,\ldots,h_p}(x_1,\ldots,x_p) = 1$. In the special case when π is degenerate, $\lambda_{h_1\ldots h_p}(y)$ is just the conditional probability of $Y = y$ given the observations in cluster $h_1(X_1) = h_1, \ldots, h_p(X_p) = h_p$ (for details, refer to the descriptions in the paragraph before (4.7)). If π's are allowed to be continuous, then our model essentially uses a kernel estimate that allows borrowing of information across clusters via a weighted average of the cluster frequencies.

To illustrate the strength of this, consider a simplified example involving one covariate X with m categories and a binary response Y. In fact, each category of X can correspond to a cluster as in the preceding paragraph and the implications can be extended to our model by changing the notations. Let $P_j = P(Y = 1|X = j)$ for $j = 1, \ldots, m$. Then the MLE for $(P_1, \ldots, P_m)$ is sample frequencies $(s_1/n_1, \ldots, s_m/n_m)$, denoted by $(\hat{P}_1, \ldots, \hat{P}_m)$, where $s_j = \sharp\{i : y_i = 1 \text{ and } x_i = j\}$ and $n_j = \sharp\{i : x_i = j\}$. Instead, kernel estimates (4.9) are

$$\tilde{P}_k = \left\{1 - \sum_{j\neq k} w_{jk}\right\}\hat{P}_k + \sum_{j\neq k} w_{jk}\hat{P}_j,\ k = 1, \ldots, m,$$

where w_{jk} could be considered as the weight of the contribution to cluster k by cluster j. MLE corresponds to a special case when $w_{jk} = 0$ for all $j \neq k$. We use squared loss to compare these two estimators. After some calculations,

$$E\{L(\hat{P}, P)\} = \sum_{j=1}^{m} E(\hat{P}_j - P_j)^2 = \sum_{j=1}^{m} \frac{P_j(1-P_j)}{n_j},$$

and $E\{L(\tilde{P}, P)\} = \sum_{j=1}^{m} E(\tilde{P}_j - P_j)^2$ is a function of w_{jk}'s, whose partial derivative with respect to $w_{jk}(j \neq k)$ at zero is

$$\left.\frac{\partial E\{L(\tilde{P}, P)\}}{\partial w_{jk}}\right|_{w_{st}=0,\forall s\neq t} = -2\frac{P_k(1-P_k)}{n_k}.$$

This implies that $E\{L(\tilde{P}, P)\}$ will be reduced by $2\frac{P_k(1-P_k)}{n_k}$ for every unit increasing of w_{jk} near zero. Particularly when n_k is small, borrowing information from other

cluster $j(\neq k)$ will considerably reduce $E\{L(\tilde{P}, P)\}$ compare to MLE. In the special case when all w_{jk} are equal, $E\{L(\tilde{P}, P)\}$ can attain a minimum

$$E\{L(\hat{P}, P)\}\left[1 - \left(1 - \frac{1}{m}\right)\frac{E\{L(\hat{P}, P)\}}{E\{L(\hat{P}, P)\} + \frac{1}{m-1}\sum_{i<j}(P_i - P_j)^2}\right]$$
$$\in \left(\frac{1}{m}E\{L(\hat{P}, P)\}, E\{L(\hat{P}, P)\}\right).$$

This suggests that when P_j's are similar, the estimate $\tilde{P}$ can reduce the risk up to only $1/m$ the risk of estimating $\hat{P}$ separately. If P_j's are not similar, $\tilde{P}$ can still reduce the risk considerably when the cell counts $\{n_j\}$ are small.

Another interesting feature of our tensor model is the special structure of the weights w's in (4.9). Consider a class of continuous $\tilde{\pi}$'s indexed by a single parameter $c \in (0, 1)$ characterizing the strength of borrowing information,

$$\tilde{\pi}^{(j)}_{h_j}(x_j) = (1 - k_j c)I\big\{h_j = h_j(x_j)\big\} + cI\big\{h_j \neq h_j(x_j)\big\},$$

for $h_j \leqslant k_j$ and all possible x_j's. This $\tilde{\pi}$ still satisfies constraint (C.10) and the weight becomes

$$\tilde{w}_{h_1,\ldots,h_p}(x_1, \ldots, x_p) = \prod_{j=1}^{p}(1 - k_j c)^{I\{h_j = h_j(x_j)\}} c^{I\{h_j \neq h_j(x_j)\}}.$$

When c is small, given x, the weight of the contribution by the cluster indexed by $(h_1, \ldots, h_p)$ is approximately equal to c^s, where $s = \sum_{j=1}^{p} I\{h_j \neq h_j(x_j)\}$ is the number of latent classes not shared by $(h_1, \ldots, h_p)$ and $(h_1(x_1), \ldots, h_p(x_p))$, i.e. the Hamming distances between the latent class indices. This special structure in the weights suggests that similar clusters should share more information.

4.3 Bayesian Tensor Factorization

In this section, we will provide a Bayesian implementation of the tensor factorization model and prove the corresponding posterior convergence rate.

4.3.1 Prior specification

To complete a Bayesian specification of our model, we choose independent Dirichlet priors for the parameters $\Lambda = \{\lambda_{h_1,\ldots,h_p}, h_j = 1,\ldots,k_j, j = 1,\ldots,p\}$ and $\pi = \{\pi_{h_j}^{(j)}(x_j), h_j = 1,\ldots,k_j, x_j = 1,\ldots,d_j, j = 1,\ldots,p\}$,

$$\begin{aligned} \left\{\lambda_{h_1,\ldots,h_p}(1),\ldots,\lambda_{h_1,\ldots,h_p}(d_0)\right\} &\sim \text{Diri}(1/d_0,\ldots,1/d_0), \\ \left\{\pi_1^{(j)}(x_j),\ldots,\pi_{k_j}^{(j)}(x_j)\right\} &\sim \text{Diri}(1/k_j,\ldots,1/k_j), j = 1,\ldots,p. \end{aligned} \quad (4.10)$$

These priors have the advantages of imposing non-negative and sum to one constraints, while leading to conditional conjugacy in posterior computation. The hyperparameters in the Dirichlet priors are chosen to favor placing most of the probability on a few elements, inducing near sparsity in these vectors.

If $k_j = 1$ in (C.12), by constraints (C.10) $\pi_1^{(j)}(x_j) = 1$, $P(y|x_1,\ldots,x_p)$ will not depend on x_j and $Y \perp X_j|X_{j'}, j' \neq j$. Hence, $I(k_j > 1)$ are variable selection indicators. In addition, k_j can be interpreted as the number of latent classes for the jth covariate. Levels of X_j are clustered according to their relations with the response variable in a soft probabilistic manner, with $k_1,\ldots,k_p$ controlling the complexity of the latent structure as well as sparsity. Because we are faced with extreme data sparsity in which the vast majority of combinations of $Y, X_1,\ldots,X_p$ are not observed, it is critical to impose sparsity assumptions. Even if such assumptions do not hold, they have the effect of massively reducing the variance, making the problem tractable. A sparse model that discards predictors having less impact and parameters having small values may still explain most of the variation in the data, resulting in a useful classifier that has good performance in terms of the bias-variance tradeoff even when sparsity assumptions are not satisfied.

To embody our prior belief that only a small number of k_j's are greater than one,

we want

$$P(k_j = k) \approx Q(j,k) \triangleq \left(1 - \frac{r}{p}\right) I(k=1) + \frac{r}{(d_j - 1)p} I(k > 1),$$

for $j = 1, \ldots, p$, where $I(A)$ is the indicator function for the event A and r is the expected number of predictors included. This specification accommodates variable selection. To further include a low rank constraint on the conditional probability tensor, we impose $|k| = \prod_{j=1}^{p} k_j$ to be less than or equal to M. Intuitively, M controls the effective number of parameters in the model. This low rank constraint in turn restricts the maximum number of predictors to be $\log_2 M$. We note that in the setting in which $p > n$ some such constraint is necessary.

To summarize, the effective prior on the k_j's is

$$P(k_1 = l_1, \ldots, k_p = l_p) \propto Q(1, l_1) \cdots Q(p, l_p) \; I\left\{ \prod_{j=1}^{p} l_j \leqslant M \right\}. \tag{4.11}$$

Let $\gamma = (\gamma_1, \ldots, \gamma_p)'$ be a vector having elements $\gamma_j = I(k_j > 1)$ indicating inclusion of the jth predictor. Since $\prod_{j=1}^{p} l_j \leqslant M$ implies inclusion of at most $\log_2 M$ predictors, the induced prior for γ resembles the prior in Jiang (2006). Potentially, we can put a more structured prior on the components in the conditional tensor factorization, including sparsity in Λ. However, the theory shown in the next part provides strong support for prior (4.10)-(4.11).

4.3.2 Posterior convergence rates

Before formally describing the sparsity and low rank assumptions, we first introduce some notations and definitions. Suppose we obtain data for n observations $y^n = (y_1, \ldots, y_n)'$, which are conditionally independent given $X^n = (x_1, \ldots, x_n)'$ with $x_i = (x_{i1}, \ldots, x_{ip_n})'$, $x_{ij} \in \{1, \ldots, d\}$ and $p_n \gg n$. We exclude the n subscript on p and other quantities when convenient and assume that $d = \max_j\{d_j\}$ is finite and does

not depend on n. An important special case is when all d_j's are the same. Let P_0 denote the true data generating model, which can be dependent on n. Let ϵ_n be a sequence converging to zero while keeping $n\epsilon_n^2 \to \infty$. This sequence will serve as the convergence rate in the sense that under a certain metric d to be defined later, the posterior of the conditional probability tensor P will asymptotically concentrate within an ϵ_n d-ball centered on the truth P_0. We use the notation $f \prec g$ to mean $f/g \to 0$ as $n \to \infty$. Next, we describe all the assumptions that are needed for the main theorem.

To determine the posterior convergence rate, two things are competing with each other: 1. variable selection among the high dimension covariates; 2. the approximation abilities of near low rank tensors. The assumption below characterizes the first.

Assumption A. $\log p_n \prec n\epsilon_n^2 / \log D_n$.

Recalling the definition of D_n as the prior upper threshold for the size $|k| = \prod_{j=1}^{p} k_j$, $\log D_n$ can be interpreted as the maximum number of predictors to be selected and cannot exceed $\log n$. As a result, Assumption A implies that the high dimensional variable selection per se imposes a lower bound for ϵ_n as $\sqrt{\log n \log p_n / n}$. As a result, to obtain a convergence rate of $n^{-(1-\alpha)/2}$ up to some logarithmic factor, p_n is allowed to increase with n as fast as $o(e^{n^\alpha})$.

To characterize the low rank tensor assumption, rather than assume that most of the predictors have no impact on Y, we consider the situation similar to Jiang (2006) that most have nonzero but very small influence. Specifically, parameterizing the true model P_0 in our tensor form with $k_j = d_j$ for $j = 1, \ldots, p_n$ (this is always possible for any P_0), we assume:

Assumption B. $D_n \log(1/\epsilon_n) \prec n\epsilon_n^2$ and there exists a multirank sequence

$k^{(1)}, k^{(2)}, \ldots$ with $|k^{(n)}| \leqslant D_n$, such that

$$\sum_{j=1}^{p_n} \max_{x_j} \sum_{h_j > k_j^{(n)}}^{d_j} \pi_{h_j}^{(j)}(x_j) \prec \epsilon_n^2,$$

where $f \prec g$ means $f/g \to 0$ as $n \to \infty$.

This is a near low rank restriction on P_0. This assumption intuitively means that the true tensor P_0 could be approximated within error ϵ_n^2 by a truncated tensor with multirank $k^{(n)}$, whose size is less than $n\epsilon_n^2/\log(1/\epsilon_n)$. Theoretically, a lower bound of ϵ_n attributed to the low rank approximation could be identified as the minimum ϵ such that

$$\exists \text{ multirank } k,\ s.t. |k| < n\epsilon^2/\log(1/\epsilon) \text{ and } \sum_{j=1}^{p_n} \max_{x_j} \sum_{h_j > k_j}^{d_j} \pi_{h_j}^{(j)}(x_j) \leqslant \epsilon^2.$$

The overall ϵ_n will be the minimum of this lower bound and the one determined by Assumption A. Assumption B includes the special case when P_0 is exactly of low multirank $k^{(0)}$. In such case, all $k^{(n)}$ could be chosen as $k^{(0)}$ and Assumption B puts no constraint on ϵ_n, leading the convergence rate entirely determined by the variable selection in Assumption A as $\sqrt{\log p_n/n}$ (Corollary 6 below). In section 6 of real data applications, we will provide empirical evidence of this near low multirank assumption.

The last assumption can be considered as a regularity condition.

Assumption C. $P_0(y|x) \geqslant \epsilon_0$ for any x, y for some $\epsilon_0 > 0$.

Under this assumption, the Kullback-Leibler divergence would be bounded by the sup norm up to a constant, where the latter is easier to characterize in case of our model.

The next theorem states the posterior contraction rate under our prior (4.10)-(4.11) and Assumption A-C. Recall that r_n is the hyperparameters in the prior.

Theorem 20. *Assume the design points $x_1, \ldots, x_n$ are independent observations from an unknown probability distribution G_n on $\{1, \ldots, d\}^{p_n}$. Moreover, assume the prior is specified as in (4.10)-(4.11). Let ϵ_n be a sequence with $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$ and $\sum_n \exp(-n\epsilon_n^2) < \infty$, with which Assumptions A, B and C hold. Denote $d(P, P_0) = \int \sum_{y=1}^{d_0} \big|P(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)\big| G_n(dx_1, \ldots, dx_p)$, then*

$$\Pi_n\big\{P : d(P, P_0) \geqslant M\epsilon_n | y^n, X^n\big\} \to 0 \ a.s. P_0^n,$$

where $\Pi_n(A|y^n, X^n)$ stands for the posterior probability of A given the observations.

The following corollary tells us that the posterior convergence rate of our model can be very close to $n^{-1/2}$ under appropriate near low rank conditions.

Corollary 21. *For $\alpha \in (0, 1)$, $\epsilon_n = n^{-(1-\alpha)/2} \log n$ will satisfy the conditions in Theorem 20 if $M_n \prec n^\alpha \log n$, $p_n \prec \exp(n^\alpha / \log n)$ and there exists a sequence of multiranks $k^{(n)}$ with size at most M_n such that*

$$\sum_{j=1}^{p_n} \max_{x_j} \sum_{h_j > k_j^{(n)}}^{d_j} \pi_{h_j}^{(j)}(x_j) \prec n^{-(1-\alpha)} \log^2 n.$$

As mentioned after Assumption B, if the truth is exactly lower multirank, then with a small modification to the proof of Theorem 20, we can eliminate the $\log D_n$ factor in Assumption A and leading to the following result.

Corollary 22. *If the truth P_0 has multirank k with a finite number of components $k_j > 1$, then with the choice of M_n to be a sufficiently large fixed number, the posterior convergence rate ϵ_n could be at least $\sqrt{\log p_n / n}$.*

Since $(d_0 - 1)M_n$ could be interpreted as the maximum effective number of parameters in the model, which should be at most the same order as the sample size n, we suggest to set $M_n = n$ as a default for the prior defined in section 3.1 to conceptually

provide as loose an *a priori* upper bound as possible. Results tend to be robust to the choice of M_n as long as it is not chosen to be small. Since $M \geqslant |k| \geqslant 2^{\sharp\{j:k_j>1\}}$, the maximum number of predictors included in the model is $\log_2 n$. This suggests that we can choose $(\log_2 n)/2 = \log_4 n$ as a default value for r in the prior.

4.4 Posterior Computation

In section 4.1, we consider fixed $k = (k_1, \ldots, k_p)'$ and use a Gibbs sampler to draw posterior samples. Generalizing this Gibbs sampler, we developed a reversible jump Markov Chain Monte Carlo (RJMCMC) algorithm (Green, 1995) to draw posterior samples from the joint distribution of $k = \{k_j : j = 1, \ldots, p\}$ and (Λ, π, z). However, for n and p equal to several hundred or more, we were unable to design an RJMCMC algorithm that was sufficiently efficient to be used routinely. Hence, in section 4.2, we propose a faster two stage procedure based on approximated marginal likelihood.

4.4.1 Gibbs sampling for fixed k

Under (4.10) the full conditional posterior distributions of Λ, π and z all have simple forms, which we sample from as follows.

1. For $h_j = 1, \ldots, k_j, j = 1, \ldots, p$, update $\lambda_{h_1,\ldots,h_p}$ from the Dirichlet conditional,

$$\{\lambda_{h_1,\ldots,h_p}(1), \ldots, \lambda_{h_1,\ldots,h_p}(d)\}|- \sim \text{Diri}\bigg(\frac{1}{d} + \sum_{i=1}^{n} 1(z_{i1} = h_1, \ldots, z_{ip} = h_p, y_i = 1), \ldots, \frac{1}{d} + \sum_{i=1}^{n} 1(z_{i1} = h_1, \ldots, z_{ip} = h_p, y_i = d)\bigg).$$

2. Update $\pi^{(j)}(k)$ from the Dirichlet full conditional posterior distribution,

$$\{\pi_1^{(j)}(k), \ldots, \pi_{k_j}^{(j)}(k)\}|- \sim \text{Diri}\bigg(\frac{1}{k_j} + \sum_{i=1}^{n} 1(z_{ij} = 1)1(x_{ij} = k), \ldots, \frac{1}{k_j} + \sum_{i=1}^{n} 1(z_{ij} = k_j)1(x_{ij} = k)\bigg).$$

3. Update z_{ij} from the multinomial full conditional posterior, with

$$P(z_{ij} = h|-) \propto \pi_h^{(j)}(x_{ij}) \lambda_{z_{i,1},\ldots,z_{i,j-1},h,z_{i,j+1},\ldots,z_{i,p}}(y_i).$$

4.4.2 Two step approximation

We propose a two stage algorithm, which identifies a good model in the first stage and then learns the posterior distribution for this model in a second stage via the Gibbs sampler of section 4.1. We first propose an approximation to the marginal likelihood. For simplicity in exposition, we focus on binary Y with $d_0 = 2$, but the approach generalizes in a straightforward manner, with the beta functions in the below expression for the marginal likelihood replaced with functions of the form $\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_{d_0})/\Gamma(a_1 + \cdots + a_{d_0})$. To motivate our approach, we first note that $\pi_{h_j}^{(j)}(x_j)$ can be viewed as providing a type of *soft* clustering of the jth feature X_j, controlling borrowing of information among probabilities conditional on combinations of predictors. To obtain approximated marginal likelihoods to be used only in the initial model selection stage, we propose to force $\pi_{h_j}^{(j)}(x_j)$ to be either zero or one, corresponding to a hard clustering of the predictors. The example in Section 3.2 gives a heuristic argument on the variance-bias tradeoff by using the degenerate approximation. Under this approximation, the marginal likelihood has a simple expression.

For a given model indexed by $k = \{k_j, j = 1, \ldots, p\}$, we assume that the levels of X_j are clustered into k_j groups $A_1^{(j)}, \ldots, A_{k_j}^{(j)}$. For example, with levels $\{1, 2, 3, 4, 5\}$, $A_1^{(j)} = \{1, 2, 3\}$ and $A_2^{(j)} = \{4, 5\}$. Then it is easy to see that the marginal likelihood conditional on k and A is $\mathcal{L}(y|k, A) =$

$$\prod_{h_1,\ldots,h_p} \frac{1}{\text{Beta}(1/2, 1/2)} \text{Beta}\Bigg(\frac{1}{2} + \sum_{i=1}^{n} I(x_{i1} \in A_{h_1}^{(1)}, \ldots, x_{ip} \in A_{h_p}^{(p)}, y_i = 1),$$

$$\frac{1}{2} + \sum_{i=1}^{n} I(x_{i1} \in A_{h_1}^{(1)}, \ldots, x_{ip} \in A_{h_p}^{(p)}, y_i = 0)\Bigg).$$

Having an expression for the marginal likelihood, we apply a stochastic search MCMC algorithm (George and McCulloch, 1997) to obtain samples of $(k_1, \ldots, k_p)$ from the approximated posterior distribution. This proceeds as follows.

1. For $j = 1$ to p, do the following. Given the current model indexed by $k = \{k_j : j = 1, \ldots, p\}$ and clusters $A = \{A_h^{(j)} : h = 1, \ldots, k_j, j = 1, \ldots, p\}$, propose to increase k_j to $k_j + 1$ (if $k_j < d$) or reduce it to $k_j - 1$ (if $k_j > 1$) with equal probability.
2. If increase, randomly split a cluster of X_j into two clusters (all splits have equal probability). For example, if $d_j = 5$, $k_j = 2$ and the levels of X_j are clustered as $\{1, 2, 3\}$ and $\{4, 5\}$. There are 4 possible splitting schemes: three ways to split $\{1, 2, 3\}$ and one way to split $\{4, 5\}$. We randomly choose one. Accept this move with acceptance rate based on the approximated marginal likelihood.
3. If decrease, randomly merge two clusters and accept or reject this move.
4. If k_j remains 1, propose an additional switching step that switches k_j with a currently "active predictor" j' whose $k_{j'} > 1$ and randomly divide the cluster of X_j into $k_{j'}$ clusters.

Estimating approximated marginal inclusion probabilities of $k_j > 1$ based on this algorithm, we keep predictors having inclusion probabilities great than 0.5; this leads to selecting the median probability model, which in simpler settings has been shown to have optimality properties in terms of predictive performance (Barbieri and Berger, 2004).

4.5 Simulation Studies

To assess the performance of the proposed approach, we conducted a simulation study and calculated the misclassification rate on the testing samples. Each simulated

dataset consisted of $N = 3,000$ instances with p of the covariates $X_1, \ldots, X_p$, each of which has $d = 4$ levels, and a binary response Y. Two scenarios were considered: moderate dimension setting where $p = 3$, 4, 5 and high dimension setting where $p = 20$, 100, 500. Note that although $p = 20$ appears less than the training size n, the effective number of parameters is equal to 4^{20}. Similarly, we can call $p = 3$ moderate since the effective number of parameters is equal to $4^3 = 64$. Fixing p, four training sizes $n = 200$, 400, 600 and 800 were considered. In the moderate (high) dimension settings, 100 (10) datasets were simulated for each combination of training size n and covariate dimension p. We assumed that the true model had three important predictors X_1, X_2 and X_3, and generated $P(Y = 1|X_1 = x_1, X_2 = x_2, X_3 = x_3)$ independently for each combination of (x_1, x_2, x_3); this was done once for each simulation replicate prior to generating the data conditionally on $P(Y|X)$. To obtain an average Bayes error rate (optimal misclassification rate) around 15% (standard deviation is around 2%), we generated the conditional probabilities from $f(U) = U^2/\{U^2 + (1 - U)^2\}$, where $U \sim \text{Unif}(0, 1)$. For each dataset, we randomly chose n samples as training with the remaining $N - n$ as testing. We implemented the two stage algorithm on the training set and calculated the misclassification rate on the testing set.

According to our theoretical results, we chose $r = \lceil \log_4 n \rceil$ as the expected number of important predictors in the prior and $M = \log n$ as the maximum model size, where $\lceil x \rceil$ stands for the minimal integer greater equal than x. We ran 1,000 iterations for the first stage and 2,000 iterations for the second stage, treating the first half as burn-in. We compared the results applied to the same training-test split data with classification and regression trees (CART), random forests (RF) (Breiman, 2001), neural networks (NN) with two layers of hidden units, lasso penalized logistic regression (LASSO) (Park and Hastie, 2007), support vector machines (SVM) and Bayesian additive regression trees (BART) (Chipman et al., 2010). All these models

were fitted by using existed R-packages. The penalizing regularization parameter for LASSO was chosen by cross validation. The tunable parameters for other methods were chosen by their default settings. In the moderate dimension scenario, we enumerated all orders of interactions as input covariates for NN, LASSO and SVM. NN was not implemented for $p = 5$ since the available R code was unable to fit the model with $4^5 = 1024$ covariates. In the high dimension scenario, since the number of interactions grows exponentially fast, we only included $(d-1) \times p$ dummy variables for the main effects as input covariates for NN, LASSO and SVM. Under $p = 5(500)$ and $n = 800(800)$, the first stage of our algorithm took about 1s(2s) to draw 40(1) iterations and the second stage took about 1s(1s) to draw 50(50) iterations in matlab. The sampler was quite efficient, with a burn-in of 100 iterations in the first stage and 200 iterations in the second stage sufficient and autocorrelations rapidly decreasing to zero with increasing lag time.

Table 1 displays the results under moderate dimension settings. When $p = 3$, the effective number $4^3 = 64$ of parameters is much smaller than the sample size, resulting in the good performances of all methods, among which LASSO was the best under $n = 200$ and 400. Nevertheless, our method had a rapid decreasing misclassification rate and achieved comparable performance to the best competitors when $n = 400$ and 600. As p increases to 4 and 5, irrelevant covariates are included. As can be seen from table 1, the best methods under $p = 3$, including NN, LASSO and SVM, had noticeably worse performance than our method and RF. Especially, it was interesting that RF had better performance under $p = 4$ and 5 than under $p = 3$. We guess that when all covariates were important, RF tended to overfit the model and lead to poor classification performance on the test samples. Nonetheless, our methods still had the best performance and tended to be robust to the inclusion of irrelevant covariates.

Table 4.1: Simulation study results for moderate dimension case. RF: random forests, NN: neural networks, SVM: support vector machine, BART: Bayesian additive regression trees, TF: Our tensor factorization model. Misclassification rates and their standard deviations over 100 simulations are displayed.

		$n = 200$	$n = 400$	$n = 600$	$n = 800$
$p = 3$	CART	0.371(0.056)	0.357(0.066)	0.341(0.072)	0.335(0.064)
	RF	0.277(0.034)	0.254(0.039)	0.243(0.034)	0.235(0.032)
	NN	0.212(0.033)	0.188(0.038)	0.181(0.043)	0.175(0.037)
	LASSO	**0.206**(0.031)	**0.178**(0.027)	**0.169**(0.023)	**0.167**(0.021)
	SVM	0.320(0.065)	0.195(0.065)	**0.168**(0.023)	**0.167**(0.026)
	BART	0.354(0.044)	0.311(0.041)	0.279(0.036)	0.266(0.036)
	TF	0.243(0.041)	**0.181**(0.031)	**0.168**(0.023)	**0.165**(0.021)
$p = 4$	CART	0.376(0.055)	0.360(0.066)	0.342(0.072)	0.336(0.071)
	RF	0.278(0.028)	0.223(0.029)	0.195(0.025)	0.189(0.026)
	NN	0.353(0.044)	0.266(0.039)	0.235(0.039)	0.223(0.037)
	LASSO	0.323(0.036)	0.256(0.030)	0.219(0.025)	0.201(0.023)
	SVM	0.325(0.032)	0.257(0.024)	0.219(0.025)	0.202(0.023)
	BART	0.378(0.042)	0.329(0.041)	0.282(0.035)	0.269(0.034)
	TF	**0.241**(0.041)	**0.183**(0.031)	**0.170**(0.023)	**0.164**(0.021)
$p = 5$	CART	0.384(0.054)	0.364(0.067)	0.342(0.071)	0.342(0.063)
	RF	0.324(0.031)	0.267(0.031)	0.230(0.028)	0.218(0.063)
	NN	-	-	-	-
	LASSO	0.415(0.046)	0.366(0.048)	0.314(0.032)	0.298(0.025)
	SVM	0.414(0.042)	0.374(0.036)	0.335(0.029)	0.306(0.029)
	BART	0.395(0.027)	0.353(0.036)	0.335(0.031)	0.306(0.029)
	TF	**0.242**(0.042)	**0.184**(0.031)	**0.168**(0.022)	**0.164**(0.022)

Table 2 displays the results under high dimension settings. The differences become more perceptible. All the competing methods broke down and had worse performance than TF. In the very challenging case in which the training sample size was only 200 and $p = 500$, all methods had poor performance. However, as the training sample size increased, the proposed conditional tensor factorization method rapidly approached the optimal 15%, with excellent performance even in the $n = 600$, $p = 500$ case. In contrast, the competitive methods had consistently poor performance. In this challenging setting involving a low signal strength, a modest sample size, and moderately large numbers of candidate predictors, CART appeared to be

Table 4.2: Simulation study results in the high dimension setting. RF: random forests, NN: neural networks, SVM: support vector machine, BART: Bayesian additive regression trees, TF: Our tensor factorization model. Misclassification rates and their standard deviations over 100 simulations are displayed.

		$n = 200$	$n = 400$	$n = 600$	$n = 800$
$p = 20$	CART	0.448(0.025)	0.367(0.042)	0.342(0.063)	0.337(0.087)
	RF	0.461(0.022)	0.444(0.025)	0.412(0.026)	0.393(0.023)
	NN	0.501(0.009)	0.494(0.008)	0.507(0.043)	0.482(0.021)
	LASSO	0.440(0.040)	0.418(0.025)	0.372(0.032)	0.357(0.044)
	SVM	0.503(0.011)	0.485(0.012)	0.494(0.012)	0.472(0.024)
	BART	0.450(0.026)	0.401(0.037)	0.374(0.032)	0.345(0.031)
	TF	**0.249**(0.036)	**0.182**(0.036)	**0.172**(0.026)	**0.162**(0.022)
$p = 100$	CART	0.478(0.023)	0.428(0.042)	0.389(0.046)	0.361(0.052)
	RF	0.468(0.022)	0.472(0.027)	0.433(0.025)	0.421(0.022)
	NN	0.504(0.010)	0.492(0.008)	0.495(0.015)	0.479(0.013)
	LASSO	0.450(0.036)	0.430(0.033)	0.410(0.042)	0.404(0.032)
	SVM	0.507(0.011)	0.483(0.011)	0.490(0.013)	0.463(0.024)
	BART	0.465(0.017)	0.450(0.024)	0.410(0.013)	0.404(0.032)
	TF	**0.323**(0.120)	**0.179**(0.027)	**0.169**(0.021)	**0.164**(0.024)
$p = 500$	CART	0.489(0.09)	0.461(0.048)	0.404(0.032)	0.380(0.080)
	RF	0.480(0.023)	0.468(0.020)	0.446(0.028)	0.434(0.019)
	NN	0.496(0.013)	0.488(0.021)	0.466(0.028)	0.446(0.019)
	LASSO	**0.459**(0.012)	0.466(0.025)	0.392(0.020)	0.419(0.016)
	SVM	0.492(0.016)	0.493(0.021)	0.482(0.017)	0.468(0.016)
	BART	0.475(0.013)	0.466(0.025)	0.427(0.027)	0.431(0.015)
	TF	**0.454**(0.105)	**0.205**(0.083)	**0.173**(0.022)	**0.164**(0.021)

the best competing method.

In addition to the clearly superior classification performance, our method had the advantage of providing variable selection results. Table 3 provides the average approximated marginal inclusion probabilities for the three important predictors and remaining predictors in the high dimension settings. Consistently with the results in Table 2, the method fails to detect the important predictors when $p = 500$ and the training sample size is only $n = 200$. But as the sample size increases appropriately, TF assigns high marginal inclusion probabilities to the important predictors and low ones to the unimportant predictors. In addition, to access the fitting performances,

Table 4.3: Simulation study variable selection results in the high dimensional case. Rows 1-3 within each fixed p are approximated inclusion probabilities of the 1st,2nd,3rd predictors. Max is the maximum inclusion probability across the remaining predictors. Ave is the average inclusion probability across the remaining predictors. These quantities are averages over 10 trials.

		$n = 200$	$n = 400$	$n = 600$	$n = 800$
$p = 20$	X_1	1.00	1.00	1.00	1.00
	X_2	1.00	1.00	1.00	1.00
	X_3	1.00	1.00	1.00	1.00
	Max	0.00	0.00	0.00	0.00
	Ave	0.00	0.00	0.00	0.00
	aMSE	0.074(0.013)	0.025(0.005)	0.014(0.004)	0.009(0.002)
$p = 100$	X_1	0.74	1.00	1.00	1.00
	X_2	0.70	1.00	1.00	1.00
	X_3	0.72	1.00	1.00	1.00
	Max	0.21	0.00	0.00	0.00
	Ave	0.01	0.00	0.00	0.00
	aMSE	0.089(0.026)	0.027(0.003)	0.014(0.002)	0.009(0.002)
$p = 500$	X_1	0.23	0.91	1.00	1.00
	X_2	0.24	0.90	1.00	1.00
	X_3	0.21	0.91	1.00	1.00
	Max	0.28	0.07	0.00	0.00
	Ave	0.00	0.00	0.00	0.00
	aMSE	0.134(0.034)	0.036(0.037)	0.014(0.003)	0.009(0.002)

we calculated the empirical average MSE defined as

$$\text{aMSE} = \frac{1}{N}\sum_{i=1}^{N}\left\{P(Y=1|x_{i1},\ldots,x_{ip}) - \hat{P}(Y=1|x_{i1},\ldots,x_{ip})\right\}^2,$$

where $(x_{i1},\ldots,x_{ip})$ is the vector of covariates of the ith sample and $\hat{P}$ is the fitted conditional probability. The aMSE approached to zero rapidly as testing size increased and tended to be robust to the covariate dimension as long as the method could identify the important predictors.

4.6 Applications

We compare our method with other competing methods in three data sets from the UCI repository. The first data set is Promoter Gene Sequences (abbreviated as promoter data below). The data consists of A, C, G, T nucleotides at $p = 57$ positions for $N = 106$ sequences and a binary response indicating instances of promoters and non-promoters. We use 5-fold cross validation with $n = 85$ training samples and $N - n = 21$ test samples in each training-test split.

The second data set is the Splice-junction Gene Sequences (abbreviated as splice data below). These data consist of A, C, G, T nucleotides at $p = 60$ positions for $N = 3,175$ sequences. Each sequence belongs to one of the three classes: exon/intron boundary (EI), intron/exon boundary (IE) or neither (N). Since its sample size is much larger than the first data set, we compare our approach with competing methods in two scenarios: a small sample size and a moderate sample size. In the small sample size case, each time we randomly select $n = 200$ instances as training and calculate the misclassification rate on the testing set composed of the remaining $2,975$ instances. We repeat this for each method for five training-test splits and report the average misclassification rate. In the moderate sample size case, we use 5-fold cross validation so that each time $n = 2,540$ instances are treated as training data.

The third data set describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) has 22 binary feature patterns. This data set has been previously divided into a training set of size 80 and a testing set of size 187.

We considered the same competitors as those in the simulation part. Among them, BART was not implemented in the splice data since we were unable to find a multi-class implementation of their approach.

Table 4.4: UCI Data Example. RF: random forests, NN: neural networks, SVM: support vector machine, BART: Bayesian additive regression trees, TF: Our tensor factorization model. Misclassification rates are displayed.

Data	CART	RF	NN	LASSO	SVM	BART	TF
Promoter (n=85)	0.236	**0.066**	0.170	0.075	0.151	0.113	**0.066**
Splice (n=200)	0.161	0.122	0.226	0.141	0.286	-	**0.112**
Splice (n=2540)	0.059	**0.046**	0.165	0.123	0.059	-	0.058
SPECT (n=80)	0.312	0.235	0.278	0.277	0.246	0.225	**0.198**

Table 4 shows the results. Our method produced at worst comparable classification accuracy to the best of the competitors in each of the cases considered. Among the competitors, Random Forests (RF) provided the best competitor overall, which is consistent with our previous experiment under high dimensional settings. We expect our approach to do particularly well when there is a modest training sample size and high-dimensional predictors. We additionally have an advantage in terms of interpretability over several of these approaches, including RF and BART, in conducting variable selection.

Table 5 displays the selected variables along with their associated mode ranks. As can been seen, in the promoter data and splice data, nearby nucleotide sequences are selected. These results are reasonable since for nucleotide sequences, nearby nucleotides form a motif regulating important functions. For the splice data, the number of variables selected by our model increases from 4 under $n = 200$ to 6 under $n = 200$. This gradually increase in the model size suggests that the splice data may possess a near low multirank structure characterized by Assumption B, where the optimal number of selected variables is determined by the bias-variance tradeoff. As the training size further grows, more important variables would be selected into the model. In the contrast, the number of selected variables in the SPECT data remains the same as the training size grows, suggesting that an exact low multirank assumption maybe valid. It is notable that in each of these cases we obtained excel-

Table 4.5: Variable selection results. The selected variables are displayed, with their associated mode ranks k_j's included in the parenthesis.

	Important variables selected
Promoter (n=106)	15th(2), 16th(2), 17th(3), 39th(3)
Splice (n=200)	29th(2), 30th(2), 31st(2), 32nd(2)
Splice (n=2540)	28th(2), 29th(2), 30th(2), 31st(2), 32nd(2), 35th(2)
SPECT (n=80)	11th(2), 13th(2), 16th(2)
SPECT (n=267)	11th(2), 13th(2), 16th(2)

lent classification performance based on a small subset of the predictors. Moreover, for the nucleotide sequences data, most selected variables have low mode ranks k_j comparing to the full size $d_j = 4$. Therefore, these variable selection results provide empirical verifications of the near low multirank assumption B in section 3.2.

4.7 Discussion

This article proposes a framework for nonparametric Bayesian classification relying on a novel class of conditional tensor factorizations. The nonparametric Bayes framework is appealing in facilitating variable selection and uncertainty about the core tensor dimensions in the Tucker-type factorization, while avoiding the need for parameter tuning. In particular, we have recommended a single default prior setting that can be used in general applications without relying on cross-validation or other approaches for estimating tuning parameters. One of our major contributions is the strong theoretical support we provide for our proposed approach. Although it has been commonly observed that Bayesian parametric and nonparametric methods have practical gains in numerous applications, there is a clear lack of theory supporting these empirical gains.

Interesting ongoing directions include developing faster approximation algorithms and generalizing the conditional tensor factorization model to accommodate broader feature modalities. In the fast algorithms direction, online variational methods (Hoff-

man et al., 2010) provide a promising direction. Regarding generalizations, we can potentially accommodate continuous predictors and more complex *object* predictors (text, images, curves, etc) through probabilistic clustering of the predictors in a first stage, with X_j then corresponding to the cluster index for feature j.

5

Minimax optimal Bayesian aggregation

5.1 Introduction

In many applications, it is not at all clear how to pick one most suitable method out of a list of possible models or learning algorithms $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. Each model/algorithm has its own set of implicit or explicit assumptions under which that approach will obtain at or near optimal performance. However, in practice verifying which if any of these assumptions hold for a real application is problematic. Hence, it is of substantial practical importance to have an aggregating mechanism that can automatically combine the estimators $\hat{f}_1, \ldots, \hat{f}_M$ obtained from the M different approaches $\mathcal{M}_1, \ldots, \mathcal{M}_M$, with the aggregated estimator potentially better than any single one.

Towards this goal, three main aggregation strategies receive most attention in the literature: model selection aggregation (MSA), convex aggregation (CA) and linear aggregation (LA), as first stated by Nemirovski (2000). MSA aims at selecting the optimal single estimator from the list; CA considers searching for the optimal convex combination of the estimators; and LA focuses on selecting the optimal linear

combination. Although there is an extensive literature (Juditsky and Nemirovski, 2000; Tsybakov, 2003; Wegkamp, 2003; Yang, 2000, 2001, 2004; Bunea and Nobel, 2008; Bunea and Tsybakov, 2007; Guedj and Alquier, 2013; van der Laan et al., 2007) on aggregation, there has been limited consideration of Bayesian approaches.

In this chapter, we study Bayesian aggregation procedures and their performance in regression. Consider the regression model

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{5.1}$$

where Y_i is the response variable, $f : \mathcal{X} \to \mathbb{R}$ is an unknown regression function, $\mathcal{X}$ is the feature space, X_i's are the fixed- or random-designed elements in $\mathcal{X}$ and the errors are iid Gaussian.

Aggregation procedures typically start with randomly dividing the sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ into a training set for constructing estimators $\hat{f}_1, \ldots, \hat{f}_M$, and a learning set for constructing $\hat{f}$. Our primary interest is in the aggregation step, so we adopt the convention (Bunea and Tsybakov, 2007) of fixing the training set and treating the estimators $\hat{f}_1, \ldots, \hat{f}_M$ as fixed functions $f_1, \ldots, f_M$. Our results can also be translated to the context where the fixed functions $f_1, \ldots, f_M$ are considered as a functional basis (Juditsky and Nemirovski, 2000), either orthonormal or overcomplete, or as "weak learners" (van der Laan et al., 2007). For example, high-dimensional linear regression is a special case of LA where f_j maps an M-dimensional vector into its jth component.

Bayesian model averaging (BMA) (Hoeting et al., 1999) provides an approach for aggregation, placing a prior over the ensemble and then updating using available data to obtain posterior model probabilities. For BMA, $\hat{f}$ can be constructed as a convex combination of estimates $\hat{f}_1, \ldots, \hat{f}_M$ obtained under each model, with weights corresponding to the posterior model probabilities. If the true data generating model f_0 is one of the models in the pre-specified list ("$\mathcal{M}$-closed" view), then as the

sample size increases the weight on f_0 will typically converge to one. With a uniform prior over $\mathcal{M}$ in the regression setting with Gaussian noise, $\hat{f}$ coincides with the exponentially weighted aggregates (Tsybakov, 2003). However, BMA relies on the assumption that $\mathcal{M}$ contains the true model. If this assumption is violated ("$\mathcal{M}$-open"), then $\hat{f}$ tends to converge to the single model in $\mathcal{M}$ that is closest to the true model in Kullback-Leibler (KL) divergence. For example, when f_0 is a weighted average of f_1 and f_2, under our regression setting $\hat{f}$ will converge to $f \in \{f_1, f_2\}$ that minimizes $||f - f_0||_n^2 = n^{-1} \sum_{i=1}^n |f(X_i) - f_0(X_i)|^2$ under fixed design or $||f - f_0||_Q^2 = E_Q|f(X) - f_0(X)|^2$ under random design where $X \sim Q$. Henceforth, we use the notation $||\cdot||$ to denote $||\cdot||_n$ or $||\cdot||_Q$ depending on the context.

In this chapter, we primarily focus on Bayesian procedures for CA and LA. Let

$$\mathcal{F}^H = \Big\{ f_\lambda = \sum_{j=1}^M \lambda_j f_j : \lambda = (\lambda_1, \ldots, \lambda_M) \in H \Big\}$$

be the space of all aggregated estimators for f_0 with index set H. For CA, H takes the form of $\Lambda = \{(\lambda_1, \ldots, \lambda_M) : \lambda_j \geqslant 0, j = 1, \ldots, M, \ \sum_{j=1}^M \lambda_j = 1\}$ and for LA, $H = \Omega = \{(\lambda_1, \ldots, \lambda_M) : \lambda_j \in \mathbb{R}, j = 1, \ldots, M, \ \sum_{j=1}^M |\lambda_j| \leqslant L\}$, where $L > 0$ can be unknown but is finite. In addition, for both CA and LA we consider sparse aggregation with $\mathcal{F}^{H_s}$, where an extra sparsity structure $||\lambda||_0 = s$ is imposed on the weight $\lambda \in H_s = \{\lambda \in H : ||\lambda||_0 = s\}$. Here, for a vector $\theta \in \mathbb{R}^M$, we use $||\theta||_p = (\sum_{j=1}^M |\theta_j|^p)^{1/p}$ to denotes its l_p-norm for $0 \leqslant p \leqslant \infty$. In particular, $||\theta||_0$ is the number of nonzero components of θ. The sparsity level s is allowed to be unknown and expected to be learned from data. In the sequel, we use the notation $f_{\lambda*}$ to denote the best $||\cdot||$-approximation of f_0 in $\mathcal{F}^H$. Note that if $f_0 \in \mathcal{F}^H$, then $f_0 = f_{\lambda*}$.

One primary contribution of this work is to propose a new class of priors, called Dirichlet aggregation (DA) priors, for Bayesian aggregation. Bayesian approaches

with DA priors are shown to lead to the minimax optimal posterior convergence rate over $\mathcal{F}^H$ for CA and LA, respectively. More interestingly, DA is able to achieve the minimax rate of sparse aggregation (see Section 5.1.1), which improves the minimax rate of aggregation by utilizing the extra sparsity structure on λ^*. This suggests that DA is able to automatically adapt to the unknown sparsity structure when it exists but also has optimal performance in the absence of sparsity. Such sparsity adaptive properties have also been observed in Bunea and Tsybakov (2007) for penalized optimization methods. However, in order to achieve minimax optimality, the penalty term, which depends on either the true sparsity level s or a function of λ^*, needs to be tuned properly. In contrast, the DA does not require any prior knowledge on λ^* and is tuning free.

Secondly, we also consider an "M-open" view for CA and LA, where the truth f_0 can not only fall outside the list $\mathcal{M}$, but also outside the space of all convex/linear combinations of the models in $\mathcal{M}$. Under the "M-open" view, our theory suggests that the posterior measure tends to put all its mass into a ball around the best approximation $f_{\lambda*}$ of f_0 with a radius proportional to the minimax rate. The metric that defines that ball will be made clear later. This is practically important because the true model in reality is seldom correctly specified and a convergence to $f_{\lambda*}$ is the best one can hope for. Bayesian asymptotic theory for misspecified models is under developed, with most existing results assuming that the model class is either known or is an element of a known list. One key step is to construct appropriate statistical tests discriminating $f_{\lambda*}$ from other elements in $\mathcal{F}^H$. Our tests borrow some results from Kleijn and van der Vaart (2006) and rely on concentration inequalities.

The proposed prior on λ induces a novel shrinkage structure, which is of independent interest. There is a rich literature on theoretically optimal models based on discrete (point mass mixture) priors (Ishwaran and Rao, 2005; Castillo and van der Vaart, 2012) that are supported on a combinatorial model space, leading to heavy

computational burden. However, continuous shrinkage priors avoid stochastic search variable selection algorithms (George and McCulloch, 1997) to sample from the combinatorial model space and can potentially improve computational efficiency. Furthermore, our results include a rigorous investigation on M-dimensional symmetric Dirichlet distributions, Diri$(\rho, \ldots, \rho)$ when $M \gg 1$ and $\rho \ll 1$. Here Diri$(\alpha_1, \ldots, \alpha_M)$ denotes a Dirichlet distribution with concentration parameters $\alpha_1, \ldots, \alpha_M$. In machine learning, Diri$(\rho, \ldots, \rho)$ with $\rho \ll 1$ are widely used as priors for latent class probabilities (Blei et al., 2003). However, little rigorous theory has been developed for the relationship between its concentration property and the hyperparameter ρ. Rousseau and Mengersen (2011) consider a related problem of overfitted mixture models and show that generally the posterior distribution effectively empties the extra components. However, our emphasis is to study the prediction performance instead of model selection. Moreover, in Rousseau and Mengersen (2011) the number M of components is assumed to be fixed as n increases, while in our setting we allow M to grow in the order of $e^{o(n)}$. In this large-M situation, the general prior considered in Rousseau and Mengersen (2011) is unable to empty the extra components and we need to impose sparsity. In this chapter, we show that if we choose $\rho \sim M^{-\gamma}$ with $\gamma > 1$, then Diri$(\rho, \ldots, \rho)$ could lead to the optimal concentration rate for sparse weights (Section 5.2.1). Moreover, such concentration is shown to be adaptive to the sparsity level s.

The rest of the chapter is organized as follows. In Section 1.1, we review the minimax results for aggregation. In Section 2, we describe the new class of priors for CA and LA based on symmetric Dirichlet distributions. In Section 3, we study the asymptotic properties of the proposed Bayesian methods. In Section 4, we show some simulations and applications. The proofs of the main theorems appear in Section 5 and some technical proofs are deferred to Section 6. We provide details of the MCMC implementation of our Bayesian aggregation methods in the appendix.

5.1.1 *A brief review of the minimax risks for aggregation*

It is known (Tsybakov, 2003) that for CA, the minimax risk for estimating the best convex combination $f_{\lambda *}$ within $\mathcal{F}^{\Lambda}$ is

$$\sup_{f_1,\ldots,f_M \in \mathcal{F}_0} \inf_{\hat{f}} \sup_{f_\lambda^* \in \mathcal{F}^{\Lambda}} E||\hat{f} - f_\lambda^*||^2 \asymp \begin{cases} M/n, & \text{if } M \leqslant \sqrt{n}, \\ \sqrt{\frac{1}{n}\log\left(M/\sqrt{n}+1\right)}, & \text{if } M > \sqrt{n}, \end{cases} \tag{5.2}$$

where $\mathcal{F}_0 = \left\{f : ||f||_\infty \leqslant 1\right\}$ and $\hat{f}$ ranges over all possible estimators based on n observations. Here, for any two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means that there exists a constant $C > 0$, such that $a_n \leqslant C b_n$ and $b_n \leqslant C a_n$ for any n. The norm is understood as the L_2-norm for random design and the $||\cdot||_n$-norm for fixed design. If we have more information that the truth f_λ^* also possesses a sparse structure $||\lambda^*||_0 \triangleq \#\{j : \lambda_j > 0\} = s \ll n$, then we would expect a faster convergence rate of estimating f_λ^*. For example, in the "M-closed" case where $f_\lambda^* = f_j$ for some $j \in \{1,\ldots,M\}$, $\lambda_i^* = I(i=j)$ and $||\lambda^*||_0 = 1$. Let $\mathcal{F}_s^{\Lambda} = \left\{f = \sum_{j=1}^M \lambda_j f_j : \lambda \in \Lambda, ||\lambda||_0 = s\right\}$ be the space of all s-sparse convex aggregations of $f_1,\ldots,f_M$. By extending the results in Tsybakov (2003), it can be shown that when the sparsity level s satisfies $s \leqslant \sqrt{n/\log M}$, the minimax risk of estimating an element in $\mathcal{F}_s^{\Lambda}$ is given by

$$\sup_{f_1,\ldots,f_M \in \mathcal{F}_0} \inf_{\hat{f}} \sup_{f_\lambda^* \in \mathcal{F}_s^{\Lambda}} E||\hat{f} - f_\lambda^*||^2 \asymp \frac{s}{n}\log\left(\frac{M}{s}\right). \tag{5.3}$$

From the preceding results, $\sqrt{n/\log M}$ serves as the sparsity/non-spasrsity boundary of the weight λ^* as there is no gain in the estimation efficiency if $s > \sqrt{n/\log M}$.

From Tsybakov (2003), the minimax risk for LA with $H = \mathbb{R}^M$ is

$$\sup_{f_1,\ldots,f_M \in \mathcal{F}_0} \inf_{\hat{f}} \sup_{f_\lambda^* \in \mathcal{F}^{\mathbb{R}^M}} E||\hat{f} - f_\lambda^*||^2 \asymp M/n.$$

As a result, general LA is only meaningful when $M/n \to 0$, as $n \to \infty$. Similarly, the above minimax risk can be extended to s-sparse LA $\mathcal{F}_s^{\mathbb{R}^M} = \left\{f = \sum_{j=1}^M \lambda_j f_j : \lambda \in\right.$

$\mathbb{R}^M, ||\lambda||_0 = s\}$ for $s \in \{1, \ldots, M\}$ as

$$\sup_{f_1,\ldots,f_M \in \mathcal{F}_0} \inf_{\hat{f}} \sup_{f_\lambda^* \in \mathcal{F}_s^{\mathbb{R}^M}} E||\hat{f} - f_\lambda^*||^2 \asymp \frac{s}{n} \log\left(\frac{M}{s}\right).$$

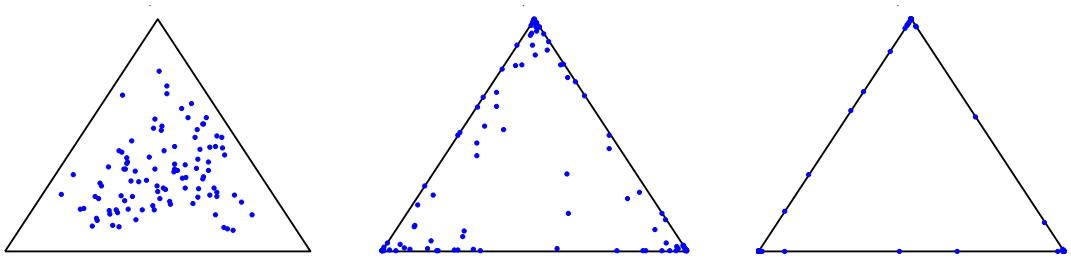
Note that for sparse LA, the sparsity level s can be arbitrary. A simple explanation is that the constraint $||\lambda^*||_1 = 1$ ensures that every element in $\mathcal{F}^\Lambda$ can be approximated with error at most $\sqrt{\frac{1}{n}\log\left(M/\sqrt{n}+1\right)}$ by some $\sqrt{n/\log M}$-sparse element in $\mathcal{F}^\Lambda$ (see Lemma 82). However, if we further assume that $||\lambda^*|| \leqslant A$ and restrict $f^{\lambda^*} \in \mathcal{F}^\Omega$, then by extending Tsybakov (2003), it can be shown that the minimax risks of LA of $\mathcal{F}_A^{\mathbb{R}^M}$ is the same as those of convex aggregation under a non-sparse structure as (5.2) and a sparse structure as (5.3).

5.2 Bayesian approaches for aggregation

5.2.1 Concentration properties of high dimensional symmetric Dirichlet distributions

Consider an M-dimensional symmetric Dirichlet distribution $\text{Diri}(\rho, \ldots, \rho)$ indexed by a concentration parameter $\rho > 0$, whose pdf at $\lambda \in \Lambda$ is given by $\Gamma(M\rho)\{\Gamma(\rho)\}^{-M}$ $\prod_{j=1}^M \lambda_j^{\rho-1}$, where $\Gamma(\cdot)$ is the Gamma function. M-dimensional Dirichlet distributions are commonly used in Bayesian procedures as priors over the $M-1$-simplex. For example, Dirichlet distributions can be used as priors for probability vectors for latent class allocation. In this subsection, we investigate the concentration properties of $\text{Diri}(\rho, \ldots, \rho)$ when $M \gg 1$ and $\rho \ll 1$. Fig. 5.2.1 displays typical patterns for 3-dimensional Dirichlet distributions $\text{Diri}(\rho, \rho, \rho)$ with ρ changing from moderate to small. As can be seen, the Dirichlet distribution tends to concentrate on the boundaries for small ρ, which is suitable for capturing sparsity structures.

To study the concentration of $\text{Diri}(\rho, \ldots, \rho)$, we need to characterize the space of sparse weight vectors. Since Dirichlet distributions are absolutely continuous, the



(a) $\rho = 1$. (b) $\rho = 0.1$. (c) $\rho = 0.01$.

FIGURE 5.1: Symmetric Dirichlet distributions with different values for the concentration parameter. Each plot displays 100 independent draws from $\text{Diri}(\rho, \rho, \rho)$.

probability of generating an exactly s-sparse vector is zero for any $s < M$. Therefore, we need to relax the definition of s-sparsity. Consider the following set indexed by a tolerance level $\epsilon > 0$ and a sparsity level $s \in \{1, \ldots, M\}$: $\mathcal{F}^{\Lambda}_{s,\epsilon} = \{\lambda \in \Lambda : \sum_{j=s+1}^{M} \lambda_{(j)} \leqslant \epsilon\}$, where $\lambda_{(1)} \geqslant \lambda_{(2)} \geqslant \cdots \geqslant \lambda_{(M)}$ is the ordered sequence of $\lambda_1, \ldots, \lambda_M$. $\mathcal{F}^{\Lambda}_{s,\epsilon}$ consists of all vectors that can be approximated by s-sparse vectors with l_1-error at most ϵ. The following theorem shows the concentration property of the symmetric Dirichlet distribution $\text{Diri}(\rho, \ldots, \rho)$ with $\rho = \alpha/M^{\gamma}$. This theorem is a easy consequence of Lemma 26 and Lemma 29 in Section 5.5.

Theorem 23. *Assume that $\lambda \sim \text{Diri}(\rho, \ldots, \rho)$ with $\rho = \alpha/M^{\gamma}$ and $\gamma > 1$. Let $\lambda^* \in \Lambda_s$ be any s-sparse vector in the $M-1$-dimensional simplex Λ. Then for any $\epsilon \in (0, 1)$ and some $C > 0$,*

$$P(||\lambda - \lambda^*||_2 \leqslant \epsilon) \gtrsim \exp\left\{ - C\gamma s \log \frac{M}{\epsilon} \right\}, \tag{5.4}$$

$$P(\lambda \notin \mathcal{F}^{\Lambda}_{s,\epsilon}) \lesssim \exp\left\{ - C(\gamma - 1) s \log \frac{M}{\epsilon} \right\}. \tag{5.5}$$

The proof of (5.5) utilizes the stick-breaking representation of Dirichlet processes (Sethuraman, 1994) and the fact that $\text{Diri}(\rho, \ldots, \rho)$ can be viewed as the joint distribution of $\big(G([0, 1/M)), \ \ldots, \ G([(M-1)/M, 1))\big)$ where $G \sim$ Dirichlet process $\text{DP}((M\rho)U)$ with U the uniform distribution on $[0, 1]$. The condition $\gamma > 1$ in Theorem 23 reflects the fact that the concentration parameter $M\rho = \alpha M^{-(\gamma-1)}$ should

decrease to 0 as $M \to \infty$ in order for DP$((M\rho)U)$ to favor sparsity. (5.5) validates our observations in Fig. 5.2.1 and (5.4) suggests that the prior mass around every sparse vector is uniformly large since the total number of s-sparse patterns (locations of nonzero components) in Λ is of order $\exp\{Cs\log(M/s)\}$. In fact, both (5.4) and (5.5) play crucial roles in the proofs in Section 5.5.1 on characterizing the posterior convergence rate ϵ_n for the Bayesian method below for CA (also true for more general Bayesian methods), where $\{\epsilon_n\}$ is a sequence satisfying $P(||\lambda - \lambda^*||_2 \leqslant \epsilon_n) \gtrsim \exp(-n\epsilon_n^2)$ and $P(\lambda \notin \mathcal{F}_{s,\epsilon}^{\Lambda}) \lesssim \exp(-n\epsilon_n^2)$. Assume the best approximation f_{λ^*} of the truth f_0 to be s-sparse. (5.5) implies that the posterior distribution of λ tends to put almost all its mass in $\mathcal{F}_{s,\epsilon}^{\Lambda}$ and (5.4) is required for the posterior distribution to be able to concentrate around λ^* at the desired minimax rate given by (5.2).

5.2.2 Using Dirichlet priors for convex aggregation

In this subsection, we assume X_i to be random with distribution Q and $f_0 \in L_2(Q)$. Here, for a probability measure Q on a space $\mathcal{X}$, we use the notation $||\cdot||_Q$ to denote the norm associated with the square integrable function space $L_2(Q) = \{f : \int_{\mathcal{X}} |f(x)|^2 dQ(x) \leqslant \infty\}$. We assume the random design for theoretical convenience and the procedure and theory for CA can also be generalized to fixed design problems. Assume the M functions $f_1, \ldots, f_M$ also belong to $L_2(Q)$. Consider combining these M functions into an aggregated estimator $\hat{f} = \sum_{j=1}^{M} \hat{\lambda}_j f_j$, which tries to estimate f_0 by elements in the space $\mathcal{F}^{\Lambda} = \left\{f = \sum_{j=1}^{M} \lambda_j f_j : \lambda_j \geqslant 0, \sum_{j=1}^{M} \lambda_j = 1\right\}$ of all convex combinations of $f_1, \ldots, f_M$. The assumption that $f_1, \ldots, f_M$ are fixed is reasonable as long as different subsets of samples are used for producing $f_1, \ldots, f_M$ and for aggregation. For example, we can divide the data into two parts and use the first part for estimating $f_1, \ldots, f_M$ and the second part for aggregation.

We propose the following Dirichlet aggregation (DA) prior:

$$\text{(DA)} \qquad f = \sum_{j=1}^{M} \lambda_j f_j,\ (\lambda_1, \ldots, \lambda_M) \sim \text{Diri}\left(\frac{\alpha}{M^\gamma}, \ldots, \frac{\alpha}{M^\gamma}\right),$$

where (γ, α) are two positive hyperparameters. As Theorem 23 and the results in Section 5.5 suggest, such a symmetric Dirichlet distribution is favorable since $\text{Diri}(\alpha_1, \ldots, \alpha_M)$ with equally small parameters $\alpha_1 = \ldots = \alpha_M = \alpha/M^\gamma$ for $\gamma > 1$ has nice concentration properties under both sparse and nonsparse L_1 type conditions, leading to near minimax optimal posterior contraction rate under both scenarios.

We also mention a related paper (Bhattacharya et al., 2013) that uses Dirichlet distributions in high dimensional shrinkage priors, where they considered normal mean estimating problems. They proposed a new class of Dirichlet Laplace priors for sparse problems, with the Dirichlet placed on scaling parameters of Laplace priors for the normal means. Our prior is fundamentally different in using the Dirichlet directly for the weights λ, including a power γ for M. This is natural for aggregation problems, and we show that the proposed prior is simultaneously minimax optimal under both sparse and nonsparse conditions on the weight vector λ as long as $\gamma > 1$.

5.2.3 Using Dirichlet priors for linear aggregation

For LA, we consider a fixed design for $X_i \in \mathbb{R}^d$ and write (5.1) into vector form as $Y = F_0 + \epsilon$, $\epsilon \sim N(0, \sigma^2 I_n)$, where $Y = (Y_1, \ldots, y_n)$ is the $n \times 1$ response vector, $F_0 = (f_0(X_1), \ldots, f_0(X_n))^T$ is the $n \times 1$ vector representing the expectation of Y and I_n is the $n \times n$ identity matrix. Let $F = (F_{ij}) = (f_j(X_x))$ be the $n \times M$ prediction matrix, where the jth column of F consists of all values of f_j evaluated at the training predictors $X_1, \ldots, X_n$. LA estimates F_0 as $F\lambda$ with $\lambda = (\lambda_1, \ldots, \lambda_M)^T \in \mathbb{R}^M$ the $p \times 1$ the coefficient vector. Use the notation F_j to denote the jth column of F and $F^{(i)}$ the ith row. Notice that this framework of linear aggregation includes (high-dimensional) linear models as a special case where $d = M$ and $f_j(X_i) = X_{ij}$.

Let $A = ||\lambda||_1 = \sum_{j=1}^{M} |\lambda_j|$, $\mu = (\mu_1, \ldots, \mu_M) \in \Lambda$ with $\mu_j = |\lambda_j|/A$, $z = (z_1, \ldots, z_M) \in \{-1, 1\}^M$ with $z_j = \text{sgn}(\lambda_j)$. This new parametrization is identifiable and (A, μ, z) uniquely determines λ. Therefore, there exists a one-to-one correspondence between the prior on (A, μ, z) and the prior on λ. Under this parametrization, the geometric properties of λ transfer to those of μ. For example, a prior on μ that induces sparsity will produce a sparse prior for λ. With this in mind, we propose the following double Dirichlet Gamma (DDG) prior for λ or (A, μ, z):

$$\text{(DDG1)} \qquad A \sim \text{Ga}(a_0, b_0),\ \mu \sim \text{Diri}\left(\frac{\alpha}{M^\gamma}, \ldots, \frac{\alpha}{M^\gamma}\right),$$
$$z_1, \ldots, z_M \text{ iid with } P(z_i = 1) = \frac{1}{2}.$$

Since μ follows a Dirichlet distribution, it can be equivalently represented as

$$\left(\frac{T_1}{\sum_{j=1}^{p} T_j}, \ldots, \frac{T_M}{\sum_{j=1}^{p} T_j}\right),\ \text{with } T_j \overset{\text{iid}}{\sim} \text{Ga}\left(\frac{\alpha}{M^\gamma}, 1\right).$$

Let $\eta = (\eta_1, \ldots, \eta_M)$ with $\eta_j = z_j \lambda_j$. By marginalizing out the z, the prior for μ can be equivalently represented as

$$\left(\frac{T_1}{\sum_{j=1}^{M} |T_j|}, \ldots, \frac{T_M}{\sum_{j=1}^{M} |T_j|}\right),\ \text{with } T_j \overset{\text{iid}}{\sim} \text{DG}\left(\frac{\alpha}{M^\gamma}, 1\right). \tag{5.6}$$

where $\text{DG}(a, b)$ denotes the double Gamma distribution with shape parameter a, rate parameter b and pdf $\{2\Gamma(a)\}^{-1} b^a |t|^{a-1} e^{-b|t|}$ $(t \in \mathbb{R})$, where $\Gamma(\cdot)$ is the Gamma function. More generally, we call a distribution as the double Dirichlet distribution with parameter $(a_1, \ldots, a_M)$, denoted by $\text{DD}(a_1, \ldots, a_M)$, if it can be represented by (5.6) with $T_j \sim \text{DG}(a_j, 1)$. Then, the DDG prior for λ has an alternative form as

$$\text{(DDG2)} \qquad \lambda = A\eta,\ A \sim \text{Ga}(a_0, b_0),\ \eta \sim \text{DD}\left(\frac{\alpha}{M^\gamma}, \ldots, \frac{\alpha}{M^\gamma}\right).$$

We will use the form (DDG2) for studying the theoretical properties of the DDG prior and focus on the form (DDG1) for posterior computation.

5.3 Theoretical properties

In this section, we study the prediction efficiency of the proposed Bayesian aggregation procedures for CA and LA in terms of convergence rate of posterior prediction.

We say that a Bayesian model $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$, with a prior distribution Π over the parameter space Θ, has a posterior convergence rate at least ϵ_n if

$$\Pi\big(d(\theta, \theta^*) \geqslant D\epsilon_n \big| X_1, \ldots, X_n\big) \xrightarrow{P_{\theta_0}} 0, \tag{5.7}$$

with a limit $\theta^* \in \Theta$, where d is a metric on Θ and D is a sufficiently large positive constant. For example, to characterize prediction accuracy, we use $d(\lambda, \lambda') = ||f_\lambda - f_{\lambda'}||_Q$ and $||n^{-1/2}F(\lambda - \lambda')||_2$ for CA and LA, respectively. Let $P_0 = P_{\theta_0}$ be the truth under which the iid observations $X_1, \ldots, X_n$ are generated. If $\theta_0 \in \Theta$, then the model is well-specified and under mild conditions, $\theta^* = \theta_0$. If $\theta_0 \notin \Theta$, then the limit θ^* is usually the point in Θ so that P_θ has the minimal Kullback-Leibler (KL) divergence to P_{θ_0}. (5.7) suggests that the posterior probability measure puts almost all its mass over a sequence of d-balls whose radii shrink towards θ^* at a rate ϵ_n. In the following, we make the assumption that σ is known, which is a standard assumption adopted in Bayesian asymptotic proofs to avoid long and tedious arguments. de Jonge and van Zanten (2013) studies the asymptotic behavior of the error standard deviation in regression when a prior is specified for σ. Their proofs can also be used to justify our setup when σ is unknown. In the rest of the chapter, we will frequently use C to denote a constant, whose meaning might change from line to line.

5.3.1 Posterior convergence rate of Bayesian convex aggregation

Let $\Sigma = (E_Q[f_i(X)f_j(X)])_{M\times M}$ be the second order moment matrix of $(f_1(X), \ldots, f_M(X))$, where $X \sim Q$. Let $f^* = \sum_{j=1}^M \lambda_j^* f_j$ be the best $L_2(Q)$-approximation of f_0 in the space $\mathcal{F}^\Lambda = \big\{f = \sum_{j=1}^M \lambda_j f_j : \lambda_j \geqslant 0, \sum_{j=1}^M \lambda_j = 1\big\}$ of all convex combinations of $f_1, \ldots, f_M$, i.e. $\lambda^* = \arg\min_{\lambda\in\Lambda} ||f_\lambda - f_0||_Q^2$. This misspecified framework also

includes the well-specified situation as a special case where $f_0 = f^* \in \mathcal{F}^\Lambda$. Denote the jth column of Σ by Σ_j.

We make the following assumptions:

(A1) There exists a constant $0 < \kappa < \infty$ such that $\sup_{1 \leqslant j \leqslant M} |\Sigma_{jj}| \leqslant \kappa$.

(A2) (Sparsity) There exists an integer $s > 0$, such that $||\lambda^*||_0 = s < n$.

(A3) There exists a constant $0 < \kappa < \infty$ such that $\sup_{1 \leqslant j \leqslant M} \sup_{x \in \mathcal{X}} |f_j(x)| \leqslant \kappa$.

- If $E_Q[f_j(X)] = 0$ for each j, then Σ is the variance covariance matrix. (A1) assumes the second moment Σ_{jj} of $f_j(X)$ to be uniformly bounded. By applying Cauchy's inequality, the off-diagonal elements of Σ can also be uniformly bounded by the same κ.

- (A3) implies (A1). This uniformly bounded condition is only used in Lemma 30 part a. As illustrated by Birgé (2004), such a condition is necessary for studying the $L_2(Q)$ loss of Gaussian regression with random design, since under this condition the Hellinger distance between two Gaussian regression models is equivalent to the $L_2(Q)$ distance between their mean functions.

- Since $\lambda^* \in \Lambda$, the l_1 norm of λ^* is always equal to one, which means that λ^* is always l_1-summable. $(A2)$ imposes an additional sparse structure on λ^*. We will study separately the convergence rates with and without (A2). It turns out that the additional sparse structure improves the rate if and only if $s \ll \sqrt{\frac{n}{\log M}}$.

The following theorem suggests that the posterior of f_λ concentrates on an $||\cdot||_Q$-ball around the best approximation f^* with a radius proportional to the minimax rate of CA. In the special case when $f_* = f_0$, the theorem suggests that the proposed Bayesian procedure is minimax optimal.

Theorem 24. *Assume (A3). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n iid copies of (X, Y) sampled from $X \sim Q$, $Y|X \sim N(f_0(X), \sigma^2)$. If $f^* = \sum_{j=1}^M \lambda_j^* f_j$ is the minimizer of $f \mapsto ||f - f_0||_Q$ on $\mathcal{F}^\Lambda$, then under the prior (DA), for some $D > 0$, as $n \to \infty$,*

$$E_{0,Q}\Pi\left(||f - f^*||_Q \geqslant D \min\left\{ \sqrt{\frac{M}{n}}, \sqrt[4]{\frac{\log(M/\sqrt{n}+1)}{n}} \right\} \middle| X_1, Y_1, \ldots, X_n, Y_n \right) \to 0.$$

Moreover, if (A2) is also satisfied, then as $n \to \infty$,

$$E_{0,Q}\Pi\left(||f - f^*||_Q \geqslant D \sqrt{\frac{s \log(M/s)}{n}} \middle| X_1, Y_1, \ldots, X_n, Y_n \right) \to 0.$$

5.3.2 Posterior convergence rate of Bayesian linear aggregation

Let $\lambda^* = (\lambda_1^*, \ldots, \lambda_M^*)$ be the coefficient such that $F\lambda^*$ best approximates F_0 in $||\cdot||_2$ norm, i.e. $\lambda^* = \arg\min_{\lambda \in \mathbb{R}^M} ||F\lambda - F_0||_2^2$. Similar to the CA case, such a misspecified framework also includes the well-specified situation as a special case where $F_0 = F\lambda^* \in \mathcal{F}^{\mathbb{R}^M}$. It is possible that there exists more than one such a minimizer and then we can choose λ^* with minimal nonzero components. This non-uniqueness will not affect our theorem quantifying the prediction performance of LA since any minimizers of $||F\lambda - F_0||_2^2$ will give the same prediction $F\lambda$. Our choice of λ^*, which minimizes $||\lambda^*||_0$, can lead to the fastest posterior convergence rate.

We make the following assumptions:

(B1) There exists a constant $0 < \kappa < \infty$ such that $\frac{1}{\sqrt{n}} \sup_{1 \leqslant j \leqslant M} ||F_j||_2 \leqslant \kappa$.

(B2a) (Sparsity) There exists an integer $s > 0$, such that $||\lambda^*||_0 = s < n$.

(B2b) (l_1-summability) There exists a constant $A_0 > 0$, such that $A^* = ||\lambda^*||_1 < A_0$.

(B3) For $m_0 = \lceil \sqrt{n} \rceil$, there exists a constant $\kappa_0 > 0$ such that $\frac{1}{\sqrt{n}} ||F\lambda||_2 \geqslant \kappa_0 ||\lambda||_1$ for all $\lambda \in \mathbb{R}^M$ with $||\lambda||_0 = m_0$.

- (B1) is the column normalizing condition for the design matrix. This assumption is mild since the predictors can always be normalized to satisfy it. This condition can also be considered as the empirical version of (A1), where the matrix Σ is replaced by its empirical estimator $\frac{1}{n}F^TF$.

- (B2a) is a counterpart of the sparsity condition (A2) of the aggregation problem. This assumption is commonly made in the high dimensional linear regression literature. (B2b) is assumed by Bühlmann (2006) in studying consistency of boosting for high dimensional linear regression. This condition includes the sparsity condition (B2a) as a special case while also including the case in which many components of λ^* are nonzero but small in magnitude. Similar to the aggregation problem, under (B2b), the sparsity gains only when $s \ll \sqrt{\frac{n}{\log M}}$. (B2a) also implies a sparsity constraint on $\eta^* = \lambda^*/A^*$, where η^* always satisfies $||\eta^*||_1 = 1$.

- (B3) is the same in spirit as the sparse eigenvalue condition made in Raskutti et al. (2011), which provides identifiability for m_0-sparse vectors. This assumption is only made for the l_1-summable case, where any l_1-summable $\lambda \in \mathbb{R}^M$ can be approximated by an m_0-sparse vector with error at most $O(||\lambda_1||\epsilon_n)$ under d_F (Lemma 28 part b), with ϵ_n given in (DA-PC), where $d_F(\lambda, \lambda') = ||n^{-1/2}F(\lambda - \lambda')||_2$. Under this assumption, we show that the posterior probability of $\{||\lambda||_1 \leqslant KA^*\}$ converges to zero as $n \to \infty$ for some constant K and therefore with high posterior probability, λ can be approximated by an m_0-sparse vector with error at most $O(\epsilon_n)$.

The following theorem is a counterpart of Theorem 24 for LA.

Theorem 25. *Assume (B1). Let Y be an n-dimensional response vector sampled from $Y \sim N(F_0, \sigma^2 I_n)$. Let λ^* be any one of the minimizers of $\lambda \mapsto ||F\lambda - F_0||_2$ in*

$\mathbb{R}^M$. *If (B2b) and (B3) are true, then under the prior (DDG2), for some* $D > 0$, *as* $n \to \infty$,

$$E_0\Pi\left(||n^{-\frac{1}{2}}F(\lambda - \lambda^*)||_2 \geqslant D \min\left\{ \sqrt{\frac{M}{n}}, \sqrt[4]{\frac{\log(M/\sqrt{n}+1)}{n}} \right\} \middle| \ Y \right) \to 0.$$

If (B2a) is true, then as $n \to \infty$,

$$E_0\Pi\left(||n^{-\frac{1}{2}}F(\lambda - \lambda^*)||_2 \geqslant D\sqrt{\frac{s\log(M/s)}{n}} \ \middle| \ Y \right) \to 0.$$

Theorem 25 suggests that in order to obtain the fastest posterior convergence rate for prediction, we can choose the λ^* having the minimal $||\lambda^*||_0$ among all minimizers of $||F\lambda - F_0||_2$. This suggests that the posterior measure tends to concentrate on the sparsest λ^* that achieves the same prediction accuracy, which explains the sparse adaptivity. The non-uniqueness happens when $M > n$.

5.4 Experiments

As suggested by Yang (2001), the estimator $\hat{f}_n$ depends on the order of the observations and one can randomly permute the order a number of times and average the corresponding estimators. In addition, one can add a third step of estimating $f_1, \ldots, f_M$ with the full dataset as $\hat{f}_1, \ldots, \hat{f}_M$ and setting the final estimator as $\tilde{f} = \sum_{j=1}^M \hat{\lambda}_j \hat{f}_j$. We will adopt this strategy and our splitting and aggregation scheme can be summarized as follows. First, we randomly divide the entire n samples into two subsets S_1 and S_2 with $|S_1| = n_1$ and $|S_2| = n_2$. As a default, we set $n_1 = 0.75n$ and $n_2 = 0.25n$. Using S_1 as a training set, we obtain M base learners $\hat{f}_1^{(n_1)}, \ldots, \hat{f}_M^{(n_1)}$. Second, we apply the above MCMC algorithms to aggregate these learners into $\hat{f}^{(n_1)} = \sum_{j=1}^M \hat{\lambda}_j \hat{f}_j^{(n_1)}$ based on the n_2 aggregating samples. Finally, we use the whole dataset to train these base learners, which gives us $\hat{f}_j^{(n)}$, and the final estimator is $\hat{f}^{(n)} = \sum_{j=1}^M \hat{\lambda}_j \hat{f}_j^{(n)}$. Therefore, one basic requirement on the base

learners is that they should be stable in the sense that $\hat{f}_j^{(n)}$ can not be dramatically different from $\hat{f}_j^{(n_1)}$ (e.g. CART might not be a suitable choice for the base learner).

5.4.1 Bayesian linear aggregation

In this subsection, we apply the Bayesian LA methods to the linear regression $Y = X\lambda + \epsilon$, with $X \in \mathbb{R}^M$ and $\epsilon \sim N(0, \sigma^2 I_n)$. Since every linear aggregation problem can be reformed as a linear regression problem, this is a simple canonical setting for testing our approach. We consider two scenarios: 1. the sparse case where the number of nonzero components in the regression coefficient λ is smaller than M and the sample size n; 2. the non-sparse case where λ can have many nonzero components, but the l_1 norm $||\lambda||_1 = \sum_{j=1}^{M} |\lambda_j|$ remains constant as M changes. We vary model dimensionality by letting $M = 5$, 20, 100 and 500.

We compare the Bayesian LA methods with lasso, ridge regression and horseshoe. Lasso (Tibshirani, 1996) is widely used for linear models, especially when λ is believed to be sparse. In addition, due to the use of l_1 penalty, the lasso is also minimax optimal when λ is l_1-summable (Raskutti et al., 2011). Ridge regression (Hoerl and Kennard, 1970) is a well-known shrinkage estimator for non-sparse settings. Horseshoe (Carvalho et al., 2010) is a Bayesian continuous shrinkage prior for sparse regression from the family of global-local mixtures of Gaussians (Polson and Scott, 2010). Horseshoe is well-known for its robustness and excellent empirical performance for sparse regression, but there is a lack of theoretical justification. n training samples are used to fit the models and $N - n$ testing samples are used to calculate the prediction root mean squared error (RMSE) $\left\{(N-n)^{-1}\sum_{i=n+1}^{N}(\hat{y}_i - y_i)^2\right\}^{1/2}$, where $\hat{y}_i$ denotes the prediction of y_i.

The MCMC algorithm for the Bayesian LA method is run for 2,000 iterations, with the first 1,000 iterations as the burn-in period. We set $\alpha = 1$, $\gamma = 2$, $a_0 = 0.01$ and $b_0 = 0.01$ for the hyperparameters. The tuning parameters in the MH steps are

chosen so that the acceptance rates are around 40%. The lasso is implemented by the `glmnet` package in `R`, the ridge is implemented by the `lm.ridge` function in `R` and horseshoe is implemented by the `monomvn` package in `R`. The iterations for horseshoe is set as the default 1,000. The regularization parameters in Lasso and ridge are selected via cross-validation.

Sparse case

In the sparse case, we choose the number of non-zero coefficients to be 5. The simulation data are generated from the following model:

$$(S) \qquad y = -0.5x_1 + x_2 + 0.4x_3 - x_4 + 0.6x_5 + \epsilon, \quad \epsilon \sim N(0, 0.5^2),$$

with M covariates $x_1, \ldots, x_M \sim$ i.i.d $N(0,1)$. The training size is set to be $n = 100$ and testing size $N - n = 1000$. As a result, (S) with $M = 5$ and 20 can be considered as moderate dimensional, while $M = 100$ and $M = 500$ are relatively high dimensional.

Table 5.1: RMSE for the sparse linear model (S). The numbers in the parentheses indicate the standard deviations. All results are based on 100 replicates.

M	5	20	100	500
LA	.511 (0.016)	.513 (0.016)	.529 (0.020)	.576 (0.023)
Lasso	.514 (0.017)	.536 (0.020)	.574 (0.039)	.613 (0.042)
Ridge	.514 (0.017)	.565 (0.019)	1.23 (0.139)	2.23 (0.146)
Horseshoe	.512 (0.016)	.519 (0.014)	.525 (0.019)	.590 (0.022)

From Table 5.1, all the methods are comparable when there is no nuisance predictor ($M = 5$). However, as more nuisance predictors are included, the Bayesian LA method and horseshoe have noticeably better performance than the other two methods. For example, for $M = 100$, the Bayesian LA method has 8% and 53%

improvements over lasso and ridge, respectively. In addition, as expected, ridge deteriorates more dramatically than the other two as M grows. It appears that Bayesian LA is more computationally efficient than horseshoe. For example, under $m = 100$ it takes horseshoe 50 seconds to draw 1,000 iterations but only takes LA about 1 second to draw 2,000 iterations.

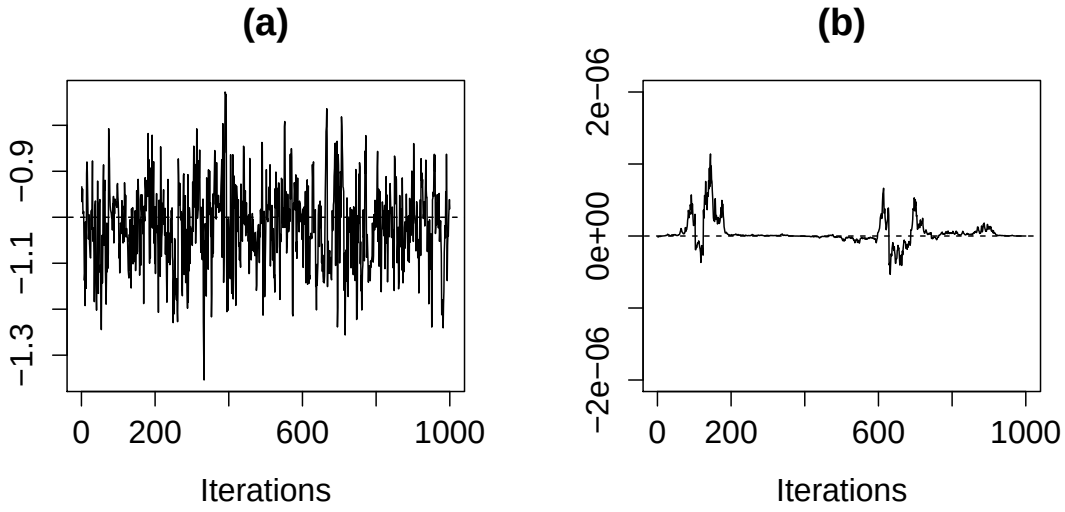


FIGURE 5.2: Traceplots for a non-zero regression coefficient and a zero coefficient.

Fig. 5.2 displays the traceplots after the burn-in for a typical non-zero and a typical zero regression coefficient respectively under $M = 100$. The non-zero coefficient mixes pretty well according to its traceplot. Although the traceplot of the zero coefficient exhibits some small fluctuations, their magnitudes are still negligible compared to the non-zero ones. We observe that these fluctuant traceplots like Fig. 5.2(b) only happens for those λ_j's whose posterior magnitudes are extremely small. The typical orders of the posterior means of those λ_j's in LA that correspond to unimportant predictors range from 10^{-17} to 10^{-2}. However, the posterior medians of unimportant predictors are less than 10^{-4} (see Fig. 5.3). This suggests that although the coefficients are not exactly to zero, the estimated regression coefficients with zero true values are still negligible compared to the estimators of the nonzero coefficients.

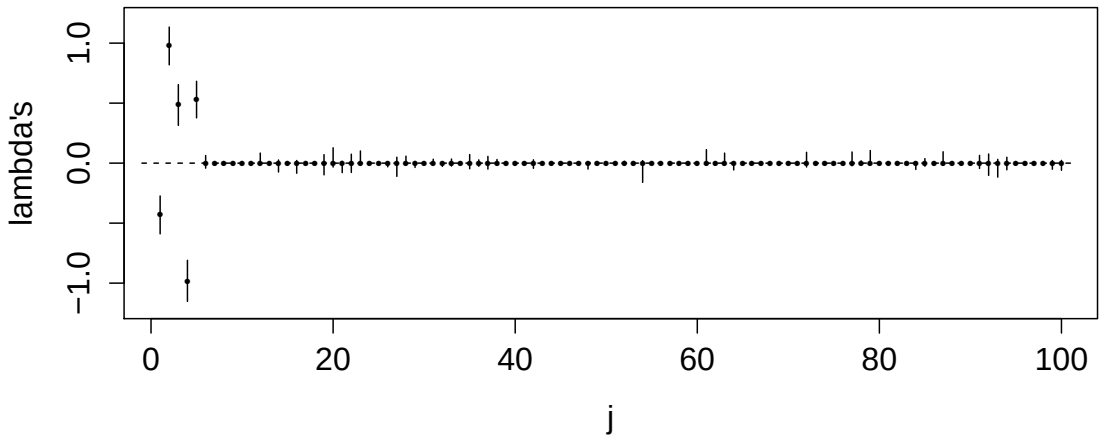


FIGURE 5.3: 95% posterior credible intervals for $\lambda_1, \ldots, \lambda_{100}$ in sparse regresion. The solid dots are the corresponding posterior medians.

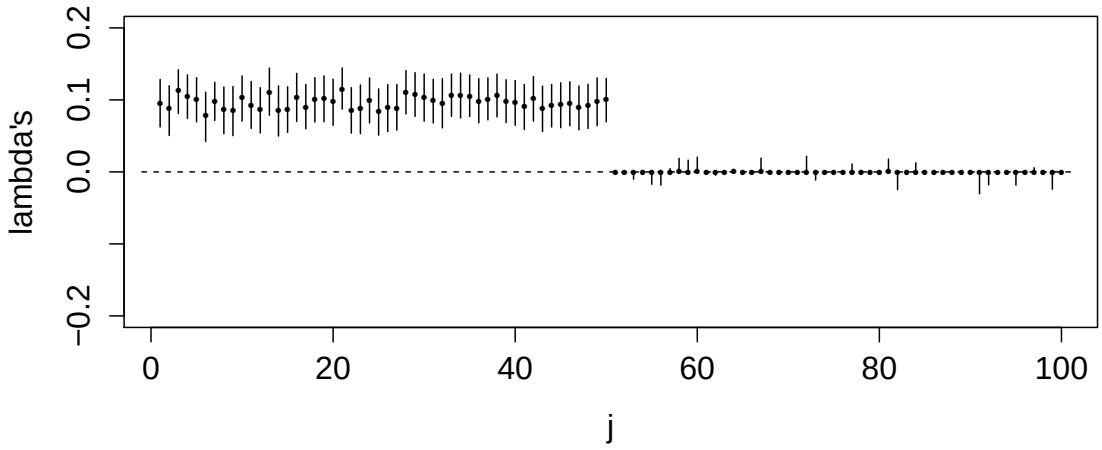


FIGURE 5.4: 95% posterior credible intervals for $\lambda_1, \ldots, \lambda_{100}$ in non-sparse regression. The solid dots are the corresponding posterior medians.

In addition, for LA the posterior median appears to be a better and more robust estimator for sparse regression than the posterior mean.

Non-sparse case

In the non-sparse case, we use the following two models as the truth:

$$(NS1) \qquad y = \sum_{j=1}^{M} \frac{3(-1)^j}{j^2} x_j + \epsilon, \quad \epsilon \sim N(0, 0.1^2),$$

$$(NS2) \qquad y = \sum_{j=1}^{\lfloor M/2 \rfloor} \frac{5}{\lfloor M/2 \rfloor} x_j + \epsilon, \quad \epsilon \sim N(0, 0.1^2),$$

with M covariates $x_1, \ldots, x_M \sim$ i.i.d $N(0,1)$. In (NS1), all the predictors affect the response and the impact of predictor x_j decreases quadratically in j. Moreover, λ satisfies the l_1-summability since $\lim_{p\to\infty} ||\lambda||_1 = \pi^2/3 \approx 4.9$. In (NS2), half of the predictors have the same influence on the response with $||\lambda||_1 = 5$. The training size is set to be $n = 200$ and testing size $N - n = 1000$ in the following simulations.

From Table 5.2, all the methods have comparable performance when M is moderate (i.e 5 or 20) in both non-sparse settings. In the non-sparse settings, horseshoe also exhibits excellent prediction performance. In most cases, LA, lasso and horseshoe have similar performance. As M increases to an order comparable to the sample size, LA and horseshoe tend to be more robust than lasso and ridge. As M becomes much greater than n, LA, lasso and horseshoe remain good in (NS1) while breaking down in (NS2); ridge breaks down in (NS1) while becoming the best in (NS2). It might be because in (NS1), although all λ_j's are nonzero, the first several predictors still dominate the impact on y. In contrast, in (NS2), half of λ_j's are nonzero and equally small. Fig. 5.4 plots 95% posterior credible intervals for $\lambda_1, \ldots, \lambda_{100}$ of (NS2) under $M = 100$. According to Section 5.1.1, the spasrse/non-sparse boundary for $(NS2)$ under $M = 100$ is $\sqrt{200}/\log 100 \approx 3 \ll 50$. Therefore, the results displayed

Table 5.2: RMSE for the non-sparse linear models (NS1) and (NS2). All results are based on 100 replicates.

	M	5	20	100	500
NS1	LA	.101 (0.002)	.112 (0.003)	.116 (0.005)	.129 (0.007)
	Lasso	.105 (0.006)	.110 (0.005)	.116 (0.005)	.155 (0.006)
	Ridge	.102 (0.003)	.107 (0.004)	.146 (0.008)	2.42 (0.053)
	Horseshoe	.102 (0.003)	.111 (0.003)	.114 (0.004)	.136 (0.005)
NS2	LA	.101 (0.002)	.104 (0.003)	.121 (0.005)	.326 (0.008)
	Lasso	.111 (0.006)	.106 (0.003)	.131 (0.007)	.323 (0.008)
	Ridge	.103 (0.003)	.107 (0.003)	.140 (0.008)	.274 (0.010)
	Horseshoe	.102 (0.003)	.104 (0.003)	.124 (0.004)	.308 (0.007)

in Fig. 5.4 can be classified into the non-sparse regime. A simple variable selection based on these credible intervals correctly identifies all 50 nonzero components.

Robustness against the hyperparameters

Since changing the hyperparameter α in the Dirichlet prior is equivalent to changing the hyperparameter γ, we perform a sensitivity analysis for γ in the above two regression settings with $M = 100$.

From Figure 5.5, the Bayesian LA method tends to be robust against the change in γ at a wide range. As expected, the Bayesian LA method starts to deteriorate as γ becomes too small. In particular, when γ is zero, the Dirichlet prior no longer favors sparse weights and the RMSE becomes large (especially for the sparse model) in all three settings. However, the Bayesian LA methods tend to be robust against increase in γ. As a result, we would recommend choosing $\gamma = 2$ in practice.

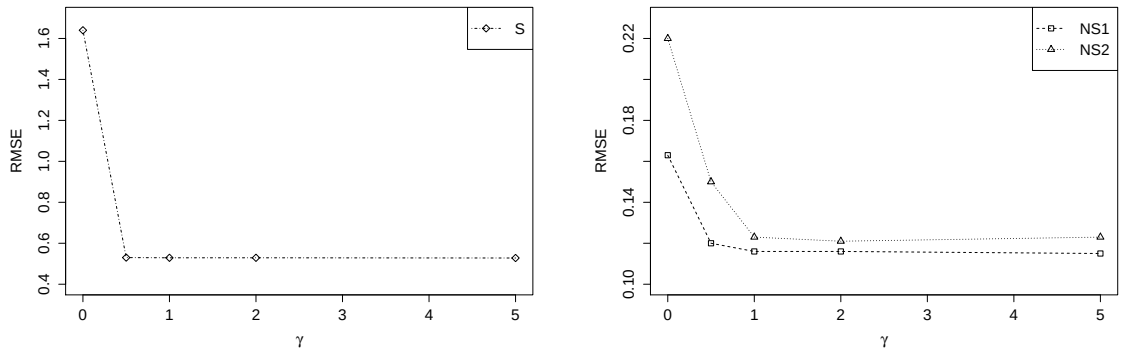


FIGURE 5.5: Robustness of the Bayesian LA methods against the hyperparameter γ. The results are based on 100 replicates.

5.4.2 Bayesian convex aggregation

In this subsection, we conduct experiments for the Bayesian convex aggregation method.

Simulations

The following regression model is used as the truth in our simulations:

$$y = x_1 + x_2 + 3x_3^2 - 2e^{-x_4} + \epsilon, \quad \epsilon \sim N(0, 0.5), \tag{5.8}$$

with p covariates $x_1, \ldots, x_d \sim$ i.i.d $N(0, 1)$. The training size is set to be $n = 500$ and testing size $N - n = 1000$ in the following simulations.

In the first simulation, we choose $M = 6$ base learners: CART, random forest (RF), lasso, SVM, ridge regression (Ridge) and neural network (NN). The Bayesian aggregation (BA) is compared with the super learner (SL). SL is implemented by the `SuperLearner` package in `R`. The implementations of the base learners are described in Table 5.3. The MCMC algorithm for the Bayesian CA method is run for 2,000 iterations, with the first 1,000 iterations as the burn-in period. We set $\alpha = 1$, $\gamma = 2$ for the hyperparameters. The simulation results are summarized in Table 5.4, where square roots of mean squared errors (RMSE) of prediction based on 100 replicates

are reported.

Table 5.3: Descriptions of the base learners.

Base learner	CART	RF	Lasso
`R` package	`rpart`	`randomForest`	`glmnet`
SVM	Ridge	NN	GAM
`e1071`	`MASS`	`nnet`	`gam`

Table 5.4: RMSE for the first simulation. All results are based on 100 replicates.

d	CART	RF	Lasso	SVM	Ridge	NN	SL	BA
5	3.31	3.33	5.12	2.71	5.12	3.89	2.66	**2.60**
	(0.41)	(0.42)	(0.33)	(0.49)	(0.33)	(0.90)	(0.48)	(0.48)
20	3.32	3.11	5.18	4.10	5.23	5.10	3.13	**3.00**
	(0.41)	(0.49)	(0.37)	(0.46)	(0.38)	(1.57)	(0.54)	(0.48)
100	3.33	3.17	5.17	5.48	5.64	7.12	3.19	**3.03**
	(0.38)	(0.45)	(0.32)	(0.35)	(0.33)	(1.31)	(0.45)	(0.45)

In the second simulation, we consider the case when M is moderately large. We consider $M = 26$, 56 and 106 by introducing $(M - 6)$ new base learners in the following way. In each simulation, for $j = 1, \ldots, M - 6$, we first randomly select a subset S_j of the covariates $\{x_1, \ldots, x_d\}$ with size $p = \lfloor \min\{n^{1/2}, d/3\} \rfloor$. Then the jth base learner f_j is fitted by the general additive model (GAM) with the response y and covariates in S_j as predictors. This choice of new learners is motivated by the fact that the truth is sparse when M is large and brutally throwing all covariates into the GAM tends to have a poor performance. Therefore, we expect that a base learner based on GAM that uses a small subset of the covariates containing the important predictors x_1, x_2 x_3 and x_4 tends to have better performance than the full model. In addition, with a large M and moderate p, the probability that one of the randomly selected $(M - 6)$ models contains the truth is high. In this simulation, we compare BA with SL and a voting method using the average prediction across all base learners. For illustration, the best prediction performance among the $(M - 6)$ random-subset base learners is also reported. Table 5.5 summarizes the results.

Table 5.5: RMSE for the second simulation study. All results are based on 100 replicates.

d	M	Best	Voting	SL	BA
20	26	3.14	4.63	3.40	**2.78**
		(0.82)	(0.48)	(0.60)	(0.52)
	56	2.86	4.98	3.33	**2.79**
		(1.57)	(0.71)	(0.86)	(0.78)
	106	2.79	4.95	3.23	**2.73**
		(0.62)	(0.60)	(0.70)	(0.61)
100	26	3.14	4.72	3.09	**2.78**
		(0.71)	(0.44)	(0.50)	(0.45)
	56	2.95	4.93	3.07	**2.78**
		(0.46)	(0.45)	(0.50)	(0.47)
	106	2.84	4.90	2.98	**2.69**
		(0.45)	(0.47)	(0.55)	(0.51)
500	26	5.21	5.75	3.77	**3.19**
		(0.75)	(0.50)	(0.650)	(0.59)
	56	4.86	5.92	4.02	**3.18**
		(0.78)	(0.59)	(0.73)	(0.70)
	106	4.65	5.98	4.18	**3.13**
		(0.69)	(0.45)	(0.52)	(0.49)

Applications

We apply BA to four datasets from the UCI repository. Table 5.6 provides a brief description of those datasets. We use CART, random forest, lasso, support vector machine, ridge regression and neural networks as the base learners. We run 40,000 iterations for the MCMC for the BA for each dataset and discard the first half as the burn-in. Table 5.7 displays the results. As we can see, for 3 datasets (auto-mpg, concrete and forest), the aggregated models perform the best. In particular, for the auto-mpg dataset, BA has 3% improvement over the best base learner. Even for the slump dataset, aggregations still have comparable performance to the best base learner. The two aggregation methods SL and BA have similar performance for all the datasets.

Table 5.6: Descriptions of the four datasets from the UCI repository. CCS: concrete compressive strength.

dataset	sample size	# of predictors	response variable
auto-mpg	392	8	mpg
concrete	1030	8	CCS*
slump	103	7	slump
forest	517	12	log(1+area)

Table 5.7: RMSE of aggregations for real data applications. All results are based on 10-fold cross-validations.

dataset	Cart	RF	Lasso	SVM	Ridge	NN	GAM	SL	BA
auto-mpg	3.42	2.69	3.38	2.68	3.40	7.79	2.71	**2.61**	**2.61**
concrete	9.40	5.35	10.51	6.65	10.50	16.64	7.95	**5.31**	5.33
slump	7.60	**6.69**	7.71	7.05	8.67	7.11	6.99	7.17	7.03
forest	.670	.628	.612	.612	.620	.613	.622	.606	**.604**

5.5 Proofs of the main results

Let $K(P,Q) = \int \log(dP/dQ)dP$ be the KL divergence between two probability distributions P and Q, and $V(P,Q) = \int |\log(dP/dQ) - K(P,Q)|^2 dP$ be a discrepancy measure.

5.5.1 Concentration properties of Dirichlet distribution and double Dirichlet distribution

According to the posterior asymptotic theory developed in Ghosal et al. (2000) (for iid observations, e.g. regression with random design, such as the aggregation problem in section 5.2.2), to ensure a posterior convergence rate of at least ϵ_n, the prior has to put enough mass around θ^* in the sense that

$$\text{(PC1)} \qquad \Pi(B(\theta^*, \epsilon_n)) \geqslant e^{-n\epsilon_n^2 C}, \text{ with}$$

$$B(\theta^*, \epsilon) = \{\theta \in \Theta : K(P_{\theta^*}, P_\theta) \leqslant \epsilon^2,\ V(P_{\theta^*}, P_\theta) \leqslant \epsilon^2\},$$

for some $C > 0$. For independent but non-identically distributed (noniid) observations (e.g. regression with fixed design, such as the linear regression problem in section 5.2.3), where the likelihood takes a product form $P_\theta^{(n)}(Y_1, \ldots, Y_n) = \prod_{i=1}^n P_{\theta,i}(Y_i)$, the corresponding prior concentration condition becomes (Ghosal and van der Vaart, 2007)

(PC2) $$\Pi(B_n(\theta^*, \epsilon_n)) \geqslant e^{-n\epsilon_n^2 C}, \text{ with}$$

$$B_n(\theta^*, \epsilon) = \left\{ \theta \in \Theta : \frac{1}{n}\sum_{i=1}^n K(P_{\theta^*,i}, P_{\theta,i}) \leqslant \epsilon^2, \ \frac{1}{n}\sum_{i=1}^n V(P_{\theta^*,i}, P_{\theta,i}) \leqslant \epsilon^2 \right\}.$$

If a (semi-)metric d_n, which might depend on n, dominates KL and V on Θ, then (PC) is implied by $\Pi(d_n(\theta, \theta^*) \leqslant c\epsilon_n) \geqslant e^{-n\epsilon_n^2 C}$ for some $c > 0$. In the aggregation problem with a random design and parameter $\theta = \lambda$, we have $K(P_{\theta*}, P_\theta) = V(P_{\theta*}, P_\theta) = \frac{1}{2\sigma^2}||\sum_{j=1}^M (\lambda_j - \lambda_j^*) f_j||_Q^2 = \frac{1}{2\sigma^2}(\lambda - \lambda^*)^T \Sigma (\lambda - \lambda^*)$. Therefore, we can choose $d_n(\theta, \theta^*)$ as $d_\Sigma(\lambda, \lambda^*) = ||\Sigma^{1/2}(\lambda - \lambda^*)||_2$. In the linear aggregation problem with a fixed design and $\theta = \lambda$, $\sum_{j=1}^n K(P_{\theta*,i}, P_{\theta,i}) = \sum_{j=1}^n V(P_{\theta*,i}, P_{\theta,i}) = \frac{1}{2\sigma^2}||F(\lambda - \lambda^*)||_2$, where $P_{\theta,i}(Y) = P_\lambda(Y|F^{(i)})$. Therefore, we can choose $d_n(\theta, \theta^*)$ as $d_F(\lambda, \lambda^*) = ||\frac{1}{\sqrt{n}} F(\lambda - \lambda^*)||_2$.

For CA and LA, the concentration probabilities can be characterized by those of $\lambda^* \in \Lambda$ and $\eta^* \in D_{M-1} = \{\eta \in \mathbb{R}^M, ||\eta||_1 = 1\}$. Therefore, it is important to investigate the concentration properties of the Dirichlet distribution and the double Dirichlet distribution as priors over Λ and D_{M-1}. The concentration probabilities $\Pi(d_\Sigma(\lambda, \lambda^*) \leqslant c\epsilon)$ and $\Pi(d_F(\eta, \eta^*) \leqslant c\epsilon)$ depend on the location of the centers λ^* and η^*, which are characterized by their geometrical properties, such as sparsity and l_1-summability. The next lemma characterizes these concentration probabilities and is of independent interest.

Lemma 26. *Assume (A1) and (B1).*

a. Assume (A2). Under the prior (DA), for any $\gamma \geqslant 1$,

$$\Pi(d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon) \geqslant \exp\left\{ - C\gamma s \log \frac{M}{\epsilon} \right\}, \quad \textit{for some } C > 0.$$

b. Under the prior (DA), for any $m > 0$, *any* $\lambda \in \Lambda$ *and any* $\gamma \geqslant 1$,

$$\Pi\left(d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon + \frac{C}{\sqrt{m}} \right) \geqslant \exp\left\{ - C\gamma m \log \frac{M}{\epsilon} \right\},$$

$$\Pi(d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon) \geqslant \exp\left\{ - C\gamma M \log \frac{M}{\epsilon} \right\}, \quad \textit{for some } C > 0.$$

c. Assume (B2a). Under the prior for η *in (DDG2), for any* $\gamma \geqslant 1$,

$$\Pi(d_F(\eta, \eta^*) \leqslant \epsilon) \geqslant \exp\left\{ - C\gamma s \log \frac{M}{\epsilon} \right\}, \quad \textit{for some } C > 0.$$

d. Under the prior for η *in (DDG2), for any* $m > 0$, *any* $\eta \in D_{M-1}$ *and any* $\gamma \geqslant 1$,

$$\Pi\left(d_F(\eta, \eta^*) \leqslant \epsilon + \frac{C}{\sqrt{m}} \right) \geqslant \exp\left\{ - C\gamma m \log \frac{M}{\epsilon} \right\},$$

$$\Pi(d_F(\eta, \eta^*) \leqslant \epsilon) \geqslant \exp\left\{ - C\gamma M \log \frac{M}{\epsilon} \right\}, \quad \textit{for some } C > 0.$$

- The lower bound $\exp\{-C\gamma s \log(M/\epsilon)\}$ in Lemma 26 can be decomposed into two parts: $\exp\{-C\gamma s \log M\}$ and $\exp\{Cs \log \epsilon\}$. The first part has the same order as $1/\binom{M}{s}$, one over the total number of ways to choose s indices from $\{1, \ldots, M\}$. The second part is of the same order as ϵ^s, the volume of an ϵ-cube in $\mathbb{R}^s$. Since usually which s components are nonzero and where the vector composed of these s nonzero components locates in $\mathbb{R}^s$ are unknown, this prior lower bound cannot be improved.

- The priors (DA) and (DDG2) do not depend on the sparsity level s. As a result, Lemma 26 suggests that the prior concentration properties hold simultaneously

for all λ^* or η^* with different s and thus these priors can adapt to an unknown sparsity level.

By the first two parts of Lemma 26, the following is satisfied for the prior (DA) with D large enough,

$$\text{(DA-PC)}\quad \Pi(d_\Sigma(\lambda,\lambda^*)\leqslant\epsilon_n)\geqslant e^{-n\epsilon_n^2 C},\ \text{with } \epsilon_n=\begin{cases} D\sqrt{\frac{s\log(M/s)}{n}}, & \text{if } ||\lambda^*||_0=s;\\ D\sqrt{\frac{M}{n}}, & \text{if } M\leqslant\sqrt{n};\\ D\sqrt[4]{\frac{\log(M/\sqrt{n}+1)}{n}}, & \text{if } M>\sqrt{n}.\end{cases}$$

This prior concentration property will play a key role in characterizing the posterior convergence rate of the prior (DA) for Bayesian aggregation.

Based on the prior concentration property of the double Dirichlet distribution provided in Lemma 26 part c and part d, we have the corresponding property for the prior (DDG2) by taking into account the prior distribution of $A=||\lambda||_1$.

Corollary 27. *Assume (B1).*

a. Assume (B2a). Under the prior (DDG2), for any $\gamma\geqslant 1$,

$$\Pi(d_F(\lambda,\lambda^*)\leqslant\epsilon)\geqslant\exp\left\{-C\gamma s\log\frac{M}{\epsilon}\right\},\quad \textit{for some } C>0.$$

b. Assume (B2b). Under the prior (DDG2), for any $m>0$, any $\eta\in D_{M-1}$ and any $\gamma\geqslant 1$,

$$\Pi\left(d_F(\lambda,\lambda^*)\leqslant\epsilon+\frac{C}{\sqrt{m}}\right)\geqslant\exp\left\{-C\gamma m\log\frac{M}{\epsilon}\right\},$$

$$\Pi(d_F(\lambda,\lambda^*)\leqslant\epsilon)\geqslant\exp\left\{-C\gamma M\log\frac{M}{\epsilon}\right\},\quad \textit{for some } C>0.$$

Based on the above corollary, we have a similar prior concentration property for the prior (DDG2):

(DDG2-PC) $\Pi(d_F(\theta,\theta^*)\leqslant c\epsilon_n)\geqslant e^{-n\epsilon_n^2 C}$, with the same ϵ_n in (DA-PC).

5.5.2 *Supports of the Dirichlet distribution and the double Dirichlet distribution*

By Ghosal et al. (2000), a second condition to ensure the posterior convergence rate of $\theta^* \in \Theta$ at least ϵ_n is that the prior Π should put almost all its mass in a sequence of subsets of Θ that are not too complex. More precisely, one needs to show that there exists a sieve sequence $\{\mathcal{F}_n\}$ such that $\theta^* \in \mathcal{F}_n \subset \Theta$, $\Pi(\mathcal{F}_n^c) \leqslant e^{-n\epsilon_n^2 C}$ and $\log N(\epsilon_n, \mathcal{F}_n, d_n) \leqslant n\epsilon_n^2$ for each n, where for a metric space $\mathcal{F}$ associated with a (semi-)metric d, $N(\epsilon, \mathcal{F}, d)$ denotes the minimal number of d-balls with radii ϵ that are needed to cover $\mathcal{F}$.

For the priors (DA) and (DDG2), the probability of the space of all s-sparse vectors is zero. We consider the approximate s-sparse vector space $\mathcal{F}_{s,\epsilon}^{\Lambda}$ defined in Section 5.2.1 for CA. For LA, we define $\mathcal{F}_{B,s,\epsilon}^{D} = \{\theta = A\eta : \eta \in D_{M-1}, \sum_{j=s+1}^{M} |\eta_{(j)}| \leqslant B^{-1}\epsilon;\ 0 \leqslant A \leqslant B\}$, where $|\eta_{(1)}| \geqslant \cdots \geqslant |\eta_{(M)}|$ is the ordered sequence of $\eta_1, \ldots, \eta_M$ according to their absolute values.

The following lemma characterizes the complexity of Λ, $D_{M-1}^{B} = \{A\eta : \eta \in D_{M-1};\ 0 \leqslant A \leqslant B\}$, $\mathcal{F}_{s,\epsilon}^{\Lambda}$ and $\mathcal{F}_{B,s,\epsilon}^{D}$ in terms of their covering numbers.

Lemma 28. *Assume (A1) and (B1).*

a. For any $\epsilon \in (0,1)$, integer $s > 0$ and $B > 0$, we have

$$\log N(\epsilon, \mathcal{F}_{s,\epsilon}^{\Lambda}, d_{\Sigma}) \lesssim s \log \frac{M}{\epsilon},$$

$$\log N(\epsilon, \mathcal{F}_{B,s,\epsilon}^{D}, d_F) \lesssim s \log \frac{M}{\epsilon} + s \log B.$$

b. For any $\epsilon \in (0,1)$ and integer $m > 0$, we have

$$\log N(C/\sqrt{m}, \Lambda, d_\Sigma) \lesssim m \log M,$$

$$\log N(\epsilon, \Lambda, d_\Sigma) \lesssim M \log \frac{M}{\epsilon},$$

$$\log N(CB/\sqrt{m}, BD_{M-1}, d_F) \lesssim m \log M,$$

$$\log N(B\epsilon, BD_{M-1}, d_F) \lesssim M \log \frac{M}{\epsilon}.$$

The next lemma provides upper bounds to the complementary prior probabilities of $\mathcal{F}^{\Lambda}_{s,\epsilon}$ and $\mathcal{F}^{D}_{B,s,\epsilon}$. The proof utilizes the connection between the Dirichlet distribution and the stick-breaking representation of the Dirichlet processes (Sethuraman, 1994).

Lemma 29. *a. For any $\epsilon \in (0,1)$, under the prior (DA) with $\gamma > 1$, we have*

$$\Pi(\lambda \notin \mathcal{F}^{\Lambda}_{s,\epsilon}) \lesssim \exp\left\{ -Cs(\gamma - 1) \log \frac{M}{\epsilon} \right\}.$$

b. For any $\epsilon \in (0,1)$, under the prior (DDG2) with $\gamma > 1$, we have

$$\Pi(\theta \notin \mathcal{F}^{D}_{B,s,\epsilon}) \lesssim \exp\left\{ -Cs(\gamma - 1) \log \frac{M}{\epsilon} - Cs \log B \right\} + \exp\{-CB\}.$$

5.5.3 Test construction

For CA, we use the notation $P_{\lambda,Q}$ to denote the joint distribution of (X, Y), whenever $X \sim Q$ and $Y|X \sim N(\sum_{j=1}^{M} \lambda_j f_j(X), \sigma^2)$ for any $\lambda \in \Lambda$ and $E_{\lambda,Q}$ the expectation with respect to $P_{\lambda,Q}$. Use $P^{(n)}_{\lambda,Q}$ to denote the n-fold convolution of $P_{\lambda,Q}$. Let $X^n = (X_1, \ldots, X_n)$ and $Y^n = (Y_1, \ldots, X_n)$ be n copies of X and Y. Recall that f_0 is the true regression function that generates the data. We use $P_{0,Q}$ to denote the corresponding true distribution of Y. For LA, we use P_λ to denote the distribution of Y, whenever $Y \sim N(F\lambda, \sigma^2 I_n)$ and E_λ the expectation with respect to P_λ.

For both aggregation problems, we use the "M-open" view where f_0 might not necessarily belong to $\mathcal{F}^{\Lambda}$ or $\mathcal{F}^{\mathbb{R}^M}$. We apply the result in Kleijn and van der Vaart

(2006) to construct a test under misspecification for CA with random design and explicitly construct a test under misspecification for LA with fixed design. Note that the results in Kleijn and van der Vaart (2006) only apply for random-designed models. For LA with fixed design, we construct a test based on concentration inequalities for Gaussian random variables.

Lemma 30. *Assume (A3).*

a. *Assume that* $f^* = \sum_{j=1}^M \lambda_j^* f_j$ *satisfies* $E_Q(f - f^*)(f^* - f_0) = 0$ *for every* $f \in \mathcal{F}^\Lambda$. *Then there exist* $C > 0$ *and a measurable function* ϕ_n *of* X^n *and* Y^n, *such that for any other vector* $\lambda_2 \in \Lambda$,

$$P_{0,Q}^{(n)} \phi_n(X^n, Y^n) \leqslant \exp\big\{ - C n d_\Sigma^2(\lambda_2, \lambda^*)\big\}$$

$$\sup_{\lambda \in \Lambda:\ d_\Sigma(\lambda, \lambda_2) < \frac{1}{4} d_F(\lambda^*, \lambda_2)} P_{\lambda,Q}^{(n)}(1 - \phi_n(X^n, Y^n)) \leqslant \exp\big\{ - C n d_\Sigma^2(\lambda_2, \lambda^*)\big\}.$$

b. *Assume that* $\lambda^* \in \mathbb{R}^d$ *satisfies* $F^T(F\lambda^* - F_0) = 0$ *for every* $\lambda \in \mathbb{R}^d$. *Then there exists a measurable function* ϕ_n *of* Y *and some* $C > 0$, *such that for any other* $\lambda_2 \in \mathbb{R}^d$,

$$P_0 \phi_n(Y) \leqslant \exp\big\{ - C n d_F^2(\lambda_2, \lambda^*)\big\}$$

$$\sup_{\lambda \in \mathbb{R}^M:\ d_F(\lambda, \lambda_2) < \frac{1}{4} d_F(\lambda^*, \lambda_2)} P_\lambda(1 - \phi_n(Y)) \leqslant \exp\big\{ - C n d_X^2(\lambda_2, \lambda^*)\big\}.$$

- As we discussed in the remark in section 5.3.1, in order to apply Kleijn and van der Vaart (2006) for Gaussian regression with random design, we need the mean function to be uniformly bounded. For the convex aggregation space $\mathcal{F}^\Lambda$, this uniformly bounded condition is implied by (A3). For the linear regression with fixed design, we do not need the uniformly bounded condition. This property ensures that the type I and type II errors in b do not deteriorate as $||\lambda_2||_1$ grows, which plays a critical role in showing that the posterior probability

of $\{A > CA^*\}$ converges to zero in probability for C sufficiently large, where $A = ||\lambda||_1$ and $A^* = ||\lambda^*||_1$. Similarly, if we consider CA with a fixed design, then only an assumption like (B1) on the design points is needed.

- The assumption on f^* in CA is equivalent to that f^* is the minimizer over $f \in \mathcal{F}^\Lambda$ of $||f - f_0||_Q^2$, which is proportional to the expectation of the KL divergence between two normal distributions with mean functions $f_0(X)$ and $f(X)$ with $X \sim Q$. Therefore, f^* is the best $L_2(Q)$-approximation of f_0 within the aggregation space $\mathcal{F}^\Lambda$ and the lemma suggests that the likelihood function under f^* tends to be exponentially larger than other functions in $\mathcal{F}^\Lambda$. Similarly, the condition on λ^* in LA is equivalent to that λ^* is the minimizer over $\lambda \in \mathbb{R}^d$ of $||F\lambda - F_0||_2^2$, which is proportional to the KL divergence between two multivariate normal distributions with mean vectors $F\lambda$ and F_0.

5.5.4 Proof of Theorem 24

The proof follows similar steps as the proof of Theorem 2.1 in Ghosal et al. (2000). The difference is that we consider the misspecified framework where the asymptotic limit of the posterior distribution of f is f^* instead of the true underlying regression function f_0. As a result, we need to apply the test condition in Lemma 30 part a in the model misspecified framework. We provide a sketched proof as follows.

Let ϵ_n be given by (DA-PC) and $\Pi^B(\lambda) = \Pi\big(\lambda | B(\lambda^*, \epsilon_n)\big)$ with $B(\lambda^*, \epsilon_n)$ defined in (PC1). By Jensen's inequality applied to the logarithm,

$$\log \int_{B(\lambda^*, \epsilon_n)} \prod_{i=1}^n \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i, Y_i) d\Pi^B(\lambda) \geqslant \sum_{i=1}^n \int_{B(\lambda^*, \epsilon_n)} \log \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i, Y_i) d\Pi^B(\lambda).$$

By the definition of $B(\lambda^*, \epsilon_n)$ and an application of Chebyshev's inequality, we have

that for any $C > 0$,

$$P_0\Big\{ \sum_{i=1}^n \int_{B(\lambda^*,\epsilon_n)} \Big(\log \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) + K(P_{\lambda_0,Q},P_{\lambda,Q}) \Big) d\Pi^B(\lambda)$$

$$\leqslant -(1+C)n\epsilon_n^2 + n\int_{B(\lambda^*,\epsilon_n)} K(P_{\lambda_0,Q},P_{\lambda,Q}) d\Pi^B(\lambda) \Big\}$$

$$\leqslant P_0\Big\{ \sum_{i=1}^n \int_{B(\lambda^*,\epsilon_n)} \Big(\log \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) + K(P_{\lambda_0,Q},P_{\lambda,Q}) \Big) d\Pi^B(\lambda) \leqslant -Cn\epsilon_n^2 \Big\}$$

$$\leqslant \frac{n\int_{B(\lambda^*,\epsilon_n)} V(P_{\lambda_0,Q},P_{\lambda,Q}) d\Pi^B(\lambda)}{(Cn\epsilon_n^2)^2} \leqslant \frac{1}{C^2 n\epsilon_n^2} \to 0, \text{ as } n\to\infty.$$

Combining the above two yields that on some set A_n with P_0-probability converging to one,

$$\int_{B(\lambda^*,\epsilon_n)} \prod_{i=1}^n \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) d\Pi(\lambda) \geqslant \exp(-(1+C)n\epsilon_n^2)\Pi(B(\lambda^*,\epsilon_n)) \geqslant \exp(-C_0 n\epsilon_n^2), \tag{5.9}$$

for some $C_0 > 0$, where we have used the fact that $\Pi(B(\lambda^*,\epsilon_n)) \geqslant \Pi(d_\Sigma(\lambda,\lambda^*) \leqslant C\epsilon_n) \geqslant \exp(-Cn\epsilon_n^2)$ for some $C > 0$.

Let $\mathcal{F}_n = \mathcal{F}^\lambda_{as,\epsilon_n}$ for some $a > 0$ if (A2) holds and otherwise $\mathcal{F}_n = \Lambda$. Then by Lemma 28 part a and Lemma 29 part a, for some constants $C_1 > 0$ and $C_2 > 0$,

$$\log N(\epsilon_n, \mathcal{F}_n, d_\Sigma) \leqslant C_1 n\epsilon_n^2, \quad \Pi(\lambda \notin \mathcal{F}_n) \leqslant \exp(-C_2 n\epsilon_n^2). \tag{5.10}$$

Because C_2 is increasing with the a in the definition of $\mathcal{F}_n$, we can assume $C_2 > C_0 + 1$ by properly selecting an a.

For some $D_0 > 0$ sufficiently large, let $\lambda_1^*, \ldots, \lambda_J^* \in \mathcal{F}_n - \{\lambda : d_\Sigma(\lambda,\lambda^*) \leqslant 4D_0\epsilon_n\}$ with $|J| \leqslant \exp(C_1 n\epsilon_n^2)$ be J points that form an $D_0\epsilon_n$-covering net of $\mathcal{F}_n - \{\lambda : d_\Sigma(\lambda,\lambda^*) \leqslant 4D_0\epsilon_n\}$. Let $\phi_{j,n}$ be the corresponding test function provided by Lemma 30 part a with $\lambda_2 = \lambda_j^*$ for $j = 1,\ldots,J$. Set $\phi_n = \max_j \phi_{j,n}$. Since $d_\Sigma(\lambda_j^*,\lambda^*) \geqslant 4D_0\epsilon_n$ for any j, we obtain

$$P_{0,Q}^{(n)}\phi_n \leqslant \sum_{j=1}^J P_{0,Q}^{(n)}\phi_{j,n} \leqslant |J| \exp(-C16D_0^2 n\epsilon_n^2) \leqslant \exp(-C_3 n\epsilon_n^2), \tag{5.11}$$

where $C_3 = 16CD_0^2 - 1 > 0$ for D_0 large enough. For any $\lambda \in \mathcal{F}_n - \{\lambda : d_\Sigma(\lambda, \lambda^*) \leqslant 4D_0\epsilon_n\}$, by the design, there exists a j_0 such that $d_\Sigma(\lambda^*_{j_0}, \lambda) \leqslant D_0\epsilon_n$. This implies that $d_\Sigma(\lambda^*_{j_0}, \lambda^*) \geqslant 4D_0\epsilon_n \geqslant 4d_\Sigma(\lambda^*_{j_0}, \lambda)$, therefore

$$\begin{aligned}
&\sup_{\lambda\in\mathcal{F}_n:\ d_\Sigma(\lambda,\lambda^*)\geqslant 4D_0\epsilon_n} P^{(n)}_{\lambda,Q}\phi_n \\
&\leqslant \min_j \sup_{\lambda\in\Lambda:\ d_\Sigma(\lambda,\lambda^*_j)<\frac{1}{4}d_\Sigma(\lambda^*,\lambda^*_j)} P^{(n)}_{\lambda,Q}(1-\phi_n) \leqslant \exp\big\{-C_4 n\epsilon_n^2\big\}, \qquad (5.12)
\end{aligned}$$

with $C_4 = 16CD_0^2 > C_0 + 1$ with D_0 sufficiently large. With $D = 4D_0$, we have

$$\begin{aligned}
&E_{0,Q}\Pi(d_\Sigma(\lambda,\lambda^*) \geqslant D\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n) I(A_n) \\
&\leqslant P^{(n)}_{0,Q}\phi_n + E_{0,Q}\Pi(d_\Sigma(\lambda,\lambda^*) \geqslant D\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n) I(A_n)(1-\phi_n). \qquad (5.13)
\end{aligned}$$

By (5.9), (5.10) and (5.12), we have

$$\begin{aligned}
&E_{0,Q}\Pi(d_\Sigma(\lambda,\lambda^*) \geqslant D\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n) I(A_n)(1-\phi_n) \\
&\leqslant P^{(n)}_{0,Q}(1-\phi_n) I(A_n) \frac{\int_{\lambda\in\mathcal{F}_n: d_\Sigma(\lambda,\lambda^*)\geqslant D\epsilon_n} \prod_{i=1}^n \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) d\Pi(\lambda)}{\int_{B(\lambda^*,\epsilon_n)} \prod_{i=1}^n \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) d\Pi(\lambda)} \\
&\qquad + P^{(n)}_{0,Q} I(A_n) \frac{\int_{\lambda\notin\mathcal{F}_n} \prod_{i=1}^n \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) d\Pi(\lambda)}{\int_{B(\lambda^*,\epsilon_n)} \prod_{i=1}^n \frac{P_{\lambda,Q}}{P_{0,Q}}(X_i,Y_i) d\Pi(\lambda)} \\
&\leqslant \exp(C_0 n\epsilon_n^2) \sup_{\lambda\in\mathcal{F}_n:\ d_\Sigma(\lambda,\lambda^*)\geqslant 4D_0\epsilon_n} P^{(n)}_{\lambda,Q}\phi_n + \exp(C_0 n\epsilon_n^2)\Pi(\lambda\notin\mathcal{F}_n) \leqslant 2\exp(-n\epsilon_n^2). \qquad (5.14)
\end{aligned}$$

Combining the above with (5.11), (5.13), and the fact that $E_{0,Q}I(A_n^c) \to 0$ as $n \to \infty$, Theorem 24 can be proved.

5.5.5 Proof of Theorem 25

For the sparse case where (B2a) is satisfied, we construct the sieve by $\mathcal{F}_n = \mathcal{F}^D_{bn\epsilon_n^2, as, \epsilon_n}$ with the ϵ_n given in (DDG2-PC), where $a > 0$, $b > 0$ are sufficiently large constants. Then by Lemma 28 part a and Lemma 29 part b, we have

$$\log N(\epsilon_n, \mathcal{F}_n, d_F) \leqslant C_1 n\epsilon_n^2, \quad \Pi(\lambda\notin\mathcal{F}_n) \leqslant \Pi(-C_2 n\epsilon_n^2), \qquad (5.15)$$

where C_2 is increasing with a and b. The rest of the proof is similar to the proof of Theorem 24 with the help of (5.15), Corollary 26 part b and Lemma 30 part b.

Next, we consider the dense case where (B2b) and (B3) are satisfied. By the second half of Lemma 28 part b, the approximation accuracy of BD_{M-1} degrades linearly in B. Therefore, in order to construct a sieve such that (5.15) is satisfied with the ϵ_n given in (DDG2-PC), we need to show that $E_0\Pi(A \leqslant KA^*|Y) \to 0$ as $n \to \infty$ with some constant $K > 0$. Then by conditioning on the event $\{A \leqslant KA^*\}$, we can choose $\mathcal{F}_n = BD_{M-1}$ with $B = KA^*$, which does not increase with n, and 5.15 will be satisfied. As long as 5.15 is true, the rest of the proof will be similar to the sparse case.

We only prove that $E_0\Pi(A \leqslant KA^*|Y) \to 0$ as $n \to \infty$ here. By (B1) and (B3), for any $\eta \in D_{M-1}$ and $A > 0$, $d_F(A\eta, A^*\eta^*) \geqslant \kappa_0 A - \kappa A^*$. As a result, we can choose K large enough so that $d_F(A\eta, A^*\eta^*) \geqslant 4$ for all $A \geqslant KA^*$ and all $\eta \in D_{M-1}$. Therefore, by Lemma 30 part b, for any $\lambda_2 = A_2\eta_2$ with $A_2 > KA^*$ and $\eta_2 \in D_{M-1}$, there exists a test ϕ_n such that

$$P_{\lambda^*}\phi_n(Y) \leqslant \exp\big\{-Cn\big\}$$

$$\sup_{\lambda \in \mathbb{R}^M:\ d_F(\lambda,\lambda_2) < \frac{1}{4} d_F(\lambda^*,\lambda_2)} P_\lambda(1 - \phi_n(Y)) \leqslant \exp\big\{-Cn\big\}.$$

By choosing K large enough, we can assume that $\kappa_0 KA^*/8 > \kappa + \kappa A^*/4$. For any $\lambda = A\eta$ satisfying $d_F(\eta, \eta_2) \leqslant \kappa_0/8$ and $|A - A_2| \leqslant 1$, by (B1) and $A_2 > KA^*$ we have

$$d_F(\lambda, \lambda_2) \leqslant d_F(A\eta, A_2\eta) + d_F(A_2\eta, A_2\eta_2) \leqslant \kappa + \frac{1}{8}\kappa_0 A_2$$

$$\leqslant \frac{1}{4}(\kappa_0 A_2 - \kappa A^*) \leqslant \frac{1}{4} d_F(\lambda^*, \lambda_2).$$

Combining the above, we have that for any $\lambda_2 = A_2\eta_2$ with $A_2 > KA^*$ and $\eta_2 \in$

D_{M-1},

$$P_{\lambda^*}\phi_n(Y) \leqslant \exp\big\{-Cn\big\}$$
$$\sup_{|A-A_2|\leqslant 1, d_F(\eta,\eta_2)\leqslant \kappa_0/8} P_\lambda(1-\phi_n(Y)) \leqslant \exp\big\{-Cn\big\}.$$

Let $A_1^*, \ldots, A_{J_1}^*$ be a 1-covering net of the interval $[KA^*, Cn\epsilon_n^2]$ with $J_1 \leqslant Cn\epsilon_n^2$ and $\eta_1^*, \ldots, \eta_{J_2}^*$ be a $\kappa_0/8$-covering net of D_{M-1} with $\log J_2 \leqslant Cn\epsilon_n^2$ (by Lemma 28 part b with $B = 1$). Let ϕ_j $(j = 1, \ldots, J_1 J_2)$ be the corresponding tests associated with each combination of (A_s^*, η_t^*) for $s = 1, \ldots, J_1$ and $t = 1, \ldots, J_2$. Let $\phi_n = \max_j \phi_j$. Then for n large enough,

$$P_{\lambda^*}\phi_n(Y) \leqslant \exp\big\{\log(n\epsilon_n^2) + Cn\epsilon_n^2 - Cn\big\} \leqslant \exp\big\{-Cn\big\}$$
$$\sup_{\lambda = A\eta : A\in[KA^*, Cn\epsilon_n^2], \eta\in D_{M-1}} P_\lambda(1-\phi_n(Y)) \leqslant \exp\big\{-Cn\big\}. \tag{5.16}$$

Moreover, because $A \sim \mathrm{Ga}(a_0, b_0)$, we have

$$\Pi(\lambda \notin Cn\epsilon_n^2 D_{M-1}) \leqslant \Pi(A > Cn\epsilon_n^2) \leqslant \exp\{-Cn\epsilon_n^2\}. \tag{5.17}$$

Combining (5.16) and (5.17), we can prove that $E_0\Pi(A \leqslant KA^*|Y) \to 0$ as $n \to \infty$ by the same arguments as in (5.14).

6

Sequential Markov chain Monte Carlo

6.1 Introduction

The Bayesian paradigm provides a natural formalism for optimal learning from data in a sequential manner, with the posterior distribution at one time point becoming the prior distribution at the next. Consider the following general setup. Let $\{\pi_t : t \in \mathbb{N}\}$ be a sequence of probability distributions indexed by discrete time $t \in \mathbb{N} = \{0, 1, \ldots\}$. Assume that each π_t can either be defined on a common measurable space $(E, \mathcal{E})$ or a sequence of measurable spaces $\{(E_t, \mathcal{E}_t) : t \in \mathbb{N}\}$ with non-decreasing dimensions $d_0 \leqslant d_1 \leqslant \ldots$. Without loss of generality, we assume that $(E_t, \mathcal{E}_t) = (\mathbb{R}^{d_t}, \mathcal{B}(\mathbb{R}^{d_t}))$, where $\mathcal{B}(\mathbb{R}^{d_t})$ is the Borel field on $\mathbb{R}^{d_t}$. Moreover, π_t admits a density $\pi_t(\theta^{(t)})$ with respect to the Lebesgue measure $\lambda^{d_t}(d\theta^{(t)})$, where $\theta^{(t)} = (\theta^{(t-1)}, \eta_t)$ is the quantity or parameter of interest at t and $\eta_t \in \mathbb{R}^{d_t - d_{t-1}}$ is the additional component other than $\theta^{(t)}$. This framework can be considered as a generalization of Liu and Chen (1998) from dynamic systems to arbitrary models or extension of Del Moral et al. (2006) from fixed space E to time-dependent space E_t.

Many applications can be placed within this setting. In the sequential Bayesian

inference context, $\theta^{(t)}$ corresponds to a vector composed of all the parameters and other unknowns to sample at time t. Similarly, π_t is the posterior distribution of $\theta^{(t)}$ given the data collected until time t. For example, in generalized linear models with fixed number of covariates, $\theta^{(t)}$ includes the regression coefficients and residual variance and d_t is a constant. In finite mixture models, $\theta^{(t)}$ includes both the parameters of the mixture components and mixing distribution, and the latent class indicators for each observation, so that d_t is increasing with t. In state-space models, $\theta^{(t)}$ could be a vector composed of static parameters and state space variables, where the size of the latter grows with t. Even in batch situations where a full dataset $\{y_1, \ldots, y_n\}$ has been obtained, we can still consider the sequence of posterior distributions $p(\theta^{(t)}|y_1, \ldots, y_t)$ for $t \leqslant n$. The annealing effect (Chopin, 2002) of adding data sequentially can lead to substantial improvements over usual MCMC methods, which incorporate all the data at once and sample serially.

Markov Chain Monte Carlo (MCMC) is an important statistical analysis tool, which is designed to sample from complex distributions. It can not only be used for Bayesian analysis where a normalizing constant is unknown, but also for frequentist analysis when the likelihood involves high dimensional integrals such as in missing data problems and mixed effects models. However, in general, MCMC methods have several major drawbacks. First, it is difficult to assess whether a Markov chain has reached its stationary distribution. Second, a Markov chain can be easily trapped in local modes, which in turn would impede convergence diagnostics. To speed up explorations of the state space, annealing approaches introduce companion chains with flattened stationary distributions to facilitate the moves among separated high energy regions (Geyer, 1991; Earlab and Deema, 2005; Kou et al., 2006).

An alternative to MCMC is sequential Monte Carlo (SMC). The main idea of SMC is to represent the distribution π_t through the empirical distribution $\hat{\pi}_t =$

$\sum_{i=1}^{N} W_t^{(i)} \delta_{X_t^{(i)}}$, where $\{(W_t^{(i)}, X_t^{(i)}) : i = 1, \ldots, N\}$ is a finite set of N weighted particles with $\sum_{i=1}^{N} W_t^{(i)} = 1$ and δ_x is the Dirac measure at x. As a new observation y_{t+1} arrives, both weights and states of particles are updated in order to represent the new posterior π_{t+1}. Although SMC can potentially solve many of the drawbacks of MCMC mentioned above, it suffers from the notorious weight degeneracy issue where few particles quickly dominate as t increases, causing performance based on $\hat{\pi}_t$ to degrade. Moreover, numerical errors introduced in an early stage can accumulate for some SMCs when static parameters are present (Storvik, 2002). Although many variants of SMC, such as adaptive importance sampling (West, 1993), resample-move strategies (Chopin, 2002) and annealed importance sampling (Neal, 2001), are proposed to alleviate the weight degeneracy problem, issues remain, particularly in models involving moderate to high-dimensional unknowns.

In this work, we propose a sequential MCMC algorithm to sample from $\{\pi_t : t \in \mathbb{N}\}$ that is based on parallel sequential approximation algorithms. The proposed sequential MCMC is a population-based MCMC, where each chain is constructed via specifying a transition kernel T_t for updating $\theta^{(t)}$ within time t and a jumping kernel J_t for generating additional component η_t. The annealing effect of sequential MCMC can substantially boost efficiency of MCMC algorithms with poor mixing rates with slight modifications. By exploiting multiple processors, SMCMC has comparable total computational burden as MCMC. For streaming data problems, SMCMC distributes this burden over time and allows one to extract current available information at any time point.

We develop a theoretical justification on the convergence of SMCMC and provide explicit bounds on the error in terms of a number of critical quantities. The theory indicates an opposite phenomenon as the weight degeneracy effect of SMC: the deviations or numerical errors in the early stage decay exponentially fast as t

grows, leading to estimators with increasing accuracy. One of our main theoretical contribution is to formulate the geometric ergodicity for general state-space Markov chains. Our conditions are much easier to verify compared to the usual conditions for geometric erdogic chain, such as the drift and minorization conditions (Rosenthal, 1995). In the special case of uniform ergodic chains, our conditions are weaker than the minorization condition (Meyn and Tweedie, 1993). We provide two different proofs for the uniform ergodicity. The first proof is based on the coupling techniques and the second is based on the operator theory. As an easy byproduct of this formulation, we show that for any geometrically ergodic transition kernel, starting from any initial distribution, the one step distribution always becomes closer to its stationary distribution.

This chapter has the following organization. In Section 5.2, we present a generic SMCMC algorithm to sample from a sequence of distributions $\{\pi_t : t \in \mathbb{N}\}$ and discuss possible variations. In Section 5.3, we study the convergence properties of SMCMC under various settings, including parametric and nonparametric models. Section 5.4 compares SMCMC with other methods in a finite mixture of normals simulation. In Section 5.5, we apply SMCMC to an on-line nonparametric regression problem. In section 5.6, we review and introduce some new results on the convergence of Markov chains. Technical proofs appear in Appendix D.

6.2 Sequential Markov chain Monte Carlo

We propose a sequential Markov chain Monte Carlo (SMCMC) class of algorithms in this section. The main idea of SMCMC is to run time-inhomogeneous Markov chains in parallel with the transition kernels depending on the current available data. Inferences can be made by using the ensemble composed of the last samples in those chains.

6.2.1 Notation and assumptions

Let Y_t denote the data coming in at time t, $Y^{(t)} = (Y_1, \ldots, Y_t)$ the entire data up to t, $\theta^{(t)}$ the parameters at time t, d_t the size of $\theta^{(t)}$ and $\pi^{(t)}(\theta^{(t)})$ the prior distribution, implying that we can add parameters over time. In the sequel, we will use the same notation to interchangeably denote a probability measure or its density function with respect to the Lebesgue measure λ. Throughout this chapter, we use the notation $||p|| = 2\sup_A |p(A)| = \int |p(x)|dx$ to denote the L_1-norm (total variation norm) for a signed measure p. Although not necessary, for notational simplicity we assume that the prior is compatible: $\pi^{(t)}(\theta^{(t)}) = \int \pi^{(t+1)}(\theta^{(t)}, \eta_{t+1})\lambda(d\eta_{t+1})$ with $\theta^{(t+1)} = (\theta^{(t)}, \eta_{t+1})$. Under this assumption, we can suppress the superscript t in $\pi^{(t)}$. The compatibility assumption is a consequence of the restriction that if the extra parameters in the prior at time $t+1$ are marginalized out, then we recover the prior at time t. This restriction is trivially satisfied under the special case when d_t does not grow with time, and is also true under more general priors such as hierarchical priors for mixed effects models and Gaussian process priors for nonparametric regression. We propose to conduct L Markov chains in parallel exploiting L processors to obtain samples, $\theta^{(t,l)} = \{\theta^{(1,t,l)}, \ldots, \theta^{(m_t,t,l)}\}$ for $t = 1, 2, \ldots$ and $l = 1, \ldots, L$, where m_t is the number of draws obtained at time t for each chain and $\theta^{(s,t,l)} \in \mathbb{R}^{d_t}$ is the sth draw obtained in the lth chain at t. The ensemble $\Theta_t = \{\theta^{m_t,t,l} : l = 1, \ldots, L\}$ will be treated as independent draws sampled from the posterior $\pi_t(\theta^{(t)}) = \pi(\theta^{(t)}|Y^{(t)})$ at time t.

6.2.2 Markov chain construction

At each time t, we consider two kernels: a jumping kernel J_t proposing the parameter jumping from $t-1$ to t at the beginning of time t and a transition kernel T_t specifying the parameter updating process within time t. $J_t(\cdot, \cdot)$ is defined on $\mathbb{R}^{d_{t-1}} \times \mathbb{R}^{d_t}$ and is primarily designed for the situation when the parameter grows at t. In the case when

$d_t = d_{t-1}$, J_t could be chosen as the identity map. $T_t(\cdot,\cdot)$ is defined on $\mathbb{R}^{d_t} \times \mathbb{R}^{d_t}$ so that the posterior π_t is the stationary measure of the Markov chain with transition kernel T_t, i.e. $\pi_t(\theta') = \int_{\mathbb{R}^{d_t}} \pi_t(\theta)\ T_t(\theta, \theta')\lambda(d\theta)$. T_t aims at transferring the distribution of the draws Θ_{t-1} from π_{t-1} to π_t. From standard Markov chain theory (Meyn and Tweedie, 1993), if the chain with transition kernel T_t is an aperiodic recurrent Harris chain, then $||T_t^{m_t} \circ p_0 - \pi_t|| \to 0$ as $m_t \to \infty$ for any initial distribution p_0. Therefore, as we repeat applying the transition T_t for enough times, the distribution of Θ_t will converge to π_t. Theorem 46 in section 6.3.2 quantifies such approximation error with given m_t. Section 6.2.5 provides recommendations on choosing m_t in practice.

We construct our SMCMC based on J_t and T_t as follows:

1. At $t = 0$, we set $m_t = 1$ and draw L samples from a known distribution, for example, the prior $\pi = \pi_0$. The samples at $t = 0$ are denoted as $\theta^{(1,0,1)}, \ldots, \theta^{(1,0,L)}$.

2. At $t > 0$, we first update $\theta^{(m_{t-1},t-1,l)}$ to $\theta^{(1,t,l)}$ through the jumping kernel J_t as

$$P\big(\theta^{(1,t,l)} \big| \theta^{(m_{t-1},t-1,l)}\big) = J_t\big(\theta^{(m_{t-1},t-1,l)}, \theta^{(1,t,l)}\big),$$

   in parallel for $l = 1, \ldots, L$. Then, for $s = 1, \ldots, m_t - 1$, $\theta^{(s,t,l)}$ is sequentially transited to $\theta^{(s+1,t,l)}$ through the transition kernel T_t as

$$P\big(\theta^{(s+1,t,l)} \big| \theta^{(s,t,l)}\big) = T_t\big(\theta^{(s,t,l)}, \theta^{(s+1,t,l)}\big),$$

   in parallel for $l = 1, \ldots, L$.

With the above updating scheme, the last samples $\{\theta^{(m_t,t,l)} : l = 1, \ldots, L\}$ at t would be taken as the ensemble Θ_t to approximate the posterior π_t. Let $\hat{\pi}_t$ denote the common distribution of $\theta^{(m_t,t,l)}$'s. Theorem 32 in section 6.3.2 and Theorem 36 in section 6.3.3 guarantee the error $||\hat{\pi}_t - \pi_t||$ decays to zero as t increases to infinity as long as $||\pi_t - \pi_{t-1}|| \to 0$. When d_t is growing, the π_t in the L_1 norm is understood as the marginal distribution of $\theta^{(t-1)}$ given by $\pi_t(\theta^{(t-1)}) = \int_{\mathbb{R}^{d_t - d_{t-1}}} \pi_t(\theta^{(t-1)}, \eta_t)\lambda(d\eta_t)$.

The sequential Monte Carlo sampler (Del Moral et al., 2006) could also be cast into this framework if the jumping kernel J_t is a random kernel that depends on Θ_{t-1}. However, as Theorem 32 indicates, with sufficient iterations m_t at each time point t, one can guarantee the convergence without the resampling step used in SMC algorithms as long as the posterior π_t does not change too much in t.

As the mixture model example in section 6.4 demonstrates, even in batch problems, the annealing effect of adding data sequentially will lead to substantial improvements over usual MCMC algorithms that incorporate all the data at once and sample serially. This annealing effect has also been observed in the SMC literature, for example, Chopin (2002). For streaming data problems, SMCMC avoids the need to restart the algorithm at each time point as new data arrive, and allows real time updating exploiting multiple processors and distributing the computational burden over time. For example, the SMCMC for nonparametric probit regression in section 6.5 has similar total computational burden as running MCMC chains in parallel using multiple processors. However, SMCMC distributes this burden over time, and one can extract current available information at any time point. Moreover, the samples $\{\theta^{(m_t,t,l)} : l = 1, \ldots, L\}$ within each time point are drawn from independent chains. This independence and the annealing effect can substantially boost efficiency of MCMC algorithms with poor mixing rates.

6.2.3 Choice of J_t

We shall restrict the jumping kernel J_t to be a pre-specified transition kernel that leaves $\theta^{(t-1)}$ unchanged by letting

$$P\big((\tilde{\theta}^{(t-1)}, \eta_t)|\theta^{(t-1)}\big) = J_t\big(\theta^{(t-1)}, (\tilde{\theta}^{(t-1)}, \eta_t)\big) I(\theta^{(t-1)} = \tilde{\theta}^{(t-1)}), \tag{6.1}$$

where $I(\cdot)$ denotes the indicator function. Otherwise, J_t can always be decomposed into an updating of θ^{t-1} followed by a generation of η_t, where the former step can be

absorbed into T_{t-1}. Henceforth, with slight abuse of notation, the jumping kernel J_t will be considered as a map from $\mathbb{R}^{d_{t-1}}$ to $\mathbb{R}^{d_t - d_{t-1}}$, mapping $\theta^{(t-1)}$ to η_t.

Intuitively, if $\theta^{(t-1)}$ is approximately distributed as $\pi_t(\theta^{(t-1)})$ and η_t is sampled from the conditional posterior $\pi_t(\eta_t|\theta^{(t-1)})$, then $(\theta^{(t-1)}, \eta_t)$ is approximately distributed as $\pi_t(\theta^{(t-1)}, \eta_t) = \pi_t(\theta^{(t-1)})\pi_t(\eta_t|\theta^{(t-1)})$, the exact posterior distribution. This observation is formalized in Lemma 35 in section 6.3.3, suggesting that the jumping kernel J_t should be chosen close to full conditional $\pi_t(\eta_t|\theta^{(t-1)})$ at time t. Two types of J_t can be used (some examples can be found in Del Moral et al. (2006)):

1. *Exact conditional sampling.* When draws from the full conditional $\pi_t(\eta_t|\theta^{(t-1)})$ can be easily sampled, J_t can be chosen as this full conditional. For example, $\pi_t(\eta_t|\theta^{(t-1)})$ can be recognized as some standard distribution. Even when $\pi_t(\eta_t|\theta^{(t-1)})$ is unrecognizable, if $d_t - d_{t-1}$ is small, then we can apply the accept-reject algorithm (Robert and Casella, 2004) or slice sampler (Neal, 2003).

2. *Approximate conditional sampling.* When sampling from the full conditional of η_t is difficult, we can use other transition kernels, such as blocked Metropolis-Hastings (MH) or inter-woven MH or Gibbs steps chosen to have $\pi_t(\eta_t|\theta^{(t-1)})$ as the stationary distribution.

Theorem 36 in section 6.3.3 provides an explicit expression about the impact of

$$\tau_t = \sup_{\theta^{(t-1)} \in \mathbb{R}^{d_{t-1}}} ||\pi_t(\cdot|\theta^{(t-1)}) - J_t(\theta^{(t-1)}, \cdot)||$$

on the approximation error of π_t, which basically requires $\tau_t \to 0$ as $t \to \infty$. To achieve $\tau_t \to 0$, one can run the transition kernel in approximate conditional sampling case for an increasing number of iterations as t grows. However, we observe good practical performances for a fixed small number of iterations.

6.2.4 Choice of T_t

Lemma 43 in section 6.6 suggests that a good $T_t(\theta, \theta')$ should be close to $\pi_t(\theta')$. The transition kernel T_t can be chosen as in usual MCMC algorithms. For example, T_t can be the transition kernel associated with blocked or inter-weaved MH or Gibbs samplers. For conditionally conjugate models, it is particularly convenient to use Gibbs and keep track of conditional sufficient statistics to mitigate the increase in storage and computational burden over time.

6.2.5 Choice of m_t

The number of samples in each chain per time point, m_t, should be chosen to be small enough to meet the computational budget while being large enough so that the difference between the distribution of samples in Θ_t and the posterior distribution π_t goes to zero. Formal definitions of difference and other concepts will be given in the next section. Intuitively, for a given t, if the Markov chain with transition kernel T_t has slow mixing or there are big changes in π_t from π_{t-1}, then m_t should be large. Theorem 32 in section 6.3.2 provides explicit bounds on the approximation error as a function of m_t's. Moreover, for a given $\epsilon \in (0, 1)$, Theorem 32 implies that if we select m_t to be the minimal integer k such that $r_t(k) \leqslant 1 - \epsilon$, where r_t is the rate function associated with T_t defined in (**??**), then the distribution of Θ_t converges to π_t as $t \to \infty$ under the assumption that $||\pi_t - \pi_{t-1}|| \to 0$. Typical rate functions can be chosen as $r_t(k) = \rho^k$, for some ρ^k. Since the rate functions r_t relate to the unknown mixing rate of the Markov chain with transition kernel T_t, we estimate them in an online manner.

To estimate r_t we utilize the relationship between the mixing rate of a Markov chain and its autocorrelation function. By comparing (6.10) and (6.11) in section 6.3.5, the decay rate of the autocorrelation function provides an upper bound for the mixing rate. Therefore, we can bound the rate function $r_t(k)$ with the lag-k

autocorrelation function

$$f_t(k) = \max_{j=1,\ldots,p} \operatorname{corr}(X_j^{(k)}, X_j^{(0)}),$$

where $(X_j^{(1)}, \ldots, X_j^{(p)})$ is the p-dimensional sample in the kth step of the Markov chain with transition kernel T_t.

For a single Markov chain, the common choice of estimating $f_t(k)$ by the sample average of lag-k differences over the steps from $s = s_1, \ldots, s_2$ as

$$\tilde{f}_t(k) = \max_{j=1,\ldots,p} \frac{\sum_{s=s_1}^{s_2} (X_j^{(s)} - \bar{X}_j)(X_j^{(s-k)} - \bar{X}_j)}{\sum_{s=s_1}^{s_2} (X_j^{(s)} - \bar{X}_j)^2},$$

where $\bar{X}_j = \sum_{s=s_1}^{s_2} X_j^{(s)}/(s_2 - s_1 + 1)$, could have large bias even though $s_2 - s_1$ is large. The reason is that for slow mixing Markov chains, the samples tend to be stuck in local modes, leading to high variation of $\tilde{f}_t(k)$'s with $X_j^{(s)}$ starting from different regions. Within these local modes, $\tilde{f}_t(k)$ might decay fast, inappropriately suggesting good mixing. In our algorithm, we have L chains running independently in parallel. Hence, instead of averaging over time, we can estimate the autocorrelation function $f_t(k)$ by averaging across the independent chains as

$$\hat{f}_t(k) = \max_{j=1,\ldots,p} \frac{\sum_{l=1}^{L} (X_j^{(k,l)} - \bar{X}_j^{(k)})(X_j^{(0,l)} - \bar{X}_j^{(0)})}{\big(\sum_{l=1}^{L} (X_j^{(k,l)} - \bar{X}_j^{(k)})^2\big)^{1/2} \big(\sum_{l=1}^{L} (X_j^{(0,l)} - \bar{X}_j^{(0)})^2\big)^{1/2}},$$

where $X_j^{(k,l)}$ is the jth component of the sample in the kth step of the lth chain and $\bar{X}_j^{(k)} = \sum_{l=1}^{L} X_j^{(k,l)}/L$ is the ensemble average of the draws in the kth step across the L Markov chains. $\hat{f}_t$ will be more robust than $\tilde{f}_t$ to local modes. Although by Slutsky's theorem, both estimators are asymptotically unbiased as $s_2 - s_1 \to \infty$ and $L \to \infty$ respectively, the convergence of $\tilde{f}_t$ might be much slower than that of $\hat{f}_t$ due to potential high correlations among the summands in $\tilde{f}_t$.

In our case, the estimator $\hat{f}_t(k)$ takes the form of

$$\hat{f}_t(k) = \max_{j=1,\ldots,p} \frac{\sum_{l=1}^{L} (\theta_j^{(k+1,t,l)} - \bar{\theta}_j^{(k+1,t)})(\theta_j^{(1,t,l)} - \bar{\theta}_j^{(1,t)})}{\big(\sum_{l=1}^{L} (\theta_j^{(k+1,t,l)} - \bar{\theta}_j^{(k+1,t)})^2\big)^{1/2} \big(\sum_{l=1}^{L} (\theta_j^{(1,t,l)} - \bar{\theta}_j^{(1,t)})^2\big)^{1/2}}, \tag{6.2}$$

where $\bar{\theta}_j^{(k,t)} = \sum_{l=1}^{L} \theta_j^{(k,t,l)}/L$ is the jth component of the ensemble average of the draws across the L Markov chains in the kth step at time t. For each $t > 0$, we choose m_t to be the minimal integer k such that the sample autocorrelation decreases below $1-\epsilon$, i.e. $m_t = \min\{k : \hat{f}_t(k) \leqslant 1-\epsilon\}$. In practice, we can choose ϵ according to the full sample size n and error tolerance ϵ_T based on Theorem 32. For example, for small datasets with $n \sim 10^2$, we recommend $\epsilon = 0.5$ and for large datasets, ϵ such that $\sum_{t=1}^{n} \frac{\epsilon^{n+1-t}}{\sqrt{t}} \leqslant \epsilon_T$, where $t^{-1/2}$ is a typical rate for $||\pi_t - \pi_{t-1}||$ for regular parametric models (Lemma 33). To summarize, Algorithm 1 provides pseudo code for SMCMC.

Algorithm 1 Sequential Markov Chain Monte Carlo

```
m_0 ← 1
for l = 1 to L do
    Draw θ^(1,0,l) ~ π_0
end for
for t = 1 to n do
    m_t ← 1
    ρ ← 1
    for l = 1 to L do
        Draw [η^(t,l) | θ^(m_{t-1},t-1,l)] ~ J_t(θ^(m_{t-1},t-1,l), ·)
        θ^(1,t,l) ← (θ^(m_{t-1},t-1,l), η^(t,l))
    end for
    while ρ > 1 − ϵ do
        m_t ← m_t + 1
        for l = 1 to L do
            Draw [θ^(m_t,t,l) | θ^(m_t−1,t,l)] ~ T_t(θ^(m_t−1,t,l), ·)
        end for
        Calculate f̂_t(m_t − 1) by (6.2)
        ρ ← f̂_t(m_t − 1)
    end while
    Θ_t ← {θ^(m_t,t,l) : l = 1, ..., L}
end for
```

All the loops for l in the above algorithm can be computed in parallel. Assuming the availability of a distributed computing platform with multiple processors, Algorithm 1 has comparable computational complexity to running MCMC in parallel on L processors starting with the full data at time t. The only distributed operation is computation of $\hat{f}_t$, which can be updated every s_0 iterations to reduce communication time. Moreover, the t loop can be conducted whenever t_0 (> 1) new data

points accrue, rather than as each data point arrives, as long as $||\pi_t - \pi_{t-t_0}|| \to 0$ as $t \to \infty$. More generally, for any sequence $t_1 < t_2 < \ldots < t_{k_0} = n$ such that $||\pi_{t_k} - \pi_{t_{k-1}}|| \to 0$ as $k \to \infty$, the loop for t can be changed into "for $k = 1$ to k_0 do $t \leftarrow t_k$ ... end for". Since the posterior π_t is expected to vary slower as t grows, the batch sizes $t_k - t_{k-1}$ can be increasing in k, leading to faster computations. To avoid the SMCMC becoming too complicated, we shall restrict our attention to Algorithm 1 in the rest of the chapter.

6.3 Convergence of SMCMC

In this section, we study the convergence properties of SMCMC as $t \to \infty$ by applying the convergence results for Markov chains in Section 6.6.

We introduce some notation that will be used throughout this section. For a transition kernel $T(x, y)$, we recursively define its t-step transition kernel by $T^t(x, y) = \int T^{t-1}(x, z)T(z, y)\lambda(dz)$. Similarly, given an initial density p_0, we denote by $T^t \circ p_0$ the probability measure evolved after tth steps with transition kernel T from the initial distribution p_0, which is related to T^t by $T^t \circ p_0(x) = \int T^t(z, x)p_0(z)\lambda(dz)$.

SMCMC generates L time-inhomogeneous Markov chains. To investigate its asymptotic properties, we need a notion of convergence. Existing literature on the convergence of MCMC or adaptive MCMC focuses on the case when the stationary distribution does not change with time. A nonadaptive MCMC algorithm is said to be converging if

$$||Q^t \circ p_0 - \pi|| \to 0, \quad \text{as } t \to \infty, \tag{6.3}$$

where $||\cdot||$ is the L_1 norm, Q is the time homogeneous transition kernel, p_0 is the initial distribution and π is the unique stationary measure. However, for sequential MCMC, both the stationary distribution π_t and the transition kernel Q_t is changing over time. As an extension of (6.3), a stationary-distribution-varying Markov chain

is said to be convergent if

$$||Q_t \circ \cdots \circ Q_1 \circ p_0 - \pi_t|| \to 0, \quad \text{as } t \to \infty. \tag{6.4}$$

In our case, $Q_t = T_t^{m_t} \circ J_t$, where T_t, J_t and m_t are defined in section 6.2.2.

6.3.1 Implications of the convergence

In this subsection, we illustrate the annealing effect of the SMCMC. Consider a multimodal example where each distribution $\pi_t = \sum_{s=1}^{S} w^{(s)} h_t^{(s)}$ is a mixture of S components $\{h_t^{(s)} : s = 1, \ldots, S\}$, where each probability density $h_t^{(s)} = \frac{1}{\delta_t} h^{(s)}\left(\mu_s + \frac{\cdot - \mu_s}{\delta_t}\right)$ converges to a Delta function centered at the mode μ_s of $h^{(s)}$ at rate $\delta_t \to 0$, as $t \to \infty$. As t grows, the S modes of π_t tend to be well-separated. For example, in the case when each $h^{(s)}$ is a normal density with different centers, the transition probability between different modes of a metropolis random walk decays exponentially fast in δ_t^{-2}. As a result, common MCMC algorithms might take an exponentially long time to explore the whole state space.

Assume that the goal is to estimate the mixing probabilities $(w^{(s)})$. For instances, mixture models and Bayesian model selections can be fit into this framework. As a result of the multimodality, most commonly used MCMC algorithms for sampling from π_t tend to be stuck in one of the S local modes for large t. This is a main motivation of applying L Markov chains in parallel in the SMCMC. Even though any single chain might be stuck in some local mode, the ensemble Θ_t still consists of nearly independent samples from π_t. Benefitted by the annealing effect, these chains as an ensemble have been shuffled by the frequent moves among the modes at early time. As an ensemble, roughly $Lw^{(s)}$ chains tend to get stuck in the sth local mode at t. Therefore, an estimator of $w^{(s)}$ can be formulated by counting the numbers of chains stuck in the sth mode.

More formally, the following lemma suggests that for any Markov chain convergent in the sense of (6.4), the above counting estimator of $w^{(s)}$ is consistent.

Lemma 31. *Assume that there exists $d_0 > 0$, so that $|\mu_s - \mu_t| \geqslant 3d_0$ for any $s \neq t$. Let $\hat{\pi}_t$ be an approximation of π_t so that $||\hat{\pi}_t - \pi_t|| \to 0$, as $t \to \infty$. If $\{\mu^{(t,l)} : l = 1, \ldots, L\}$ are L independent points sampled from $\hat{\pi}_t$ and $\hat{w}_t^{(s)} = \#\{l : |\mu^{(t,l)} - \mu_s| \leqslant d_0\}/L$, then as $t \to \infty$ and $L \to \infty$, $\hat{w}_t^{(s)} \to w^{(s)}$ in probability.*

The definition of $\hat{w}_t^{(s)}$ in Lemma 31 greatly simplifies the proof. In practice, (μ_s) are mostly unknown and one can calculate $\hat{w}_t^{(s)}$ as the proportion of points in the sth clusters of $\{\mu^{(t,l)}\}$. The corresponding consistency of the estimator can be obtained by modifying Lemma 31.

6.3.2 Constant parameter dimension d_t

We first focus on the case when the parameter size is fixed, i.e. J_t is the identity map. The following theorem provides guarantees for the convergence of SMCMC under certain conditions. We will use the convention that $\sum_{\varnothing} = 0$ and $\prod_{\varnothing} = 1$.

Theorem 32. *Assume the following conditions:*

1. *(Universal ergodicity) There exists $\epsilon_t \in (0, 1)$, such that for all $t > 0$ and $x \in E$, $||T_t(x, \cdot) - \pi_t|| \leqslant 2\rho_t$.*

2. *(Stationary convergence) The stationary distribution π_t of T_t satisfies $\alpha_t = \frac{1}{2}||\pi_t - \pi_{t-1}||$.*

Let $\epsilon_t = \rho_t^{m_t}$. Then for any initial distribution π_0, as $t \to \infty$

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s+1}^{t} \epsilon_u (1 - \alpha_u) \right\} \epsilon_s \alpha_s.$$

Furthermore, if $\lim_{t\to\infty} \alpha_t = 0$ and there exists an $\epsilon > 0$ such that $\epsilon_t \leqslant 1 - \epsilon$ for any $t \in \mathbb{N}$, then as $t \to \infty$, $||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \to 0$.

To illustrate the idea, we provide here a short proof for the above theorem with a weakened conclusion

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant 2 \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_t \right\} \alpha_s.$$

In fact, by the universally ergodicity condition and Lemma 43 in Section 6.6, for all $t > 0$ and any probability distribution p,

$$||Q_t \circ p - \pi_t|| = ||T_t^{m_t} \circ p - \pi_t|| \leqslant \epsilon_t ||p - \pi_t||. \tag{6.5}$$

A recursive application of (6.5) yields

$$\begin{aligned} ||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant & \epsilon_t ||Q_{t-1} \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \\ \leqslant & \epsilon_t ||Q_{t-1} \circ \cdots \circ Q_1 \circ \pi_0 - \pi_{t-1}|| + \epsilon_t ||\pi_t - \pi_{t-1}|| \\ \leqslant & \cdots \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_t \right\} ||\pi_s - \pi_{s-1}||, \end{aligned}$$

which completes the proof.

If m_t in the algorithm is chosen large enough so that

$$\sup_x ||T_t^{m_t}(x, \cdot) - \pi_t|| \leqslant 2\epsilon_t \leqslant 2(1 - \epsilon), \tag{6.6}$$

and $\lim_{t \to \infty} \alpha_t = 0$, then as $t \to \infty$,

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant 2 \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_t \right\} \alpha_s \leqslant 2 \sum_{s=1}^{t} (1 - \epsilon)^{t+1-s} \alpha_s \to 0.$$

In practice, we can choose m_t as in section 6.2.5, which provides good approximations to (6.6). Although T_t are required to be universally ergodic in the theorem, it might be possible to weaken the conditions to those in Theorem 46 with direct application of the coupling techniques in the proofs of Lemma 42 and Lemma 43. In this section, we focuses on the universally ergodic case for conciseness and easy exhibition. In Section 6.3.4, we will consider the more general geometrically ergodic

condition. Condition 2 is intuitively reasonable and can be verified for many problems. In this subsection, we provide such a verification for regular parametric cases in Lemma 33 below, where the Bernstein von-Mises theorem holds. In the next subsection when d_t is allowed to grow in t, we provide a verification for general models that may not have $n^{-1/2}$ convergence rate or Gaussian limiting distributions; for example, nonparametric models.

For simplicity, we illustrate this for a one dimensional case. Let $Y_1, \ldots, Y_n$ be i.i.d. f_θ, where f_θ is a density with respect to the Lebesgue measure λ and $\theta \in \mathbb{R}$. Let $l(y, \theta) = \log f(y|\theta)$ be the log likelihood function. We consider a regular parametric model (Lehmann and Casella, 1998), where f_θ satisfies the following conditions at the truth θ_0: 1. $\{y : f_\theta(y) > 0\}$ is the same for all θ; 2. $l(y, \theta)$ is three times differentiable with respect to θ in a neighborhood $(\theta_0 - \delta, \theta_0 + \delta)$; 3. If $\dot{l}(y, \theta)$, $\ddot{l}(y, \theta)$ and $\dddot{l}(y, \theta)$ denotes its first, second and third derivatives, then $E_{\theta_0}\dot{l}(Y, \theta)$ and $E_{\theta_0}\ddot{l}(Y, \theta)$ are finite and $\sup_{\theta \in (\theta_0 - \delta, \theta_0 + \delta)} |\dddot{l}(y, \theta)| \leqslant M(y)$ with $E_{\theta_0} M(Y) \leqslant \infty$; 4. The order of expectation and differentiation of $l(y, \theta)$ and $\dot{l}(y, \theta)$ at θ_0 is interchangeable; 5. $I = E_{\theta_0}(\dot{l}(y, \theta))^2 > 0$.

Lemma 33. *Assume the regularization conditions on f_θ. If Δ_t observations are added at time t, so that the sample size at time t is $n_t = \sum_{s=1}^{t} \Delta_s$, then $||\pi_t - \pi_{t-1}|| = O\big(\sqrt{\frac{\Delta_t}{n_t}}\big)$. In particular, if $\Delta_t = o(n_t)$, the stationary convergence condition in Theorem 32 holds.*

As a special case of Lemma 33, one can add one observation at each time, under which t is the sample size n and $||\pi_t - \pi_{t-1}|| = O(t^{-1/2})$. However, in the batch setting where the total sample size n is fixed, such an updating scheme might not be optimal when taking the time consumption into account because the additional gain $(1 - \epsilon)||\pi_t - \pi_{t-1}|| = (1 - \epsilon)O(t^{-1/2})$ of performing the transition operator T_t decays as $n_t = t$ increases. As a result, there exists a trade-off between the increasing rate of

n_t (or Δ_t) and the decaying rate of α_t. First, we look for a theoretical upper bound for n_t. Consider the extreme case when $\alpha_t = \alpha$ for any $t \in \mathbb{N}$. Under such a case, we have $\Delta_t = \alpha^2(n_{t-1} + \Delta_t)$ and hence

$$\begin{aligned}\Delta_t =& \frac{\alpha^2}{1-\alpha^2} n_{t-1} = \frac{\alpha^2}{1-\alpha^2} n_{t-2} + \frac{\alpha^2}{1-\alpha^2}\Delta_{t-1} = \frac{\alpha^2}{1-\alpha^2} n_{t-2} + \frac{\alpha^4}{(1-\alpha^2)^2} n_{t-2} \\ =& \frac{\alpha^2}{(1-\alpha^2)^2} n_{t-2} = \cdots = \frac{\alpha^2}{(1-\alpha^2)^{t-1}} n_1.\end{aligned}$$

This implies that $n_t = \sum_{s=1}^{t} \Delta_s = n_1\big[(1-\alpha^2)^{-t+1} - (1-\alpha^2)\big] \sim \exp(Dt)$ with $D = -\log(1-\alpha^2) > 0$. This upper bound cannot be improved. In fact, for any $q > 1$ and any $C > 0$, we can choose $n_t = (Ct)^q$ so that $\Delta_t = (Ct)^q - (C(t-1))^q \leqslant qC^q t^{q-1}$ and $\alpha_t = O(\sqrt{qt^{q-1}/t^q}) = O(q^{1/2}t^{-1/2}) \to 0$ as $t \to \infty$. Therefore, for any fixed $K \in \mathbb{N}$, we can also choose $n_t = \sum_{k=0}^{K} \frac{1}{k!}(Ct)^k$ so that $\lim_{t\to\infty} \alpha_t = 0$. Such an n_t can be arbitrarily close to $\exp(Ct)$ by choosing a sufficiently large K.

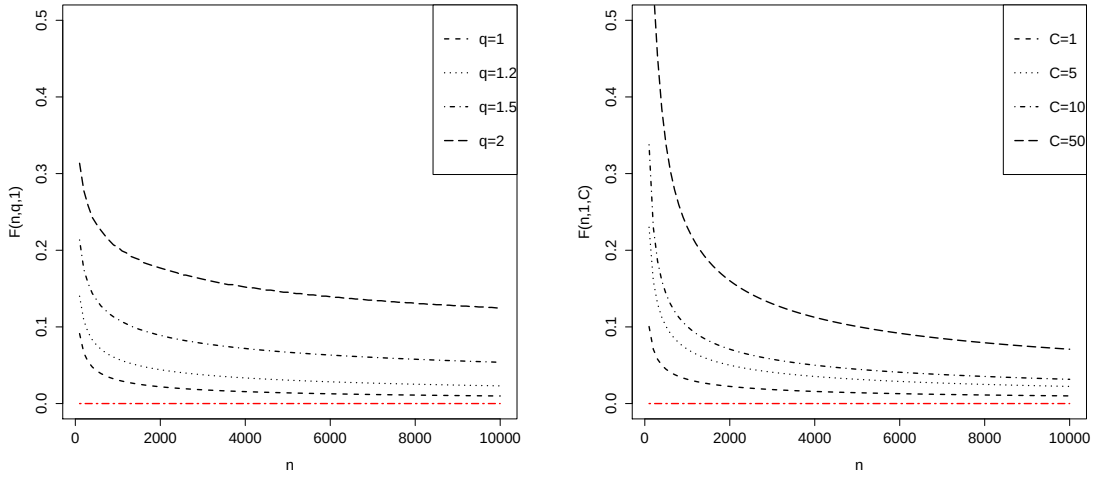


FIGURE 6.1: A plot of the error upper bound $F(n, q, C)$ as a function of (n, q, C) provided by Theorem 32.

Consider the batch setting where the total sample size n is fixed. Denote $F(n, q, C)$ be the error upper bound provided by Theorem 32 when $\epsilon = \frac{1}{2}$. We consider two

special case: 1. $C = 1$ and $n_t = t^q$ so that $\alpha_t = O(q^{1/2}t^{-1/2})$ and the total steps $T = O(n^{1/q})$; 2. $q = 1$ and $n_t = Ct$ so that $\alpha_t = O(t^{-1/2})$ and the total steps $T = O(n/C)$. The left panel in Figure 6.1 plots $F(n, q, C)$ as a function of n under $C = 1$ and $q \in \{1, 1.2, 1.5, 2\}$, where $q = 1$ corresponds to adding one observation each time. The right panel in Figure 6.1 plots $F(n, q, C)$ as a function of n under $q = 1$ and $C \in \{1, 5, 10, 50\}$, where $C = 1$ corresponds to adding one observation each time. As expected, the error bound decays slower when the batch size Δ_t increases in t than when Δ_t keeps constant. However, the total step T in the former is smaller than that in the latter. Therefore, there always exists a tradeoff between the computational complexity and the approximation accuracy.

From Figure 6.1, even in the worst case displayed, the error bound ε_t is less than 0.2. Let $\hat{\pi}_t$ denote an approximation of π_t. The following lemma suggests that for regular parametric models, as long as $\varepsilon < 1/2$, the error of an point estimator constructed by $\hat{\pi}$ is comparable to the statistical variation of the asymptotically optimal point estimator, such as the maximum likelihood estimator (MLE). Moreover, the coverage of the credible intervals created via $\hat{\pi}$ is of the same order as ε, which suggests that the uncertainty magnitude provided by $\hat{\pi}$ is correct. We will use z_α to denote the α-th quantile of the standard normal distribution. If $\alpha < 0$ or $\alpha > 1$, then we define $z_\alpha = \infty$.

Lemma 34. *Consider estimating the parameter θ of a regular parametric family $\{f_\theta\}$. Assume $||\hat{\pi}_t - \pi_t|| \leqslant \varepsilon \leqslant 1/2$. Then there exists an estimator $\tilde{\theta}_t$ based on $\hat{\pi}_t$, such that $|\hat{\theta}_t - \theta_0| = O_P\big(z_{0.5+\iota+\varepsilon} n_t^{-1/2}\big)$, where $\iota = O_P(n_t^{-1/2})$ and θ_0 is the true underlying parameter that generates the data. Moreover, let A_α be any α credible region created by $\hat{\pi}_t$, then $P_0(\theta_0 \in A_\alpha) = \alpha + O_P(\varepsilon) + O_P(\iota)$.*

Consider the case when n_t is large so that $\iota \ll \varepsilon$. If $\varepsilon = 0.2$, then $z_{0.5+\varepsilon} \approx 0.52$, suggesting that in terms of the accuracy of point estimation, using $\hat{\pi}_t$ is almost as

good as using π_t. Under the same error level, $P_0(\theta_0 \in A_{0.95}) \geqslant 0.95 - 0.2 = 0.75$, which is still a satisfactory coverage for a 0.95 credible interval. Therefore, Lemma 34 suggests that excessive reduction in the approximation error $||\hat{\pi}_t - \pi_t||$ is unnecessary in improving the estimation accuracy and it is enough to just keep this error below some pre-specified level, for example, 0.05.

6.3.3 Increasing parameter dimension d_t

Recall that the parameter at t can be written as $\theta^{(t)} = (\theta^{(t-1)}, \eta_t)$. Consider the J_t satisfying (6.1) in section 6.2.3 and $Q_t = J_t \circ T_t^{m_t}$. The following lemma links the approximation errors before and after applying the jumping kernel J_t.

Lemma 35. *For any probability density $p(\cdot)$ for $\theta^{(t-1)}$, the following holds:*

$$||\pi_t - J_t \circ p|| \leqslant ||\pi_{t-1} - p|| + \sup_{\theta^{(t-1)} \in \mathbb{R}^{d_{t-1}}} ||\pi_t(\cdot|\theta^{(t-1)}) - J_t(\theta^{(t-1)}, \cdot)||,$$

where the π_t in the second term of the right hand side stands for the marginal posterior of $\theta^{(t-1)}$ at time t.

If a Gibbs or slice sampling step is applied as J_t, then the last term in the above lemma vanishes. With Lemma 35, we can prove the following analogue of Theorem 32 for the increasing d_t scenario.

Theorem 36. *Assuming the following conditions:*

1. *(Universal ergodicity) There exists $\epsilon \in (0, 1)$, such that for all $t > 0$ and $x \in E$, $||T_t(x, \cdot) - \pi_t|| \leqslant 2\rho_t$.*

2. *(Stationary convergence) The stationary distribution π_t of T_t satisfies $\alpha_t = \frac{1}{2}||\pi_t - \pi_{t-1}||$, where π_t is the marginal posterior of $\theta^{(t-1)}$ at time t in α_t.*

3. *(Jumping consistency) For a sequence of τ_t, $\sup_{\theta^{(t-1)} \in \mathbb{R}^{d_{t-1}}} ||\pi_t(\cdot|\theta^{(t-1)}) - J_t(\theta^{(t-1)}, \cdot)|| \leqslant 2\tau_t$.*

Let $\epsilon_t = \rho_t^{m_t}$. *Then for any initial distribution* π_0,

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_u \right\} (\alpha_s + \tau_s).$$

Furthermore, if $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \tau_t = 0$, *and there exists an* $\epsilon > 0$ *such that* $\epsilon_t \leqslant 1 - \epsilon$ *for any* $t \in \mathbb{N}$, *then as* $t \to \infty$, $||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \to 0$.

Similarly, if we choose m_t such that $\sup_x ||T_t^{m_t}(x, \cdot) - \pi_t|| \leqslant 2(1 - \epsilon)$, then $||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \to 0$, as $t \to \infty$.

An increasing parameter dimension often occurs in Bayesian nonparametric models, such as Dirichlet mixture models and Gaussian process regressions. The following lemma is a counterpart of Lemma 33 for general models that may not have $n^{-1/2}$ convergence rate or normal as limiting distribution for the parameters. A function f defined on a Banach space $(V, ||\cdot||)$ is said to be Fréchet differentiable at $v \in V$ if there exists a bounded linear operator $A_v : V \to \mathbb{R}$ such that

$$f(v + h) = f(v) + A_v(h) + o(||h||), \text{ as } ||h|| \to 0,$$

where A_v is called the Fréchet derivative of f. For V being a Euclidean space, Fréchet differentiability is equivalent to the usual differentiability. The proof utilizes the notion of posterior convergence rate (Ghosal et al., 2000) and Fréchet differentiability.

Lemma 37. *Consider a Bayesian model* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ *with a prior measure* Π *on a Banach space* $(\Theta, ||\cdot||)$, *where the parameter space* Θ *can be infinite dimensional. Let* p_θ *be the density of* P_θ *with respect to some base measure* m. *Assume the following conditions:*

1. *the posterior convergence rate of the Bayesian model is at least* $\epsilon_n \to 0$ *as* $n \to \infty$, *i.e. the posterior satisfies*

   $$\Pi(||\theta - \theta_0|| > M\epsilon_n | Y_1, \ldots, Y_n) \to 0, \text{ in probability,}$$

   where $Y_1, \ldots, Y_n$ *is the observation sequence generated according to* P_{θ_0}, $M > 0$ *is a constant.*

2. *Assume that*

$$\max\Big[E_{[\theta|Y_1,\ldots,Y_n]}\{p_\theta(Y)I(||\theta-\theta_0||>M\epsilon_n)\},$$
$$E_{[\theta|Y_1,\ldots,Y_n]}\{\log p_\theta(Y)I(||\theta-\theta_0||>M\epsilon_n)\}\Big]\to 0 \text{ in probability},$$

where $Y\sim P_{\theta_0}$ *is independent of* $Y_1,\ldots,Y_n$ *and the expectation is taken with respect to the posterior distribution* $\Pi(\theta|Y_1,\ldots,Y_n)$ *for* θ.

3. *Also assume that the log likelihood function* $\log p_\theta(y)$ *is Fréchet differentiable at* θ_0 *with a Fréchet derivative* $A_{v,y}$ *satisfying* $E_{\theta_0}||A_{v,Y}||<\infty$, *where* $||\cdot||$ *is the induced operator norm and the expectation is taken with respect to* $Y\sim P_{\theta_0}$.

Then

$$||\pi(\cdot|Y_1,\ldots,Y_n)-\pi(\cdot|Y_1,\ldots,Y_{n-1})||\to 0,\ \text{as } n\to\infty.$$

The second assumption strengthens the first assumption in terms of the tail behavior of the posterior distributions and can be implied by the first if both $p_\theta(y)$ and $\log p_\theta(y)$ are uniformly bounded; for example, when Θ is compact. Since the primary goal of this chapter is the investigation of the SMCMC, we will not pursue a weakest conditions for Lemma 37 here.

The following corollary is an easy consequence of the above lemma by using the inequality $|\int f(x)\lambda(dx)|\leqslant\int|f(x)|\lambda(dx)$.

Corollary 38. *Let* ξ *be a* d_0 *dimensional component of* θ. *Denote the marginal posterior of* ξ *by* $\pi_\xi(\cdot|Y_1,\ldots,Y_n)$. *Then under the conditions in Lemma 37, we have*

$$||\pi_\xi(\cdot|Y_1,\ldots,Y_n)-\pi_\xi(\cdot|Y_1,\ldots,Y_{n-1})||\to 0,\ \text{as } n\to\infty.$$

In the case when T_t corresponds to the transition kernel of a Gibbs sampler, we can consider the marginal convergence of some fixed d_0 dimensional component ξ of θ, for example, for θ in function spaces, ξ can be the evaluations $\theta(x_1,\ldots,x_{d_0})$ on d_0 fixed points $x_1,\ldots,x_{d_0}$ in the domain of θ. Due to the special structure of the

graphical representation for a Gibbs sampler, the process of $\{\xi_s : s \geqslant 0\}$ obtained by marginalizing out other parameters in the Gibbs sampler with transition kernel T_t is still a Markov chain with another transition kernel $T_{\xi,t}$ defined on $\mathbb{R}^{d_0} \times \mathbb{R}^{d_0}$. Therefore, with this marginalized process for ξ, we can combine Theorem 32 and Corollary 38 to prove the marginal convergence of the posterior for the fixed dimensional parameter ξ under the new transition kernels $T_{\xi,t}$'s. To ensure the convergence of this marginal chain, m_t can also be chosen by the procedures in section 6.2.5, but only including the components of ξ in the calculations of (6.2).

6.3.4 Weakening the universal ergodicity condition

Both Theorem 32 and 36 rely on the strong condition of universal ergodicity. In this subsection, we generalize these results to hold under the weaker geometrically ergodic condition. We will use the following sufficient condition for geometric ergodicity (Roberts and Rosenthal, 1997) for an irreducible, aperiodic Markov chain with transition kernel T: there exists a π-a.e.-finite measurable function $V : E \to [1, \infty]$, which may be taken to satisfy $\pi(V^k) < \infty$ for any $j \in \mathbb{N}$, such that for some $\rho < 1$,

$$||T^t(x, \cdot) - \pi(\cdot)||_V \leqslant V(x)\rho^t, \quad x \in E, \quad t \in \mathbb{N}, \tag{6.7}$$

where $||\mu(\cdot)||_V = \sup_{|f| \leqslant V} |\mu(f)|$ for any signed measure μ. When $V \equiv 1$, we return to the uniform ergodic case. The following lemma generalizes Lemma 42 and 43 to geometrically ergodic chains.

Lemma 39. *Let $\{X_t\}$ be a Markov chain on E, with transition kernel T and stationary distribution π. If there exists a function $V : E \to [1, \infty)$ and $\rho \in (0, 1)$ such that for all $x \in E$,*

$$||T(x, \cdot) - \pi(\cdot)||_V \leqslant V(x)\rho, \tag{6.8}$$

then $\{X_t\}$ is geometrically ergodic. Moreover, for any initial distribution p_0, we have

$$||T^t \circ p_0 - \pi||_V \leqslant \rho^t ||p_0 - \pi||_V, \quad x \in E, \quad t \in \mathbb{N}.$$

By taking $t = 1$ in (6.7), (6.8) is also a necessary condition for geometric ergodicity. Therefore, the above lemma provides a necessary and sufficient condition for geometric ergodicity, which extends Lemma 43. By the above lemma, we can generalize Theorem 32 as follows, where $d_t = d$, for any t.

Theorem 40. *Assuming the following conditions:*

1. *(Geometric ergodicity) There exists a function $V : \mathbb{R}^d \to [1, \infty)$, $C > 0$ and $\rho_t \in (0, 1)$, such that $\pi_t(V^2) = E_{\pi_t} V^2 \leqslant C$ for any t and for all $x \in \mathbb{R}^d$, $||T_t(x, \cdot) - \pi_t(\cdot)||_V \leqslant V(x)\rho_t$.*

2. *(Stationary convergence) The stationary distribution π_t of T_t satisfies $\alpha_t = 2\sqrt{C} d_H(\pi_t, \pi_{t-1})$, where d_H is the Hellinger distance defined by $d^2(\mu, \mu') = \int (\mu^{1/2}(x) - \mu'^{1/2}(x))^2 \lambda(dx)$.*

Let $\epsilon_t = \rho_t^{m_t}$. Then for any initial distribution π_0,

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_u \right\} \alpha_s.$$

Furthermore, if $\lim_{t\to\infty} \alpha_t = 0$ and there exists an $\epsilon > 0$ such that $\epsilon_t \leqslant 1 - \epsilon$ for any $t \in \mathbb{N}$, then as $t \to \infty$, $||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \to 0$.

The first condition in the theorem is a uniform geometric ergodic condition on the collection $\{T_t : t \in \mathbb{N}\}$ of transition kernels, where a common potential V exists. The second condition is true for those π_t's in Lemma 33 and 37. In fact, Lemma 33 uses the inequality $||\pi_t - \pi_{t-1}|| \leqslant d_H(\pi_t, \pi_{t-1})$ and proves $d_H(\pi_t, \pi_{t-1}) \to 0$. Lemma 37 proves $||\pi_t - \pi_{t-1}|| \leqslant 2\sqrt{K(\pi_t, \pi_{t-1})} \to 0$, where $K(p, q)$ is the Kullback-Leibler divergence and satisfies $d_H(p, q)^2 \leqslant K(p, q)$ for any probability densities p and q.

Similarly, we have the following counterpart for Theorem 36 under geometrically ergodic condition.

Theorem 41. *Assuming the following conditions:*

1. *(Geometric ergodicity) For each t, there is a function $V_t : \mathbb{R}^{d_t} \to [1, \infty)$, $C > 0$ and $\rho_t \in (0, 1)$, such that:*

   (a) $\pi_t(V_t^2) = E_{\pi_t} V_t^2 \leqslant C$ *for any* t;

   (b) $E_{\pi_t}\big[V_t(\theta^{(t)})|\theta^{(t-1)}\big] = V_{t-1}(\theta^{(t-1)})$, *where* $\theta^{(t)} = (\theta^{(t-1)}, \eta_t)$;

   (c) *for all* $x \in \mathbb{R}^{d_t}$, $||T_t(x, \cdot) - \pi_t(\cdot)||_{V_t} \leqslant V_t(x)\rho_t$.

2. *(Stationary convergence) The stationary distribution π_t of T_t satisfies $\alpha_t = 2\sqrt{C} d_H(\pi_t, \pi_{t-1})$, where π_t is the marginal posterior of $\theta^{(t-1)}$ at time t in α_t.*

3. *(Jumping consistency) For a sequence of τ_t, the following holds:*

$$\sup_{\theta^{(t-1)} \in \mathbb{R}^{d_{t-1}}} ||\pi_t(\cdot|\theta^{(t-1)}) - J_t(\theta^{(t-1)}, \cdot)||_{\tilde{V}_t} \leqslant \tau_t,$$

   where $\tilde{V}_t$ is defined on $\mathbb{R}^{d_t - d_{t-1}}$ by $\tilde{V}_t(\eta_t) = \int_{\mathbb{R}^{d_{t-1}}} V_t(\theta^{(t-1)}, \eta_t) d\theta^{(t-1)}$.

Let $\epsilon_t = \rho_t^{m_t}$. Then for any initial distribution π_0,

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_u \right\} (\alpha_s + \tau_s).$$

Furthermore, if $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \tau_t = 0$, and there exists an $\epsilon > 0$ such that $\epsilon_t \leqslant 1 - \epsilon$ for any $t \in \mathbb{N}$, then as $t \to \infty$, $||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \to 0$.

6.3.5 Relationship between Markov chain convergence rate and the autocorrelation function

The convergence results in the previous two subsections are primarily based on a coupling technique, which can provide explicitly quantitative convergence bounds for computation. The arguments in this subsection will mainly utilize functional analysis and operator theory, which can reveal the relationship between convergence rate and

maximal correlation between two states in the Markov chain. For background details, refer to chapter 12 in Liu (2001).

For a time homogeneous Markov chain $\{X_t : t = 0, 1 \ldots\}$ with transition kernel $T(x, y)$ and stationary distribution π, consider the space of all mean zero and finite variance functions under π

$$L_0^2(\pi) = \left\{ h(x) : \int h^2(x)\pi(x)\lambda(dx) < \infty \text{ and } \int h(x)\pi(x)\lambda(dx) = 0 \right\}.$$

Being equipped with the inner product

$$\langle h, g \rangle = E_\pi\{h(x) \cdot g(x)\}, \tag{6.9}$$

$L_0^2(\pi)$ becomes a Hilbert space. On $L_0^2(\pi)$, we can define two operators, called forward and backward operators, as

$$Fh(x) \triangleq \int h(y)T(x, y)\lambda(dy) = E\{h(X_1)|X_0 = x\},$$

$$Bh(y) \triangleq \int h(y)\frac{T(x, y)\pi(x)}{\pi(y)}\lambda(dy) = E\{h(X_0)|X_1 = y\}.$$

The operator F can be considered as the continuous state generalization of the transition matrix T for finite state Markov chain (with Tv as the operation on vector space). Similarly, the operator B can be considered as the generalization of the transpose of T. With this definition, we can see that

$$E\{h(X_t)|X_0 = x\} = F^t h(x) \text{ and } E\{h(X_0)|X_t = y\} = B^t h(y).$$

Define the norm of an operator F to be the operator norm induced by the $L_0^2(\pi)$ norm defined in (6.9). By iterative variance formula

$$\text{var}\{h(X_1)\} = E[\text{var}\{h(X_1|X_0)\}] + \text{var}[E\{h(X_1)|X_0\}] \leqslant \text{var}[E\{h(X_1)|X_0\}],$$

and hence the norm of F and B are both less than or equal to one. By the Markov property, F and B are adjoint to each other, i.e. $\langle Fh, g \rangle = \langle h, Bg \rangle$. Since nonzero

constant functions are excluded from $L_0^2(\pi)$, the spectral radius r_F of F is strictly less than one under mild conditions (Liu et al., 1995), which is defined by $r_F = \lim_{t\to\infty} ||F^t||^{\frac{1}{t}} < 1$. Lemma 12.6.3 in Liu (2001) provides a Markov chain convergence bound in terms of the operator norm of F^t,

$$||T^t \circ p_0 - \pi||_{L^2(\pi)} \leqslant ||F^t|| \cdot ||p_0 - \pi||_{L^2(\pi)}, \tag{6.10}$$

where $||p-\pi||^2_{L^2(\pi)} = \int (p(z)-\pi(z))^2/\pi(z)\lambda(dz)$ and $||p-\pi|| \leqslant ||p-\pi||_{L^2(\pi)}$ holds for any probability measure p. Theorem 2.1 in Roberts and Rosenthal (1997) shows that if (6.10) is true for a time reversible Markov chain with transition kernel T, then the chain is geometric ergodic with that same rate function, i.e. there exists a potential function $V : E \to [1,\infty]$, such that $||T^t(x,\cdot) - \pi(\cdot)|| \leqslant V(x)||F^t||, \quad x \in \mathbb{X}$. Therefore, (6.10) implies a geometric convergence in L_1 norm with rate function $r(t) = ||F^t|| \sim r_F^t$. On the other side, by Lemma 12.6.4 in Liu (2001),

$$\sup_{g,h\in L^2(\pi)} \text{corr}\{g(X_0), h(X_t)\} = \sup_{||g||=1,||h||=1} \langle F^t h, g\rangle = ||F^t||. \tag{6.11}$$

This suggests the maximal autocorrelation function is of the same decay rate as the rate function $r(t)$. In practice, for multidimensional process $X_t = (X_{1,t}, \ldots, X_{p,t})$, the above quantity can often be well approximated by $\max_{j=1,\ldots,p} |\text{corr}\{X_{j,0}, X_{j,t}\}|$. Therefore, the maximal sample autocorrelation function provides a quantitative description of the mixing rate of the Markov chain, which provides the rationale for our choice of m_t in section 6.2.5.

If the Markov chain is reversible, then $F = B$ and hence F is self-adjoint. Under the further assumption that F is compact, $||F^t|| = |\tau_1|^t$, where $|\tau_1| \geqslant |\tau_2| \geqslant \cdots$ are the discrete eigenvalues of F. Therefore the rate function would be $r(t) = |\tau_1|^t$. For any $h(x) \in L_0^2(\pi)$, define the autocorrelation function as $f(t) = \text{corr}\{h(X_t), h(X_0)\}, t \geqslant 1$. Let $\alpha_1(x), \alpha_2(x), \ldots$ be the corresponding eigenfunctions. Then as long as $\langle h, \alpha_1\rangle \neq 0$, we have $\lim_{t\to\infty}\{|f(t)|\}^{1/t} = |\tau_1|$, which implies that the autocorrelation function and the rate function are very similar, i.e. $f(t) \sim r(t) \sim |\tau_1|^t$.

6.4 Simulation with finite Gaussian mixtures

The mixing rate of Gibbs samplers are notoriously slow for mixture models (Jasra et al., 2005). As an illustrative example, we consider the Bayesian Gaussian mixture model of Richardson and Green (1997), which is also considered by Del Moral et al. (2006) as a benchmark to test their method. Observations $y_1, \ldots, y_n$ are i.i.d. distributed as

$$[y_i \mid \mu_{1:k}, \tau_{1:k}, w_{1:k}] \sim \sum_{j=1}^{k} w_j N(\mu_j, \tau_j^{-1}), \tag{6.12}$$

where $\tau_{1:k}$ and $\tau_{1:k}$ are the means and inverse variances of k Gaussian components respectively, and $w_{1:k}$ are the mixing weights satisfying the constraint $\sum_{j=1}^{k} w_j = 1$. The priors for the parameters of each component $j = 1, \ldots, k$ are taken to be exchangeable as $\mu_j \sim N(\zeta, \kappa^{-1})$, $\tau_j \sim Ga(\alpha, \beta)$, $w_{1:k} \sim Diri(\delta)$, where $Ga(\alpha, \beta)$ is the gamma distribution with shape α and rate β and $Diri(\delta)$ is the Dirichlet distribution with number of categories k and concentration parameter δ. To enable a Gibbs sampler for the above model, we introduce for each observation $i = 1, \ldots, n$ a latent class indicator z_i such that

$$[y_i \mid z_i = j, \mu_{1:k}, \tau_{1:k}, w_{1:k}] \sim N(\mu_j, \tau_j^{-1}),$$
$$P(z_i = j | w_{1:k}) \propto w_j.$$

Then by marginalizing out z_i's, we can recover (6.12). With the above exchangeable prior, the joint posterior distribution $P(\mu_{1:k}|y_1, \ldots, y_n)$ of the k component means $\mu_{1:k}$ has $k!$ modes and the marginal posterior for each μ_j, $j = 1, \ldots, k$ is the same as a mixture of k components. Therefore, we can diagnose the performances of various samplers by comparing the marginal posteriors of $\mu_1, \ldots, \mu_k$. Standard MCMC algorithms tend to get stuck for long intervals in certain local modes, and even a very long run cannot equally explore all these modes (Jasra et al., 2005).

In this simulation, we generate the data with $n = 100$ samples and choose the true model as $k = 4$, $\mu_{1:4} = (-3, 0, 3, 6)$, $\tau_{1:4} = (0.55^{-2}, 0.55^{-2}, 0.55^{-2}, 0.55^{-2})$ and $w_{1:4} = (0.25, 0.25, 0.25, 0.25)$, which has the same settings as in Jasra et al. (2005) and Del Moral et al. (2006). The hyperparameters for the priors are: $\zeta = 0$, $\kappa = 0.01$, $\alpha = 1$, $\beta = 2$ and $\delta = 1$. We consider a batch setup with batch size (BS) $1, 2, 4, 6, 8$ and 10, which means that data arrive in batches of size BS. As a result, the algorithms operate $T = \lceil 100/BS \rceil = 100, 50, 25, 17, 13$ steps, where $\lceil x \rceil$ stands for the smallest integer no less than x.

In SMCMC, the dimension of the parameter $\theta^{(t)} = (\mu_{1:k}, \tau_{1:k}, w_{1:k}, z_{1:n_t})$ at time t is increasing when the latent class indicators $z_{1:t}$ are included, where $n_t = 0$ for $t = 0$ or $100 - BS \cdot (T - t)$ for $t = 1, \ldots, T$ is the data size at time t. We choose the transition kernel T_t to correspond to that for the Gibbs sampler. The jumping kernel J_t is the conditional distribution for the additional latent indicators of $y_{(n_{t-1}+1):n_t}$ given $\theta^{(t)}$ and $y_{1:n_t}$. Note that z_i are conditionally independent of z_j for $i \neq j, i, j \leqslant n_t$ given $(\mu_{1:k}, \tau_{1:k}, w_{1:k}, y_{1:n_t})$.

We compare SMCMC with two competitors. The first algorithm is the sequential Monte Carlo (SMC) sampler in Del Moral et al. (2006), which avoids data augmentation and works directly with the posterior of $(\mu_{1:k}, \tau_{1:k}, w_{1:k})$ using MH kernels. The second algorithm is the parallel Gibbs sampler (Richardson and Green, 1997) running on the full data $y_1, \ldots, y_n$, with L Gibbs samplers running in parallel, whose iterations K_{BS} equal the total Gibbs steps $\sum_{t=1}^{T} m_t$ in the SMCMC with batch size BS. The posterior distribution of each μ_j with $j = 1, 2, 3, 4$ is approximated by the empirical distribution of the L samples at K_{BS}th iteration in parallel. To demonstrate the annealing effect of SMCMC, the initial distributions of the L chains for both SMCMC and MCMC (parallel Gibbs) are centered at $(-3, 0, 3, 6)$. As a result, if no pair of labels are switched, the posterior draws will be stuck around the local mode centered at $(-3, 0, 3, 6)$, which is one of the $4! = 24$ local modes.

To compare the three algorithms, we calculate the averages of sorted estimated means across 10 trials under each setting as shown in Table 6.1. More specifically, we sort the estimated posterior means of $\mu_{1:4}$ in increasing order for each run and then average the 4 sorted estimates over 10 replicates. A good algorithm is expected to provide similar posterior means of $\mu_{1:4}$, which is approximately 1.5 in our case. The purpose for sorting the estimated means is to prevent the differences in the estimated posterior means being washed away from averaging across 10 replicates.

As can be seen from Table 6.1, SMCMC outperforms both SMC and MCMC under each setting and has satisfactory performance even when the batch size is 6, i.e. the number of time steps T is 17. Moreover, the performance of SMCMC appears stable as the batch size grows from 1 to 6, and become worse when the batch size increases to 8 and 10. A similar phenomenon is observed for SMC, with performance starting to deteriorate at batch size 6. MCMC has slightly worse performance with batch size 1 than SMCMC. However, its performance rapidly becomes bad as the number of iterations decreases. The comparison between SMCMC and MCMC illustrates substantial gains due to annealing for our method.

Figure 6.2 displays some summaries for SMCMC with batch size 1. The left plot shows the number of Gibbs iteration m_t versus time t (which is equal to the sample size at time t). m_t increases nearly at an exponential rate, which indicates the slow mixing rate of the Gibbs sampler used to construct the transition kernels T_t. As a by product of SMCMC, we can assess the convergence rate of the sampler used to construct T_t as a function of the sample size. The automatic mixing diagnostics procedure guarantees the convergence of the approximated posterior as $t \to \infty$. The right panel shows the "traceplot" for $\mu_{1:k}$ for one Markov chain among the L chains. This is not the usual traceplot since we selected the last samples of $\mu_{1:k}$ at each time t, where $\mu_{1:k}$ is approximately distributed according to a time changing posterior π_t. This "traceplot" suggests satisfactory mixing of $\mu_{1:k}$, i.e. frequent moves between

Table 6.1: Averages of sorted estimated means in mixture model by three approaches. We ran each algorithm 10 times with 1000 Markov chains or particles. We sorted the estimated means in increasing order for each run and then averaged the sorted estimates over 10 replicates. The last column reports the sample standard deviations of the first 4 numbers displayed. In the parenthesis following MCMC are the number of iterations it runs, which is equal to the average iteration the corresponding SMCMC runs across 10 replicates.

Algorithm description	Averages of sorted estimated component means				standard deviation
	μ_1	μ_2	μ_3	μ_4	
SMCMC (batch size 1)	1.38	1.50	1.57	1.67	0.12
SMC (batch size 1)	1.13	1.37	1.60	1.97	0.36
MCMC (8621 iterations)	1.31	1.42	1.56	1.77	0.20
SMCMC (batch size 2)	1.40	1.50	1.56	1.66	0.11
SMC (batch size 2)	1.22	1.46	1.75	1.99	0.34
MCMC (4435 iterations)	0.91	1.12	1.30	2.69	0.81
SMCMC (batch size 4)	1.42	1.50	1.54	1.64	0.09
SMC (batch size 4)	1.57	1.84	2.01	2.32	0.31
MCMC (2367 iterations)	0.23	0.71	1.20	3.34	1.37
SMCMC (batch size 6)	1.36	1.48	1.59	1.65	0.13
SMC (batch size 6)	1.31	1.63	1.93	2.35	0.44
MCMC (1657 iterations)	-0.23	0.53	1.32	4.45	2.05
SMCMC (batch size 8)	1.35	1.45	1.54	1.73	0.16
SMC (batch size 8)	1.43	1.69	1.99	2.35	0.40
MCMC (1390 iterations)	-0.50	0.53	1.36	4.68	2.24
SMCMC (batch size 10)	1.19	1.32	1.57	2.04	0.37
SMC (batch size 10)	1.36	1.69	1.98	2.38	0.43
MCMC (1069 iterations)	-1.00	0.38	1.60	5.11	2.62

the modes.

6.5 Sequential Bayesian estimation for heart disease data

In the following we apply SMCMC to a sequential, growing dimension nonparametric problem. We consider nonparametric probit regression with a Gaussian process (GP) prior. Let $y_1, y_2, \ldots$ be a sequence of binary responses and $x_1, x_2, \ldots$ the p dimensional covariates. The model assumes $P(y_i = 1) = \Phi(f(x_i))$, where Φ is the cdf of the standard normal distribution and f is a d-variate nonlinear function. We choose

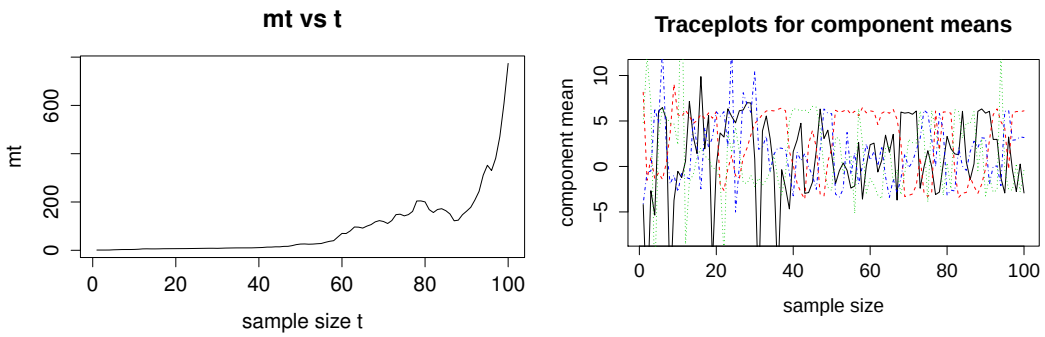


FIGURE 6.2: Summaries of SMCMC with batch size 1. The left panel displays the plot of the number of Gibbs iterations m_t versus time t (which is equal to the sample size at time t). The right panel displays the last samples of $\mu_{1:k}$ at each time t in one of L Markov chains.

a GP as a prior, $f \sim GP(\kappa, K)$, with mean function $\kappa : \mathbb{R}^p \to \mathbb{R}$ and covariance function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We consider the squared exponential kernel $K_a(x, x') = \sigma^2 \exp\{-a^2||x - x'||^2\}$ with a powered gamma prior on the inverse bandwidth, which leads to an adaptive posterior convergence rate (van der Vaart and van Zanten, 2009).

The computation of the nonparametric probit model can be simplified by introducing latent variables z_i such that

$$\begin{aligned} P(y_i = 1) &= I(z_i > 0), \\ z_i &= f(x_i) + \epsilon_i, \epsilon_i \sim N(0, 1). \end{aligned} \tag{6.13}$$

The model has simple full conditionals so that a Gibbs sampler can be used to sample the z_i's and $F_t = \{f(x_1), \ldots, f(x_t)\}$.

To alleviate the $O(n^3)$ computational burden of calculating inverses and determinants of $n \times n$ covariance matrices, we use a discrete prior to approximate the powered gamma prior for a and pre-compute those inverses and determinants over the pre-specified grid. We combine the sequential MCMC with the following off-line sequential covariance matrix updating.

Let $a_1, \ldots, a_H$ denote a grid of possible inverse bandwidths. For example, a_h can be chosen as the $\frac{h-1}{H}$th quantile of the powered gamma prior and the discrete

prior as the uniform distribution over $a_1, \ldots, a_H$. Let $C_h(x, x') = \exp\{-a_h^2||x - x'||^2\}$ and $K_{a_h} = \sigma^2 C_h$. We use the notation $C(A, B)$ to denote the matrix $(c(a_i, b_j))_{p,q}$ for a function $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and matrices $A \in \mathbb{R}^{p\times d}$, $B \in \mathbb{R}^{q\times d}$. Let $X_t = (x_1^T, \ldots, x_t^T)^T \in \mathbb{R}^{t\times d}$, $Y_t = (y_1, \ldots, y_t)$ and $Z_t = (z_1, \ldots, z_t)$ be the design matrix, response vector and latent variable vector at time t. At time t, for each $h = 1, \ldots, H$, we update the lower triangular matrix $L_h^{(t)}$ and $(L_h^{(t)})^{-1}$ in the Cholesky decomposition $C_h^{(t)} = L_h^{(t)}(L_h^{(t)})^T$ of the $t \times t$ correlation matrix $C_h^{(t)} = C_h(X_t, X_t)$. The reason is two-fold: 1. inverse and determinant can be efficiently calculated based on $L_h^{(t)}$ and $(L_h^{(t)})^{-1}$; 2. due to the uniqueness of Cholesky decomposition, $L_h^{(t+1)}$ and $(L_h^{(t+1)})^{-1}$ can be simply updated by adding $(t+1)$th row and column to $L_h^{(t)}$ and $(L_h^{(t)})^{-1}$. More precisely, if $L_h^{(t+1)}$ and $(L_h^{(t+1)})^{-1}$ are written in block forms as

$$L_h^{(t+1)} = \begin{pmatrix} L_h^{(t)} & 0 \\ B_h^{(t+1)} & d_h^{(t+1)} \end{pmatrix} \text{ and } (L_h^{(t+1)})^{-1} = \begin{pmatrix} (L_h^{(t)})^{-1} & 0 \\ E_h^{(t+1)} & g_h^{(t+1)} \end{pmatrix},$$

where $B_h^{(t+1)}$ and $E_h^{(t+1)}$ are t-dimensional row vectors and $d_h^{(t+1)}$ and $g_h^{(t+1)}$ are scalars, then we have the following recursive updating formulas: for $h = 1, \ldots, H$,

$$\begin{aligned} d_h^{(t+1)} =& \big\{C_h(x_{t+1}, x_{t+1}) - C_h(x_{t+1}, X_t)(L_h^{(t)})^{-T}(L_h^{(t)})^{-1}C_h(X_t, x_{t+1})\big\}^{1/2}, \\ B_h^{(t+1)} =& C_h(x_{t+1}, X_t)(L_h^{(t)})^{-1}, \\ g_h^{(t+1)} =& (d_h^{(t+1)})^{-1}, \\ E_h^{(t+1)} =& -g_h^{(t+1)} C_h(x_{t+1}, X_t)(L_h^{(t)})^{-T}(L_h^{(t)})^{-1}, \end{aligned}$$

where for a matrix A, A^{-T} is a shorthand for the transpose of A^{-1}. The computation complexity of the above updating procedure is $O(t^2)$.

As t increases to $t+1$, the additional component η_{t+1} is $(f(x_{t+1}), z_{t+1})$. Therefore, in the jumping step of the sequential updating, we repeat drawing $f(x_{t+1})$ and z_{t+1} from their full conditionals in turn for r times. In our algorithm, we simply choose $r = 1$ as the results do not change much with a large r. In the transition step of the

sequential updating, each full conditional is recognizable under the latent variable representation (6.13) and we can run a Gibbs sampler at each time t. Predicting draws $f(x')$ on new covariates x' can be obtained based on posterior samples of F_t.

Note that the computational complexity for the off-line updating at time t is $O(t^2)$. Therefore the total complexity due to calculating matrix inversions and determinants is $O(\sum_{t=1}^{n} t^2) = O(n^3)$, which is the same as the corresponding calculations in the MCMC with all data. However, the proposed algorithm distributes the computation over time, allowing real-time monitoring and extracting of current information.

To illustrate the above approach, we use the south African heart disease data (Rousseauw et al., 1983; Hastie and Tibshirani, 1987) to study the effects of obesity and age on the probability of suffering from hypertension. The data contains $n = 462$ observations on 10 variables, including systolic blood pressure (sbp), obesity and age. A patient is classified as hypertensive if the systolic blood pressure is higher than 139 mmHg. We use $I(\text{sbp} > 139)$ as a binary response with obesity and age as a two-dimensional covariate x.

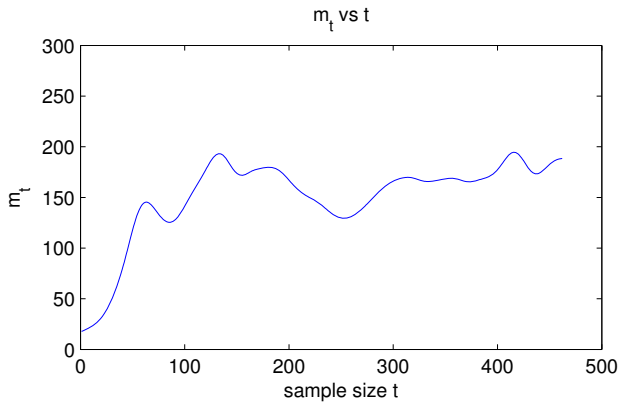


FIGURE 6.3: The iterations m_t at time t versus the sample size t is displayed. m_t has been smoothed with window width equal to 10.

Fig. 6.3 demonstrates the relationship between the number of iterations m_t and the sample size t. As can be seen, m_t keeps fluctuating between 150-200 as t becomes

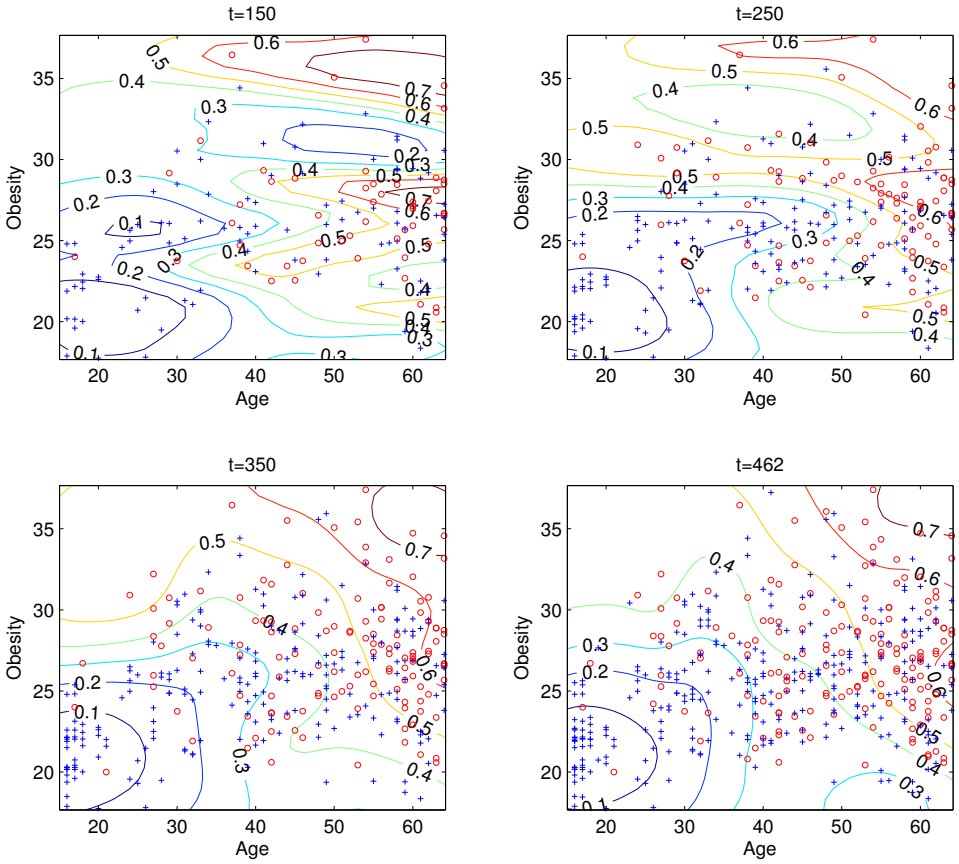


FIGURE 6.4: The fitted hypertension probability contours at $t = 150, 250, 350, 462$. The circles correspond to hypertensive patients and plus signs correspond to normal blood pressure people.

greater than 100, indicating that contrary to the mixture model example, the mixing rate of the above Markov chain designed for the nonparametric probit regression is robust to the sample size. The total number of iterations $\sum_{t=1}^{n} m_t$ is about 80k. However, the computation complexity of each SMCMC chain is much less than a 80k iterations full data MCMC since many iterations of SMCMC run with smaller sample sizes. In addition, we can reduce the iterations needed by increasing block sizes.

Fig. 6.4 shows the fitted probabilities of hypertension as a function of obesity and

age at $t = 150, 250, 350, 462$. With a relatively small sample size, the bandwidth a^{-1} tends to be small and the fitted probability contours are wiggly. As the sample size t increases, the bandwidth grows. As a result, contours begin to capture some global features and are less affected by local fluctuations. In addition, at large time point $t = 350$, the posterior changes little as the sample size further grows to $t = 462$. As expected, the probability of hypertension tends to be high when both obesity index and age are high. The gradient of the probability $P(\text{sbp} > 139|\text{obesity,age})$ as a function of obesity and age tends to be towards the 45-degree direction. The results in Fig. 6.4 are indistinguishable from those obtained running a long MCMC at each time, which are omitted here.

6.6 Convergence of Markov chain

In this section, we review some convergence results for Markov chains and introduce some new properties, which is applied to study the SMCMC convergence.

A transition kernel T is called uniformly ergodic if there exists a distribution π and a sequence $r(t) \to 0$, such that for all x, $||T^t(x, \cdot) - \pi|| \leqslant r(t)$, where $||\cdot||$ is the L_1 norm. $r(t)$ will be called the rate function. If T is ergodic, then π in the definition will be the stationary distribution associated with T. Uniformly ergodic implies geometric convergence, where $r(t) = \rho^t$ for some $\rho \in (0, 1)$ (Meyn and Tweedie, 1993).

We call a transition kernel T universally ergodic if there exists a distribution π and a sequence $r(t) \to 0$, such that for any initial distribution p_0, $||T^t \circ p_0 - \pi|| \leqslant r(t)||p_0 - \pi||$. $r(t)$ will also be called rate function. The concept of universal ergodicity plays an important role in the following study of the convergence properties of SMCMC. By choosing p_0 as a Dirac measure at x, one can see that universal ergodicity implies uniform ergodicity with rate function $2r(t)$. In addition, universal ergodicity can provide tighter bounds on the MCMC convergence than uniform ergodicity especially

when the initial distribution p_0 is already close to π. The following lemma provides the converse. The proof is based on coupling techniques.

Lemma 42. *If a transition kernel T is uniformly ergodic with rate function $r(t)$, then it is universally ergodic with the same rate function.*

The coupling in the proof of Lemma 42 is constructed through importance weights. By using the same technique, we can prove the uniform ergodicity for certain T as in the following lemma.

Lemma 43. *If the transition kernel T satisfies*

$$\sup_x ||T(x,\cdot) - \pi|| \leqslant 2\rho, \tag{6.14}$$

for some $\rho < 1$, then T is uniformly ergodic with rate function $r(t) = \rho^t$.

Note that condition (6.14) in the above lemma is weaker than the minorization condition (Meyn and Tweedie, 1993) for proving uniform ergodicity with rate function $r(t) = \rho^t$. The minorization condition assumes that there exists a probability measure ν such that,

$$T(x,y) \geqslant (1-\rho)\nu(y), \forall x, y \in E. \tag{6.15}$$

In practice, there is no rule on how to choose such measure ν. To see that (6.14) is weaker, first note that if (6.15) holds, then by the stationarity of π,

$$\pi(y) = \int T(x,y)\pi(x)\lambda(dx) \geqslant (1-\rho)\nu(y)\int \pi(x)\lambda(dx) = (1-\rho)\nu(y).$$

Therefore, for any $x \in E$, we have

$$\begin{aligned} ||T(x,\cdot) - \pi|| \leqslant & ||T(x,\cdot) - (1-\rho)\nu|| + ||\pi - (1-\rho)\nu|| \\ = & \int \big[T(x,y) - (1-\rho)\nu(y)\big]\lambda(dy) + \int \big[\pi(y) - (1-\rho)\nu(y)\big]\lambda(dy) \\ = & 1 - (1-\rho) + 1 - (1-\rho) = 2\rho. \end{aligned}$$

Therefore, condition (6.14) can lead to a tighter MCMC convergence bound than the minorization condition. Using $\sup_x ||T(x,\cdot)-\pi||$ in (6.14) also provides a tighter bound than using the Dobrushin coefficient $\beta(T) = \sup_{x,y} ||T(x,\cdot) - T(y,\cdot)||$, which is another tool used in studying the Markov chain convergence rate via operator theory. In fact, for any set $A \in \mathcal{E}$

$$\begin{aligned}\sup_x |T(x,A) - \pi(A)| &= \sup_x \left| \int_A \left\{ \int_E \pi(y)\big[T(x,z) - T(y,z)\big]\lambda(dy) \right\} \lambda(dz) \right| \\ &\leqslant \int_A \left\{ \int_E \pi(y) \big|T(x,z) - T(y,z)\big| \lambda(dy) \right\} \lambda(dz) \\ &\leqslant \beta(T)\pi(A),\end{aligned}$$

which implies that $\sup_x ||T(x,\cdot)-\pi|| \leqslant \beta(T)$. Moreover, comparing to the minorization condition and Dobrushin coefficient, (6.14) has a more intuitive explanation that the closer the transition kernel $T(x,\cdot)$ is to the stationary distribution, the faster the convergence of the Markov chain. Ideally, if $T(x,\cdot) = \pi(\cdot)$ for all $x \in E$, then the Markov chain converges in one step. The converse of Lemma 43 is also true as shown in the following lemma, which implies that condition (6.14) is also necessary for uniform ergodicity.

Lemma 44. *If T is uniformly ergodic, then there exists $\rho \in (0,1)$, such that*

$$\sup_x ||T(x,\cdot) - \pi|| \leqslant 2\rho. \tag{6.16}$$

When the condition (6.14) does not hold, we can still get a bound by applying the above coupling techniques. More specifically, assume $\{X_t : t \geqslant 0\}$ is a Markov chain with state space E, transition kernel T and initial distribution p_0 over E. Recall that π is the stationary measure associated with T. We define an accompanied transition kernel T' as

$$T'(x,y) = \frac{T(x,y) - \min\{T(x,y), \pi(y)\}}{\delta(x)},$$

where $\delta(x) = \frac{1}{2}||T(x,\cdot) - \pi||$. Let $\{X'_t : t \geqslant 0\}$ be another Markov chain with state space E, transition kernel T' and the same initial distribution p_0. The following lemma characterizes the convergence of X_t via $\tilde{X}_t$.

Lemma 45. *With the above notations and definitions, we have the following result:* $||T^t \circ p_0 - \pi|| \leqslant E\big\{\prod_{s=1}^{t} \delta(X'_s)\big\}$.

The Markov chain $\tilde{X}_t$ in the above proof is known as the trapping model in physics, where before getting trapped, a particle moves according to the transition kernel T' on E and every time the particle moves to a new location y, with probability $1 - \delta(y)$, it will be trapped there forever. Generally, the upper bound in Lemma 45 is not easy to compute. However, under the drift condition and an analogue of local minorization assumption (Rosenthal, 1995), we can obtain an explicit quantitative bound for MCMC convergence as indicated by the following theorem. The proof is omitted here, which is a combination of the result in Lemma 45 and the proof of Theorem 5 in Rosenthal (1995).

Theorem 46. *Suppose a Markov chain has transition kernel T and initial distribution p_0. Assume the following two conditions:*

1. *(Analogue of local minorization condition) There exists a subset $C \in \mathcal{E}$, such that for some $\rho < 1$, $\sup_{x \in C} ||T(x,\cdot) - \pi|| \leqslant 2\rho$.*

2. *(Drift condition) There exist a function $V : E \to [1, \infty)$ and constant b and $\tau \in (0,1)$, such that for all $x \in E$, $\int T(x,z)V(z)\lambda(dz) \leqslant \tau V(x) + b1_C(x)$.*

Then for any j, $1 \leqslant j \leqslant t$, $||T^t \circ p_0 - \pi|| \leqslant \rho^j + \tau^t B^{j-1}\bar{V}$, where $B = 1 + b/\tau$ and $\bar{V} = \int V(z)p_0(z)\lambda(dz)$.

By optimizing the j in the above theorem, we can obtain the following geometrically decaying bound on $||T^t \circ p_0 - \pi||$, which is similar to Rosenthal (1995):

$$||T^t \circ p_0 - \pi|| \leqslant \bar{V}\tilde{\rho}^t, \text{with } \log\tilde{\rho} = \frac{\log\rho\log\tau}{\log\rho - \log B}.$$

This implies that the Markov chain with transition kernel T is geometrically ergodic. Recall that a chain is geometrically ergodic if there is $\rho < 1$, and constants C_x for each $x \in E$, such that for $\pi-$a.e. $x \in E$, $||T^t(x, \cdot) - \pi(\cdot)|| \leqslant C_x \rho^t$.

6.7 Discussions

In this work, we proposed a sequential MCMC algorithm to sample from a sequence of probability distributions. Supporting theory is developed and simulations demonstrate the potential power of this method. The performance of SMCMC is closely related to the mixing behavior of the transition kernel T_t as $t \to \infty$. If T_t tends to have poor mixing as t increases, then updating the ensemble Θ_t every time a new data point arrives can lead to increasing computational burden over time. To alleviate this burden, we have three potential strategies. First, we can make the updating of Θ_t less frequent as t grows, i.e. updating Θ_t only at time $\{t_k : k = 1, \ldots\}$ with $t_k \to \infty$ as $k \to \infty$ and $t_k - t_{k-1} \to \infty$, as long as $||\pi_{t_k} - \pi_{t_{k-1}}|| \to 0$. Second, we can let the ϵ in Algorithm 1 decrease in t so that the upper bound in Theorem 32 still converges to zero. Third, we can develop 'forgetting' algorithms that only use the data within a window but still guarantee the convergence up to approximate error. The first two strategies may also be developed in an adaptive/dynamic manner, where the next step size $t_{k+1} - t_k$ or decay rate ϵ_{k+1} is optimized based on some criterion by using the past data and information.

7

Semiparametric Bernstein-von Mises Theorem: Second Order Studies

7.1 Introduction

Semiparametric modelling has provided a flexible and powerful modeling framework for modern complex data. Semiparametric models are indexed by a Euclidean parameter of interest $\theta \in \Theta \subset \mathcal{R}^k$ and an infinite-dimensional nuisance function η belonging to a Banach space $\mathcal{H}$. For example, in the Cox proportional hazards model, θ corresponds to the log hazard ratio for the regression covariate vector and η is the cumulative hazard function. In the partial linear model, θ corresponds to the regression coefficient vector for the linear component and η is the nonlinear component. By introducing a prior Π on $\Theta \times \mathcal{H}$, we are particularly interested in making Bayesian inferences for θ in semiparametric context. For example, we want to construt credible intervals for θ and test its significance using Bayes factors. These Bayesian inferences are known to be supported by the semiparametric Bernstain-von Mises (BvM) theorems (Shen, 2001; Bickel and Kleijn, 2012; Castillo and van der Vaart, 2012), which states that the marginal posterior distribution of θ converges (in

total variation norm) to a normal limit:

$$\sup_A \left|\Pi(\theta \in A|X_1,\ldots,X_n) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (n\widetilde{I}_{\theta_0,\eta_0})^{-1}\big)(A)\right| \overset{P_{\theta_0,\eta_0}}{\longrightarrow} 0, \tag{7.1}$$

where A is any measurable subset of Θ, $N_k(\mu,\Sigma)$ denotes a k-variate normal distribution with mean vector $\mu \in \mathbb{R}^k$ and variance-covariance matrix $\Sigma \in \mathbb{R}^{k\times k}$. P_{θ_0,η_0} is the true underlying distribution generating the data, where θ_0 and η_0 are the true values. Here, $\widetilde{l}_{\theta,\eta}$ is the efficient score function and $\widetilde{I}_{\theta,\eta}$ the efficient Fisher information evaluated at (θ,η) and

$$\widetilde{\Delta}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widetilde{I}^{-1}_{\theta_0,\eta_0}\widetilde{l}_{\theta_0,\eta_0}(X_i) \overset{P_{\theta_0,\eta_0}}{\rightsquigarrow} N(0,\widetilde{I}^{-1}_{\theta_0,\eta_0}). \tag{7.2}$$

Here, the notation " $\overset{P}{\rightsquigarrow}$ " and " $\overset{P}{\to}$ " denote the weak convergence and convergence in probability, respectively. A brief review of the semiparametric efficiency theory is provided in Section 7.2.1. We call (7.2) as the first order version of semiparametric BvM theorem.

The major goal of this chapter is the second order studies of semiparametric BvM theorem with an attempt to figure out the influence of nonparametric Bayesian prior on the semiparametric inference. Such results can provide us new theoretical insight, and can also be used to guide the choice of nonparametric prior. Cheng and Kosorok (2008a,b, 2009) derived the second order version of a special semiparametric BvM based on the posterior distribution of the profile likelihood in which the nuisance parameter is maximized out. In this case, no nonparametric prior is assigned. To the best of our knowledge, a comprehensive second order study of the general semiparametric Bernstein-von Mises theorem in the fully Bayesian setup does not exist.

The primary goal of this chapter is to formulate a set of necessary conditions for quantifying the second-order convergence rates of Bayesian semiparametric methods.

Intuitively, this set of conditions would be stronger than those for first-order BvM theorems. The first contribution of this chapter is that we derived the convergence rate (7.1) as $O_{P_{\theta_0,\eta_0}}(n^{1/2}\rho_n^2)$, with ρ_n the estimation error of the nuisance part. This second order term suggests that more accurate estimation of the nuisance parameter η could lead to better estimation efficiency of the parametric part. This is consistent with Cheng and Kosorok (2008b) and Cheng and Kosorok (2009) even the non-parametric prior is not assigned therein. In addition, we consider multi-dimensional nuisance function in this chapter. For example, in the partially linear model under penalization, the convergence rate for the nuisance parameter is $r = \alpha/(2\alpha + 1)$, where α is the known smoothness of the nuisance parameter. The set of conditions we formulated can also be used to study first order BvM results and appears to be weaker than that in Bickel and Kleijn (2012), where the very strong condition on the root-n convergence rate of the parametric part is replaced with a convergence rate of ρ_n.

Understanding of these conditions can conversely guide the design of the semi-parametric objective prior, by which we mean a prior that achieves the same second-order estimation and inference accuracy as frequentist approaches, such as the maximum penalized likelihood estimator (Cheng and Kosorok, 2009), could achieve. For example, a point estimator resulted from a semiparametric objective prior should match the corresponding frequentist estimators in terms of second order expansion and the resulted credible intervals/region should have the same accuracy of coverage compared to the corresponding confidence intervals. Another contribution of this chapter is to show that a new class of dependent priors for θ and η are semiparametric objective and the commonly used independent priors for θ and η might even break down the first-order consistency under some situations. The failure of independent priors has also been observed in a recent work by Castillo (2012), who proposes an interesting counter-example where the BvM does not hold due to a bias term appear-

ing in the posterior distribution. We will call such a bias term the semiparametric bias in the rest of the chapter. Intuitively, a non-negligible semiparametric bias is caused by a nonzero least favorable direction. Surprisingly, we show that by introducing prior dependence, the semiparametric bias can be eliminated by shifting the center of the prior for the nuisance parameter.

What is more surprising is that our adaptive semiparametric objective priors can be easily made adaptive. This is counted as our third contribution. In the first two conditions, we assume the smoothness α is known, which is unrealistic in reality. A third contribution of this chapter is to study the impact of the nonparametric adaptivity on the second order semiparametric efficiency under from a Bayesian perspective. Note that such nonparametric adaptive issues can only be investigated in the second order representation. Rivoirard and Rousseau (2012) propose a counter-example to rule out the BvM for independent adaptivity priors, where an independent prior achieves adaptive learning of η but fails to capture the semiparametric bias. This negative result on adaptive priors is first observed by Castillo (2012). In this chapter, we investigate sufficient conditions for a prior to be adaptive. Interestingly, we show that a dependent prior can achieve the adaptation to the second order term under mild conditions while an independent prior needs very stringent conditions.

7.2 Preliminaries

7.2.1 Semiparametric efficiency review

In this section, we review the semiparametric efficiency theory in a heuristic manner, and comment its connection to our results.

We briefly review the semiparametric efficiency theory. The score functions for θ and η are defined as

$$\dot{l}_{\theta_0,\eta_0}(X_i) = \frac{\partial}{\partial\theta}\bigg|_{\theta=\theta_0} l_{\theta,\eta}(X_i),\ A_{\theta_0,\eta_0}h(X_i) = \frac{\partial}{\partial\theta}\bigg|_{t=\theta_0} l_{\theta,\eta(t)}(X_i),$$

where h denotes a direction along which $\eta(t) \in \mathcal{H}$ approaches η_0 as $t \to \theta_0$ and $A_{\theta_0,\eta_0} : \mathbb{H} \mapsto L_2^0(P_0)$ is the score operator for η with $\mathbb{H}$ some closed and linear direction set, and $L_2^0(P_0) = \{f \in L_2(P_0) : P_0 f = 0\}$ is a subset of $L_2(P_0)$ equipped with the L_2-norm $||\cdot||_2$. The efficient score function $\tilde{l}_{\theta_0,\eta_0}$ is defined as the orthocomplement projection of $\dot{l}_{\theta_0,\eta_0}$ onto the tangent space $\mathcal{T}$, which is defined as the completion of the linear span of the tangent set $\{A_{\theta_0,\eta_0} h : h \in \mathbb{H}\}$. Therefore, the efficient score function at (θ_0, η_0) can be written as $\tilde{l}_{\theta_0,\eta_0} = \dot{l}_{\theta_0,\eta_0} - \Pi_0 \dot{l}_{\theta_0,\eta_0}$, where $\Pi_0 \dot{l}_{\theta_0,\eta_0} = \arg\min_{\tau \in \mathcal{T}} ||\tau - \dot{l}_{\theta_0,\eta_0}||_2^2$. The variance of $\dot{l}_{\theta_0,\eta_0}$ is defined as the efficient information matrix $\tilde{I}_{\theta_0,\eta_0}$, whose inverse attains the Cramér-Rao lower bound for estimating θ under a semiparametric framework (Bickel et al., 1998).

The main idea of semiparametric inference is to reduce the infinite dimensional estimation problem to a finite dimensional submodel called the least favorable submodel $\{P_{\theta,\eta^*(\theta)} : \theta \in \mathbb{R}^k\}$, where $\eta^*(\theta)$ is the so-called least favorable curve. The information matrix of the least favorable submodel attains the Cramér-Rao lower bound $\tilde{I}^{-1}_{\theta_0,\eta_0}$ and the least favorable curve $\eta^*(\theta)$ could be evaluated as the unique minimizer in $\mathcal{H}$ of the Kullback-Leibler (KL) divergence with the parametric part θ being fixed (Severini and Wong, 1992), i.e.

$$\eta^*(\theta) = \arg\inf_{\eta \in \mathcal{H}} K(P_{\theta_0,\eta_0}, P_{\theta,\eta}) = \arg\inf_{\eta \in \mathcal{H}} \left(- P_{\theta_0,\eta_0} \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta_0}} \right), \tag{7.3}$$

where $K(P,Q) = \int \log(dP/dQ) dP$ is the KL divergence between two measures P and Q. The existence of the least favorable submodel is implied by the closedness of the tangent set.

An intuitive explanation of the least favorable curve in Bayesian regime is that conditioning on θ, the posterior distribution of the nuisance parameter η tends to allocate all its mass around η^*_θ (Kleijn and van der Vaart, 2006). However, the posterior distribution of θ tends to concentrate around θ_0 (by Lemma 51 or Lemma 52 below). As a result, we only need to characterize the least favorable curve in a

small neighborhood of θ_0. In the sequel, we denote $\Delta\eta(\theta) = \eta^*(\theta) - \eta^*(\theta_0)$.

7.2.2 Model assumptions

Let $X_i = (U_i, V_i, Y_i)$, $i = 1, \ldots, n$, be i.i.d. copies of $T = (U, V, Y)$, where $Y \in \mathbb{R}$ is the response variable and $T = (U, V) \in [0,1]^k \times [0,1]^d$ is the covariance variable. Let $X^{(n)} = \{X_1, \ldots, X_n\}$. In the rest of the chapter, we stick to the notation P_0 to indicate the true underlying distribution that generates the data. Assume the following partially linear structure for a class of semi-parametric models:

$$m_0(t) \equiv E_0(Y|T = t) = F(g_0(t)), \quad g_0(t) = \theta_0^T u + \eta_0(v), \quad t = (u, v),$$

where $F : \mathbb{R} \to \mathbb{R}$ is some known link function, $\theta_0 \in \mathbb{R}^k$ is some unknown parameter of interest and η_0 is some unknown smooth function. Many statistical models can be included into this general framework. One example is the generalized partially linear models (Boente et al., 2006), where $y|t \sim p(y; m_0(t))$ for a conditional distribution p in the exponential family, such as the Guassian distribution for regression and the binomial distribution for classification. The generalized partially linear model with a Gaussian response is theoretically easiest to analysis and we will focus on it as one application of our general theory. Another example is the general partially linear model (GPLM) (Mammen and van de Geer, 1997), where the only assumption is made on the relationship between the conditional mean $m_{\theta,\eta} = F(g_{\theta,\eta})$ and the conditional variance $Var(Y|T) = V(m_{\theta,\eta}(T))$ for some known positive function V. For GPLM, the parameters (θ, η) can be estimated based on the quasi-likelihood function $Q_{\theta,\eta}(y) = \exp\{q_{\theta,\eta}(y)\}$, with $q_{\theta,\eta}(y)$ the quasi-log-likelihood function (Wedderburn, 1974)

$$q_{\theta,\eta}(y) = \int_y^{m_{\theta,\eta}(t)} \frac{(y - s)}{V(s)} ds. \tag{7.4}$$

If V is chosen as the conditional variance of the response Y and is assumed to depend only on the conditional mean m of Y, i.e. $V = V(m)$, then the quasi-

likelihood coincides with the likelihood of the corresponding generalized linear models (Wedderburn, 1974).

Despite the distinct modeling assumptions, these two classes of statistical models in the examples share many similarities and from now on, we work with a general "log-likelihood" function $l_n(\theta,\eta)=\sum_{i=1}^n l_{\theta,\eta}(X_i)$, where the general criterion function $l_{\theta,\eta}(x)$ can represent either $\log p(y;m_0(\theta,\eta))$ or $q_{\theta,\eta}(y)$. For GPLM, let

$$f(\xi)=\frac{dF(\xi)}{d\xi},\quad l(\xi)=\frac{f(\xi)}{V(F(\xi))},\quad \xi\in\mathbb{R},$$

$f_0=f(g_0)$ and $l_0=l(g_0)$. Similar to Mammen and van de Geer (1997), we make the following assumptions for GPLM:

Assumption 1. *[(a)]*

1. *There exists some positive constant C_0 such that $E_0(\exp(t|W|/C_0)|T)\leqslant C_0e^{C_0t^2}$, for all $t>0$, i.e. $W=Y-m_0(T)$ is sub-Gaussian.*

2. *There exist positive constants C_1, C_2, C_3 and C_4 such that: 1. $1/C_1\leqslant V(s)\leqslant C_1$ for all $s\in F(\mathbb{R})$; 2. $1/C_2\leqslant|l(\xi)|\leqslant C_2$ for all $\xi\in\mathbb{R}$; 3. $|l(\xi)-l(\xi_0)|\leqslant C_3|\xi-\xi_0|$ for all $|\xi-\xi_0|\leqslant\eta_0$; 4. $|f(\xi)-f(\xi_0)|\leqslant C_4|\xi-\xi_0|$ for all $|\xi-\xi_0|\leqslant\eta_0$.*

The assumption that V and l are both bounded could be restrictive and can be removed in many cases, such as the binary logistic regression model, by applying empirical process arguments similar to those in Section 7 of Mammen and van de Geer (1997). Under Assumption 1(2), the following lemma describes the local least favorable curve of the GPLM when $|\theta-\theta_0|$ is small.

Lemma 47. *Suppose Assumption 1(2) is met. Then the least favorable curve $\eta^*(\theta)$, defined as the minimizer η of*

$$E_0\log(Q_{\theta_0,\eta_0}/Q_{\theta,\eta})=E_0\int_{m_{\theta_0,\eta_0}(T)}^{m_{\theta,\eta}(T)}\frac{(Y-s)}{V(s)}ds=E_0\int_{m_{\theta_0,\eta_0}(T)}^{m_{\theta,\eta}(T)}\frac{(m_{\theta_0,\eta_0}(T)-s)}{V(s)}ds$$

as a function of θ, takes the following expression

$$\eta^*(\theta) = \eta_0 + (\theta - \theta_0)h^*(V) + O(|\theta - \theta_0|^2), \text{ as } |\theta - \theta_0| \to 0, \tag{7.5}$$

$$\text{with} \quad h^*(v) = -\frac{E_0\big[U f_0(T) l_0(T) | V = v\big]}{E_0\big[f_0(T) l_0(T) | V = v\big]}. \tag{7.6}$$

$h^*(v)$ is called the least favorable direction as it reflects the change of the least favorable curve due to a unit change in θ. The following two commonly used models are special cases of the GPLM.

Example 7.2.1 (Partially linear models)**.** *In the partially linear model (PLM), we have observations $\{X_i = (U_i, V_i, Y_i) : U_i \in \mathbb{R}^k, V_i \in \mathbb{R}^d, Y_i \in \mathbb{R}, i = 1, \ldots, n\}$ where the conditional distribution of Y given (U, V) are described by*

$$Y = U^T\theta_0 + \eta_0(V) + \epsilon, \tag{7.7}$$

where $\epsilon \sim N(0, 1)$ is assumed to be independent of (U, V). For simplicity, we focus on univariate θ, i.e., $k = 1$, and assume that (U, V) has an unknown distribution P supported on $[0, 1]^{1+d}$. $l_{\theta,\eta}(x)$ is given by

$$\log dP_{\theta,\eta}(X) = -\frac{1}{2}\big[\epsilon - (\theta - \theta_0)U - (\eta - \eta_0)(V)\big]^2, \tag{7.8}$$

where $\epsilon = Y - \theta_0 U - \eta_0(V)$ is the random error under (θ_0, η_0). For identifiability, we further assume $P(U - E[U|V])^2 > 0$. We consider the case that η_0 belongs to a Hölder function class $C^\alpha([0, 1]^d)$ with an unknown smoothness index α. For the PLM (7.7), the KL divergence between P_0 and $P_{\theta,\eta}$ is $P_0 \log(p_{\theta_0,\eta_0}/p_{\theta,\eta}) = \frac{1}{2}P\big((\theta - \theta_0)U + (\eta - \eta_0)(V)\big)^2$ and the least favorable curve is given by

$$\eta^*(\theta)(v) = \eta_0(v) - (\theta - \theta_0)E[U|V = v], \tag{7.9}$$

which satisfies assumption S0 with the least favorable direction $h^*(v) = -E[U|V = v]$ and $\Delta\eta(\theta)(v) = -(\theta - \theta_0)E[U|V = v]$. This is also a special case of Lemma 47 when $V(s) = 1$ and $F(x) = x$.

Example 7.2.2 (Partially linear logistic models)**.** *In the partially linear logistic model, $Y_i \in 0, 1$ and the conditional probability of Y given (U, V) can be described by*

$$\log \frac{P_0(Y = 1|U, V)}{P_0(Y = 0|U, V)} = U^T\theta_0 + \eta_0(V). \tag{7.10}$$

For this model, $f(\xi) = V(F(\xi)) = e^\xi(1 + e^\xi)^{-2}$ and $l(\xi) = 1$. Therefore, by Lemma 47 its least favorable curve is given by

$$\eta^*(\theta)(v) = \eta_0(v) - (\theta - \theta_0)\frac{E[Uf_0(U, V)|V = v]}{E[f_0(U, V)|V = v]} + O(|\theta - \theta_0|^2), \tag{7.11}$$

where $f_0(u, v) = e^{u^T\theta_0+\eta_0(v)}(1 + e^{u^T\theta_0+\eta_0(v)})^{-2}$.

Example 7.2.3 (Partially linear exponential models)**.** *In the partially linear exponential model, the conditional density of Y given (U, V) is*

$$p_0(y|u, v) = \lambda_0(u, v) \exp(-\lambda_0(u, v)y), \quad y > 0, \tag{7.12}$$

with $\lambda_0(u, v) = 1/m_0(u, v)$. For this model, $f(\xi) = e^\xi$, $V(F(\xi)) = e^{-2\xi}$ and $l(\xi) = e^{-\xi}$. Therefore, by Lemma 47 its least favorable curve is given by

$$\eta^*(\theta)(v) = \eta_0(v) - (\theta - \theta_0)E[U|V = v] + O(|\theta - \theta_0|^2), \tag{7.13}$$

where the least favorable direction $h^*(v) = -E[U|V = v]$ is the same as that of the PLM because $f_0l_0 \equiv 1$.

7.3 Second order semiparametric BvM theorem

7.3.1 Main results

For a general class of semiparametric models $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \mathbb{R}^k, \eta \in \mathcal{H}\}$, we consider a prior distribution Π over $\mathbb{R}^k \times \mathcal{H}$ for (θ, η). In the sequel, we use $\Pi^\theta(\eta)$ and $\Pi(\theta)$ to denote the conditional prior distribution of η given θ and the marginal prior distribution of θ, respectively. Denote $l_n(\theta, \eta)$ as the log-likelihood.

We assume the following assumptions.

Assumption 2 (Localization condition)**.** *There exist two sequences* $\{\delta_n\}$ *and* $\{\rho_n\}$ *satisfying* $\delta_n \to 0$, $\rho_n \to 0$ *and* $n\rho_n^2 \to \infty$, *and a sequence of sets* $\{\mathcal{H}_n\}$, *such that for some* $M > 0$ *as* $n \to \infty$,

$$\Pi\big(||\theta - \theta_0|| \leqslant M\rho_n, \eta \in \mathcal{H}_n \big| X_1, \ldots, X_n\big) = 1 - O_{P_0}(\delta_n).$$

Define the *localized* integrated likelihood ratio with respect to $\{\mathcal{H}_n\}$ as $\widetilde{S}_n : \mathbb{R}^k \to \mathbb{R}$:

$$\widetilde{S}_n(\theta) = \int_{\mathcal{H}_n} \exp\big(l_n(\theta, \eta) - l_n(\theta_0, \eta_0)\big) d\Pi^\theta(\eta). \tag{7.14}$$

Assumption 3 (Second order integrated local asymptotic normality)**.** *There exists a nondecreasing function* $R_n(\cdot) : \mathbb{R} \to \mathbb{R}$ *such that for every sequence* θ_n *such that* $\theta_n = \theta_0 + O_{P_0}(\rho_n)$,

$$\begin{aligned} \log \frac{\widetilde{S}_n(\theta_n)}{\widetilde{S}_n(\theta_0)} - \sqrt{n}(\theta_n - \theta_0)^T \widetilde{g}_n + \frac{n}{2}(\theta_n - \theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta_n - \theta_0) \\ = O_{P_0}(R_n\big(|\theta_n - \theta_0| \vee n^{-1/2}\log n\big)), \end{aligned} \tag{7.15}$$

where $\widetilde{g}_n = (1/\sqrt{n})\sum_{i=1}^n \widetilde{\ell}_0(X_i) \overset{P_0}{\rightsquigarrow} N_k(0, \widetilde{I}_{\theta_0,\eta_0})$.

Theorem 48. *We assume the prior for* θ *is thick at* θ_0. *Suppose* $X_1, \ldots, X_n$ *are i.i.d. observations sampled from* P_0. *Suppose that Assumption 2 & 3 are true. Then the marginal posterior for* θ *has the following expansion,*

$$\sup_A \big|\Pi(\theta \in A | X_1, \ldots, X_n) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (n\widetilde{I}_{\theta_0,\eta_0})^{-1}\big)(A)\big| = O_{P_0}(S_n), \tag{7.16}$$

where $S_n = R_n(n^{-1/2}\log n) + \delta_n$.

For regular parametric models, Johnson (1970) derived the above convergence rate as $O_{P_0}(n^{-1/2})$.

Assumption 2 is the test condition for semiparametric models, which allows us to focus on the posterior probability conditioning on the set $\{(\theta, \eta) : ||\theta - \theta_0|| \leqslant$

$M\rho_n, \eta \in \mathcal{H}_n\}$. A typical set of sufficient conditions (e.g. Lemma 51) for Assumption 2 already implies a test condition (Ghosal et al., 2000; Ghosal and van der Vaart, 2007). Therefore, Assumption 2 is stronger than the test condition for the semiparametric BvM. Further discussions on Assumption 2 are provided in Section 7.3.3.

Assumption 3 is a semiparametric extension of local asymptotic normality required for parametric BvM theorem (LeCam, 1953, LAN). Note that, by Fubini's theorem, the marginal posterior for θ can be written as

$$\begin{aligned}\Pi(\theta \in A | X_1, \ldots, X_n) = \int_A \left\{ \int_{\mathcal{H}} \exp\big(l_n(\theta, \eta) - l_n(\theta_0, \eta_0)\big) d\Pi^\theta(\eta) \right\} d\Pi(\theta) \\ \Big/ \int_\Theta \left\{ \int_{\mathcal{H}} \exp\big(l_n(\theta, \eta) - l_n(\theta_0, \eta_0)\big) d\Pi^\theta(\eta) \right\} d\Pi(\theta).\end{aligned} \tag{7.17}$$

Therefore, the following integrated likelihood ratio $S_n(\theta)$, i.e.,

$$S_n : \mathbb{R}^k \to \mathbb{R} : \theta \mapsto \int_{\mathcal{H}} \exp\big(l_n(\theta, \eta) - l_n(\theta_0, \eta_0)\big) d\Pi^\theta(\eta). \tag{7.18}$$

in a semiparametric model plays the same role as the likelihood ratio in a parametric model. To prove the first order semiparametric BvM theorems, Bickel and Kleijn (2012) assume that

$$\log \frac{S_n(\theta_0 + n^{-1/2} h_n)}{S_n(\theta_0)} = h_n^T \widetilde{g}_n - \frac{1}{2} h_n^T \widetilde{I}_{\theta_0, \eta_0} h_n + o_{P_0}(1), \tag{7.19}$$

for every random sequence $\{h_n\}$ of order $O_{P_0}(1)$.

However, accompanied with ILAN, Bickel and Kleijn (2012) requires a condition that the marginal posterior distribution of θ converges to θ_0 at rate $n^{-1/2}$. In many cases, the verification of this parametric rate condition is nontrivial. To avoid the stringent assumption on the convergence rate of θ as well as keep track of the higher-order remainders, we introduce the notion of localized integral likelihood ratio as in (7.14), where $\{\mathcal{H}_n\} \subset \mathcal{H}$ is the sequence of subsets of $\mathcal{H}$ defined

in Assumption 2. Note that each $\mathcal{H}_n$ forms a local neighborhood of η_0 such that $\Pi(\eta \in \mathcal{H}_n | X_1, \ldots, X_n) \xrightarrow{P_0} 1$. For example, $\mathcal{H}_n$ can be defined as $\{\eta : ||\eta - \eta_0||_n \leqslant M\rho_n\} \cap \mathcal{F}_n^\eta$, where $\mathcal{F}_n^\eta$ is a sieve sequence for the nuisance parameter defined after Lemma 51 and $||f||_n = n^{-1}\sum_{i=1}^n f^2(X_i)$ for a function f. ρ_n usually corresponds to the marginal posterior convergence rate of the nuisance parameter. By introducing the localization sequence $\{\mathcal{H}_n\}$, a uniform bound for the corresponding higher order term in the local asymptotic expansion as (7.19) can be developed with respect to the local neighborhood $\mathcal{H}_n$ instead of the whole space $\mathcal{H}$. Therefore, the additional information $||\eta - \eta_0||_n \leqslant M\rho_n$ and $\eta \in \mathcal{F}_n^\eta$ can be utilized when applying the maximal inequalities (van der Vaart and Wellner, 1996, Corollary 2.2.5). Moreover, we no longer need to assume a root-n marginal convergence rate for θ since we only need to focus on the posterior distribution over $\{\theta : ||\theta - \theta_0|| \leqslant M\rho_n\}$ and the posterior probability of $\{\theta : Mn^{-1/2}\log n \leqslant ||\theta - \theta_0|| \leqslant M\rho_n\}$ decays faster than S_n for sufficiently large M. (7.15) can be translated into (7.19) by letting the localization parameter $h_n = \sqrt{n}(\theta_n - \theta_0)$. The uniform remainder-bound in (7.15) is weaker than $O_{P_0}[R_n(\|\theta_n - \theta_0\|)]$ because we only require the remainder to be of order $R_n(n^{-1/2}\log n)$ when θ_n is in a $n^{-1/2}\log n$ neighborhood of θ_0. In the sequel, any mentioning of ILAN refers to (7.15).

The ILAN condition imposes constraints on both the prior through the definition of the localized integrated likelihood ratio $\tilde{S}_n$, and the semiparametric model $\{P_{\theta,\eta}\}$ through the second order LAN expansion in (7.15). Interestingly, Lemma 53 in Section 7.3.4 suggests that this convoluted condition can separated into the following condition (A1) on the semiparametric model and condition (A2) on the prior — (A1) and (A2) implies (7.15) with $R_n = G_n + G_n'$.

(A1) (Stochastically local asymptotic normality) There exists an increasing func-

tion $G_n : \mathbb{R} \to [0, \infty)$, such that for every sequence $\{\theta_n\}$ such that $\theta_n = \theta_0 + O_{P_0}(\rho_n)$,

$$\sup_{\eta \in \mathcal{H}_n} \Big| l_n\big(\theta_n, \eta + \Delta\eta(\theta_n)\big) - l_n(\theta_0, \eta) - (\theta_n - \theta_0)^T \sum_{i=1}^{n} \tilde{l}_{\theta_0,\eta_0}(X_i) + \frac{1}{2} n(\theta_n - \theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta_n - \theta_0) \Big| = O_P\big[G_n\big(\max\{|\theta_n - \theta_0|, n^{-1/2}\log n\}\big)\big].$$

In the sequel, we will say that an identity holds uniformly over $\eta \in \mathcal{H}_n$ instead of taking a sup over $\mathcal{H}_n$ to ease the exhibition. We call a prior as being thick at θ_0 if it has a Lebesgue density that is continuous and strictly positive at θ_0.

(A2) (Prior stability under perturbation) There exists an increasing function $G'_n : \mathbb{R} \to [0, \infty)$, such that for any $\theta_n = \theta_0 + O_P(\rho_n)$,

$$\frac{\int_{\mathcal{H}_n} \exp(l_n(\theta_0, \eta - \Delta\eta(\theta_n)))d\Pi^{\theta_n}(\eta)}{\int_{\mathcal{H}_n} \exp(l_n(\theta_0, \eta))d\Pi^{\theta_0}(\eta)} = 1 + O_P[G'_n(\max\{|\theta_n - \theta_0|, n^{-1/2}\log n\})].$$

As we will see, (A2) plays an important role for a prior to be semiparametric objective. The definition of semiparametric objective priors is provided in Section 7.4. More discussions on (A1) and (A2) can be found in Section 7.3.4 and Section 7.4.

7.3.2 Second order Bayesian inference

In practice, an MCMC algorithm is designed to sample a sequence of draws $\{\theta^{(l)} : l = 1, \ldots, L\}$ approximately from the marginal posterior distribution of $\theta = (\theta_1, \ldots, \theta_k)$. Then for each component θ_s with $1 \leqslant s \leqslant k$, an estimation $\widehat{\theta}^B_{n,s}$, such as the posterior median, and its corresponding α-th credible interval $(\widehat{q}_{s,\alpha/2}, \widehat{q}_{s,1-\alpha/2})$ are obtained from the samples $\{\theta^{(l)}_s\}$, which are approximately drawn from the marginal posterior distribution of θ_s. Then $\widehat{\theta}^B_n = (\widehat{\theta}^B_{n,1}, \ldots, \widehat{\theta}^B_{n,k})$ forms an point estimator of θ_0. Another way to quantify the estimation uncertainty is to construct the αth highest posterior density (HPD) region $A_{n,\alpha}$ based on $\{\theta^{(l)}\}$, which forms a αth joint credible regions for the k-dimensional vector θ.

The following result shows the frequentist validity of this procedure for the point estimator.

Corollary 49. *Consider the semiparametric model and the prior* Π *in Theorem 48. Under the same assumptions, the estimator* $\widehat{\theta}_n^B$ *of* θ_0 *constructed as above satisfies*

$$\sqrt{n}(\widehat{\theta}_n^B - \theta_0) = \widetilde{\Delta}_n + O_{P_0}(S_n),$$

where $\widetilde{\Delta}_n \overset{P_{\theta_0,\eta_0}}{\rightsquigarrow} N(0, \widetilde{I}^{-1}_{\theta_0,\eta_0})$*, as* $n \to \infty$*.*

The second order term $O_{P_0}(S_n)$ could be very close to the first order term even when the sample size n is moderate. For instance, if $S_n \sim \sqrt{n}\rho_n^2$ (which are the case in the examples) and the nuisance part converges at a cubic rate as $\rho_n \sim n^{-1/3}$ up to log terms, then $S_n \sim n^{-1/6}$ and $\sqrt{n}(\widehat{\theta}_n^B - \theta_0) = O_{P_0}(1) + O_{P_0}(n^{-1/6})$. Therefore, it is important to quantify the impact of the higher order term on BvM results.

Next we study the frequentist coverage of the individual/joint credible interval/region by the above procedures. For any $\alpha \in (0,1)$, we define the α-th marginal posterior quantile $\widehat{q}_{s,\alpha}$ of θ_s through the following equation $\Pi(\theta_s \leqslant \widehat{q}_{s,\alpha}|X_1,\ldots,X_n) = \alpha$. Consider any $1-\alpha$ credible region $A_{n,1-\alpha}$ that satisfies $\Pi(\theta \in A_{n,1-\alpha}|X_1,\ldots,X_n) = 1-\alpha$.

Corollary 50. *Consider the semiparametric model and the prior* Π *in Theorem 48. Under the same assumptions, we have*

$$P_0(\theta_{0,s} \in (\widehat{q}_{s,\alpha/2}, \widehat{q}_{s,1-\alpha/2})) = 1 - \alpha + O(S_n), \tag{7.20}$$

$$P_0(\theta_0 \in A_{n,1-\alpha}) = 1 - \alpha + O(S_n). \tag{7.21}$$

From this corollary, we see that the second order term also determines the frequentist coverage of the credible intervals/resions. Therefore, the second order properties are also important for statistical inferences and developing semiparametric objective priors that lead to the best second order term is necessary.

7.3.3 Higher-order results on posterior convergence of nuisance parameter

In this section, we provide a set of sufficient conditions for Assumption 2. We consider a general framework for investigating the posterior contraction rate based on independent but unnecessarily identically distributed observations $Y^{(n)} = (Y_1, \ldots, Y_n)$; see Ghosal and van der Vaart (2007). In this case, the statistical model in consideration is written as $\mathcal{P} = \{P_\lambda^{(n)} : \lambda \in \Lambda\}$, where $P_\lambda^{(n)}(Y^{(n)}) = \prod_{i=1}^n P_{\lambda,i}(Y_i)$ is the joint distribution of $Y^{(n)}$ with a common parameter λ.

Define a semimetric d_n by $d_n^2(\lambda, \lambda') = \frac{1}{n}\sum_{i=1}^n \int(\sqrt{p_{\lambda,i}} - \sqrt{p_{\lambda',i}})^2 d\mu_i$, which averages the squared Hellinger distances for distributions of Y_i. In the above statistical model $\mathcal{P}$, we say that the posterior convergence rate of λ is ρ_n if

$$\Pi\big(d_n(\lambda, \lambda_0) \geqslant M\rho_n \big| X_1, \ldots, X_n\big) \xrightarrow{P_0} 0, \tag{7.22}$$

where M is a sufficiently large positive constant; see Ghosal et al. (2000) and Ghosal and van der Vaart (2007). However, our second order studies of BvM theorem requires an explicit bound, called as decaying rate, characterizing the convergence rate of (7.22), e.g., Assumption 2.

In the below, we provide a Lemma for deriving a polynomial decaying rate in general cases, and further improve it to a exponential rate in the case of GPLM. The first Lemma is an immediate consequence from combining Lemma 10 in Ghosal and van der Vaart (2007) with the proof of Theorem 2.1 in Ghosal et al. (2000). Hence, we skip its proof.

For any integer k, define the discrepancy measure $V_{0,k}(P, Q) = \int |\log(dP/dQ) - K(P, Q)|^k dP$.

Lemma 51. *Let ρ_n be a sequence satisfying $\rho_n \to 0$ and $n\rho_n^2 \to \infty$. If there exists an increasing sequence of sieves $\mathcal{F}_n \subset \mathcal{F}$, such that the following conditions are satisfied:*

a. $\Pi(\mathcal{F} \backslash \mathcal{F}_n) \leqslant \exp(-n\rho_n^2(C+4))$ *for some $C > 0$;*

b. $\log N(\rho_n, \mathcal{F}_n, d_n) \leqslant n\rho_n^2$;

c. $\Pi(B_n(P_0^{(n)}, \rho_n; k)) \geqslant \exp(-Cn\rho_n^2)$,

where $B_n(P_0^{(n)}, \rho_n; k) = \left\{P \in \mathcal{P} : K(P_0^{(n)}, P^{(n)}) \leqslant n\rho_n^2, V_{0,k}(P_0^{(n)}, P^{(n)}) \leqslant n^{k/2}\rho_n^k\right\}$,

then we have

$$\Pi\big(d_n(\lambda, \lambda_0) \geqslant M\rho_n \big| X_1, \ldots, X_n\big) = O_{P_0}\big((n\rho_n^2)^{-k/2}\big). \tag{7.23}$$

By taking $k = 2$ in Lemma 51, we recover Theorem 2.1 in Ghosal et al. (2000) for iid observations without explicit characterizations on the decaying rate. Moreover, if $n^{-k/2}\rho_n^{-k} = O(n^{-\gamma})$ for some $\gamma > 1$, then by the Borel-Cantelli lemma and Lemma 51, $\Pi\big(d(P, P_0) \geqslant M\rho_n \big| X_1, \ldots, X_n\big) \to 0$ almost surely.

A typical sieve construction for semiparametric models involves sieve sequences $\{\mathcal{F}_n^\theta\}$ and $\{\mathcal{F}_n^\eta\}$ for the parametric part and nuisance part, respectively. For example, for the GPLM, the sieve takes a product form as $\mathcal{F}_n = \mathcal{F}_n^\theta \oplus \mathcal{F}_n^\eta = \{\theta^T u + \eta(v) : \theta \in \mathcal{F}_n^\theta,\ \eta \in \mathcal{F}_n^\eta\}$.

Lemma 51 provides up to polynomial decaying rates for general cases. However, for the GPLM, an exponential decaying rate can be attained by the following lemma.

Lemma 52. *Consider the GPLM under Assumption 1. If conditions (a) and (b) in Lemma 51 and the following condition are true,*

d. $\Pi(||g - g_0||_n \leqslant \rho_n) \geqslant \exp(-Cn\rho_n^2)$,

where $g(t) = \theta^T u + \eta(v)$, *then there exists a* $C_0 > 0$, *so that*

$$\Pi\big(||g - g_0||_n \geqslant M\rho_n \big| X_1, \ldots, X_n\big) = O_{P_0}\big(\exp(-C_0 n\rho_n^2)\big). \tag{7.24}$$

7.3.4 Sufficient conditions for ILAN

In this section, we discuss sufficient conditions (A1) & (A2) for Assumption 3. Condition (A1) strengthens the stochastically LAN introduced in Bickel and Kleijn (2012).

If we set $\eta = \eta_0$ in (A1), then we obtain the LAN for the least favorable submodel

$$l_n\big(\theta_n, \eta^*(\theta_n)\big) - l_n(\theta_0, \eta_0) = \sqrt{n}(\theta_n - \theta_0)^T \widetilde{g}_n - \frac{1}{2} n (\theta_n - \theta_0)^T \widetilde{I}_{\theta_0,\eta_0} (\theta_n - \theta_0) + O_P[G_n(\max\{|\theta - \theta_0|, n^{-1/2} \log n\})].$$

This explains the reason for the inclusion of the $\Delta\eta(\theta_n)$ term in (A1). Note that (A1) depends on the prior through the localization sequence $\{\mathcal{H}_n\}$, to which the posterior distribution allocates most mass. Larger the subset $\mathcal{H}_n$, greater the high order leftover in (A1). So we aim to make the $\mathcal{H}_n$ as small as possible while keeping $\Pi(\mathcal{H}_n | X_1, \ldots, X_n)$ close to one. Motivated by this, we set

$$\mathcal{H}_n = \{\eta : ||\eta - \eta_0||_n \leqslant M\rho_n\} \cap \mathcal{F}_n^{\eta}, \tag{7.25}$$

where $\{\mathcal{F}_n^{\eta}\}$ is the sieve sequence for the nuisance part η constructed in Lemma 51 and ρ_n is the corresponding posterior convergence rate of η. By Assumption 2 and condition (a) in Lemma 51, we have

$$\Pi(\mathcal{H}_n | X_1, \ldots, X_n) = 1 - O_{P_0}(\max\{\delta_n, \exp(-n\rho_n^2)\}),$$

where $\delta_n = (n\rho_n^2)^{-k/2}$ or $\exp(-n\rho_n^2)$ depends on whether Lemma 51 or Lemma 52 is satisfied. The remainder term in (A1) can be bounded from above by calculating the continuity modulus or applying the maximal inequalities from empirical process theory (van der Vaart and Wellner, 1996). The partially linear model and GPLM with quasi-likelihood examples later illustrate how to apply these tools to verify (A1).

Based on the above preparations, we can prove the following lemma which provides a sufficient condition for the ILAN.

Lemma 53. *If (A1) and (A2) hold, then we have the following ILAN,*

$$\log \frac{\widetilde{S}_n(\theta_n)}{\widetilde{S}_n(\theta_0)} = \sqrt{n}(\theta_n - \theta_0)^T \widetilde{g}_n - \frac{n}{2}(\theta_n - \theta_0)^T \widetilde{I}_{\theta_0,\eta_0} (\theta_n - \theta_0) + O_P[R_n(\max\{|\theta_n - \theta_0|, n^{-1/2} \log n\})],$$

with $R_n = G_n + G_n'$.

(A2) characterizes the stability of the prior under a small perturbation in the log likelihood function caused by the semiparametric bias $\Delta\eta(\theta_n)$ in the nuisance part. In the special case that the LFS is given by $\{P_{\theta,\eta_0} : \theta \in \mathbb{R}^k\}$, i.e., $\Delta\eta \equiv 0$, (A2) automatically holds when independent priors are specified for θ and η.

However, in general cases where $\Delta\eta \neq 0$, since S0 implies $\Delta\eta(\theta_n) = O(|\theta_n - \theta_0|)$, we have

$$\exp\left\{l_n(\theta_0, \eta - \Delta\eta(\theta_n)) - l_n(\theta_0, \eta)\right\} = O_{P_0}(n|\theta_n - \theta_0|\rho_n),$$

which does not converge to zero as we expect that $|\theta - \theta_0| = O_P(n^{-1/2})$. Hence under independent priors for θ and η, i.e. $\Pi^\theta \equiv \Pi$ for any θ, (A2) cannot be simply proved by bounding the difference between the logarithms of the integrands in the denominator and the numerator.

7.4 Semiparametric objective priors

According to Cheng and Kosorok (2008a) and Cheng and Kosorok (2009), the maximum (penalized) profile likelihood estimator $\widehat{\theta}_n$ for a semiparametric model satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \widetilde{\Delta}_n + O_{P_0}(M_n(\rho_n)), \tag{7.26}$$

where $M_n(t) = \sqrt{n}t^2$ and ρ_n corresponds to the convergence rate of the nuisance parameter. For example, $\rho_n = n^{-\alpha/(2\alpha+d)}$ if η_0 is a d-variate function with known smoothness level α. Similar to the objective prior for regular parametric models defined via probability matching (Datta and Mukerjee, 2004; Staicu and Reid, 2008), we call a prior for semiparametric models to be *semiparametric objective* if:

1. The marginal posterior median $\widehat{\theta}_n^B$ of θ satisfies $\sqrt{n}(\widehat{\theta}_n^B - \theta_0) = \widetilde{\Delta}_n + O_{P_0}(\widetilde{M}_n(\rho_n))$. where $\widetilde{M}_n$ is the same as M_n up to $\log n$ factors.

2. For any $\alpha \in (0,1)$, the αth marginal posterior quantile $\widehat{q}_{s,\alpha}$ of the sth component θ_s satisfies $P_0(\theta_{0,s} \leqslant \widehat{q}_{s,\alpha}) = \alpha + O(\widetilde{M}_n(\rho_n))$.

Different from the parametric probability matching prior where the remainder order is always $n^{-1/2}$, the remainder order of a semiparametric model also depends on the prior. Therefore, we define semiparametric objective priors in terms of higher-order matching. From the results in Section 7.3.2, a prior is semiparametric objective if the S_n term in Theorem 48 has the same order as $M_n(\rho_n)$ up to $\log n$ factors.

In this section, we first investigate the conditions under which an independent prior is semiparametric objective. These conditions become unrealistic for those independent priors whose marginal on the nonparametric part η is adaptive to the smoothness of the true function η_0, which is consistent with former negative observations (Castillo, 2012; Rivoirard and Rousseau, 2012). On the contrary, we show that a new class of dependent priors can be simultaneously semiparametric objective and adaptive under mild conditions. Throughout this section, we assume

(A3) (Local expansion for the least favorable curve) There exists a function $h^* \in L_2(P_0)$ such that as $\theta \to \theta_0$,

$$\eta^*_\theta(v) = \eta_0 + (\theta - \theta_0)h^*(v) + O(|\theta - \theta_0|^2).$$

A3 is a mild condition. For example, it is satisfied for the GPLM with quasi-likelihood (Lemma 47).

As we show in Lemma 53, the high-order remainder term S_n depends on G_n and G'_n in (A1) and (A2). (A1) appears to be intrinsic to the semiparametric model. For example, in many cases $G_n(t)$ takes a form of $nt^3 + \sqrt{n}t^2 + nt^2\rho_n + nt\rho_n^2 + \sqrt{n}\rho_n^2$. Therefore, by Theorem 48, under this situation a prior is semiparametric objective if it satisfies (A2) with $G'_n(t) = O(G_n(t))$ as $t \to 0$. As a weaker requirement compared to the semiparametric objectiveness, we call a prior to be unbiased if (A2) holds with $G'_n(n^{-1/2}\log n) \to 0$ as $n \to \infty$. Note that (A2) is also a basic requirement for the root-n convergence rate of θ.

7.4.1 Independent prior

We illustrate when an independent prior

$$\text{(PI)} \qquad \theta \sim \Pi_\Theta, \quad \eta \sim \Pi_H.$$

is semiparametric objective. Applying a change of variable $\eta \to \widetilde{\eta} + \Delta\eta(\theta_n)$ in the numerator in (A2) yields

$$\int_{\mathcal{H}_n} \exp(l_n(\theta_0, \eta - \Delta\eta(\theta_n)))d\Pi_H(\eta) = \int_{\mathcal{H}_n - \Delta\eta(\theta_n)} \exp(l_n(\theta_0, \widetilde{\eta}))f(\widetilde{\eta})d\Pi_H(\widetilde{\eta}), \quad (7.27)$$

with $f(\widetilde{\eta}) = d\Pi_{H,\cdot-\Delta\eta(\theta_n)}(\widetilde{\eta})/d\Pi_H(\widetilde{\eta})$ the Radon-Nykodym derivative between the two measures, where $\Pi_{H,\cdot-g}$ representing the distribution of $W - g$ if $W \sim \Pi_H$. Consider the following assumption:

(A4) There exists a nondecreasing function $G_n'' : \mathbb{R} \to \mathbb{R}$, such that for any $\theta_n = \theta_0 + O_{P_0}(\rho_n)$ and uniformly over $\eta \in \mathcal{H}_n$, $|\log f(\eta) - 1| = O[G_n''(\max\{|\theta_n - \theta_0|, n^{-1/2}\log n\})]$.

The following lemma provides sufficient conditions under which (PI) is unbiased.

Lemma 54. *Assume A3, Assumption 2 and A4. Then the independent prior satisfies (A2) with* $G_n' = G_n'' + \delta_n$.

$G_n'(t)$ in Lemma 54 does not converge to 0 as $t \to 0$. Therefore, we do not require the remainder term in (A1) to converge to $G_n(0) = 0$ as $\theta_n \to \theta$. Given a prior for η, A4 indeed puts a restriction on $\Delta\eta(\theta_n)$. For example, when Π corresponds to the Gaunsian process (GP) prior (Rasmussen and Williams, 2006), this restriction requires $\Delta\eta(\theta_n) \in \mathbb{H}$, where $\mathbb{H}$ is the reproducing kernel Hilbert space associated with the GP, and otherwise $|f(\eta)| = \infty$. When the smoothness α of η is known and the covariance function of the GP prior is chosen properly, ρ_n can attain $n^{-\alpha/(2\alpha+d)}(\log n)^\gamma$ for some $\gamma > 0$, which is close to the minimax rate $\Delta\eta(\theta_n)$ of estimating an d-variate α-smooth function, as long as h^* is at lease α-smooth. As a result, such an independent prior can be unbiased and semiparametric objective. On the other hand, when

a smoothness-adaptive GP prior with random length scale parameter (van der Vaart and van Zanten, 2009) is specified for η, the restriction $\Delta\eta(\theta_n) \in \mathbb{H}$ becomes very stringent under the same ρ_n. Section 7.5.1 provides a concrete example.

7.4.2 Dependent prior

Since our primary interests is a smoothness adaptive prior for η, the above analysis implies that we need to consider unbiased and semiparametric objective priors where θ and η are dependent.

Let $\widehat{h}$ be an estimator of the least favorable direction h^* that satisfies the following assumption: for any $\theta_n = \theta_0 + O_P(\rho_n)$,

$$\text{(A5)} \qquad l_n(\theta_0, \eta - \Delta\eta(\theta_n) + (\theta_n - \theta_0)\widehat{h}) - l_n(\theta_0, \eta) \\ = O_{P_0}[G_n''(\max\{|\theta - \theta_0|, n^{-1/2}\log n\})].$$

Because $\Delta\eta(\theta_n) - (\theta_n - \theta_0)\widehat{h} = (\theta_n - \theta_0)(h^* - \widehat{h}) + O(|\theta_n - \theta_0|^2)$, in many cases A5 can be implied by the following condition with $G_n''(t) = n\rho_n\kappa_n t + n\rho_n t^2$:

$$\text{(A6)} \qquad ||\widehat{h} - h^*||_n = O_P(\kappa_n),\ \kappa_n \to 0.$$

For example, assumption A6 will be made in the examples in Section 7.5.

Let Π_Θ be an appropriate marginal prior for θ and Π_H a smoothness adaptive prior for η. Consider the following prior for (θ, η),

$$\text{(PD)} \qquad \theta \sim \Pi_\Theta, \quad \eta|\theta \sim W + \theta\widehat{h} \text{ with } W \sim \Pi_H.$$

The conditional prior of η given θ in (PD) is obtained by shifting the center of Π_H by $\theta\widehat{h}$. The idea is simple: we want to compensate for the bias by adjusting the center of the prior for the nuisance part. With this bias correction, the stringent condition A4 on the least favorable direction can be avoided.

Lemma 55. *If A3, Assumption 2 and A5 are true, then the dependent prior (PD) satisfies (A2) with* $G_n' = G_n'' + \delta_n$.

Compared to A4, A6 is much weaker because it only requires a satisfactory estimator of h^*. On the contrary, A4 puts constraints on both the prior Π_H and h^*. The constraint on Π_H is too stringent to admit adaptivity. For example, see Section 7.5.1.

7.4.3 Second-order BvM theorems for unbiased priors

Applying Theorem 48 to the priors in the above subsections, we obtain the following result.

Theorem 56. *Suppose $X_1, \ldots, X_n$ are i.i.d. observations sampled from $P_0 = P_0$. Suppose that A3, S1, Assumption 2, (A1) holds. If either A4 for the independent prior (PI) or A5 for the dependent prior (PD) is true, then the marginal posterior for θ has the following expansion in total variation as $n \to \infty$,*

$$\sup_A \left|\Pi(\theta \in A|X_1, \ldots, X_n) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (n\widetilde{I}_{\theta_0,\eta_0})^{-1}\big)(A)\right| = O_P[G_n(n^{-1/2}\log n) + G_n''(n^{-1/2}\log n) + \delta_n].$$

7.5 Examples

Because the PLM is easy to analysis and sufficient for comprehension, we primarily focus on PLM as an application of the general theory. Then we provide general results for the GPLM, of which the PLM is a special case.

7.5.1 Partially linear models

Independent prior

First, we consider independent priors for θ and η. Theoretically, the prior for θ can be any continuous distribution on $\mathbb{R}^k$ that has a full support. However, for computational convenience such as conjugacy, we choose a multivariate normal distribution $N(0, I_k/\phi_0)$ for $\Pi(\theta)$, where ϕ_0 is the precision parameter of the prior. For example, one can choose $\phi_0 = 0.01$ for a vague prior for θ.

The prior Π_η is a infinite dimensional measure over the $d0$-variate Hölder function class with unknown smoothness. We consider the Gaussian process (GP) prior with a random inverse bandwidth parameter (van der Vaart and van Zanten, 2009). The inverse bandwidth parameter determines the decaying rate of the covariance function. A GP $W^a = \{W_x^a : x \in \mathbb{R}^d\}$ with a fixed inverse parameter a is denoted by $W^a \sim GP(m, K^a)$, where $m : \mathbb{R}^d \to \mathbb{R}$ is the mean function and $K^a : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the covariance function that depends on a. We primarily focus on the squared exponential covariance function, although with slight modifications, the following results could be generalized to a broader class of covariance functions with exponentially decaying spectral densities. Let $K^a(x, y) = EW_x^a W_y^a = \exp(-a^2||x - y||^2)$ be the squared exponential covariance function indexed by $a > 0$.

Given n independent observations, the minimax rate of estimating a d-variate function that is known to be Hölder α-smooth is $n^{-\alpha/(2\alpha+d)}$. van der Vaart and van Zanten (2009) shows that a hierarchical prior as

$$W^A|A \sim GP(0, K^A), \quad A^d \sim Ga(a_0, b_0), \tag{7.28}$$

for $Ga(a_0, b_0)$ the Gamma distribution with pdf $p(t) \propto t^{a_0-1}e^{-b_0 t}$ leads to the minimax rate $n^{-\alpha/(2\alpha+d)}$ up to a logarithmic factor, adaptively over the unknown smoothness level α. We slightly modify the prior for A to be a truncated $Ga(a_0, b_0)$ whose pdf $p(t) \propto t^{a_0-1}e^{-b_0 t}I(t \geqslant t_0)$ for technical reasons and this modification will not sacrifice the adaptivity of the prior. Choices of the hyperparameters have diminishing impacts on the posterior as the sample size n grows and therefore we simply set $a_0 = 1/2$ and $b_0 = 1/2$.

Properties of a GP with a covariance function K are intimately related to the reproducing kernel Hilbert space (RKHS) associated with K. For the scaling dependent covariance K^a, we use $\mathbb{H}^a$ and $||\cdot||_a$ to denote the RKHS and the RKHS norm, respectively. The unit ball in the RKHS $\mathbb{H}^a$ is denoted by $\mathbb{H}_1^a$.

By checking the assumptions A3, Assumption 2, (A1), (A2) and A4, we have the following theorem for the independent prior above for partially linear models.

Theorem 57. *Let $X_i = (U_i, V_i, Y_i) \in \mathbb{R}^k \times \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$, be n observations from the partially linear model (7.7). Consider the independent priors for θ and η as above. Assume the following conditions:*

1. *η_0 is Hölder α-smooth;*
2. *$P(U - E[U|V])^2 > 0$;*
3. *The conditional expectation $E[U|V = v]$ as a function of v belongs to the RHKS $\mathbb{H}^{t_0}$.*

Then the following second order asymptotic expansion holds:

$$\sup_A \left|\Pi(\theta \in A|X^{(n)}) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (nI_0^{-1})\big)(A)\right| = O_P(\sqrt{n}\rho_n^2 \log n), \qquad (7.29)$$

where $\widetilde{\Delta}_n = n^{-1/2}\sum_{i=1}^n I_0^{-1}\epsilon_i(U_i - E[U|V_i]) \overset{P_0}{\rightsquigarrow} N_k(0, I_0^{-1})$, $I_0 = P(U - E[U|V])^2$ and $\rho_n = n^{-\alpha/(2\alpha+d)}(\log n)^{1+d}$.

The above theorem suggests that by choosing an adaptive prior for the nuisance parameter, the second order estimating efficiency of the parametric part could also be improved adaptively. However, this theorem requires a strong constraint on the least favorable direction $h*(v)$ if an independent prior is specified. In fact, a necessary and sufficient condition for a function f belonging to $\mathbb{H}^{t_0}$ is that it has a Fourier transform $\widehat{f}$ satisfying

$$\int_{\mathbb{R}^d} |\widehat{f}(\lambda)|^2 e^{c||\lambda||^2/t_0^2} d\lambda \leqslant \infty,$$

for some $c > 0$ (van der Vaart and van Zanten, 2009). This condition implies the infinite differentiability of h^*. Since $\mathbb{H}^s \in \mathbb{H}^t$ for all $s < t$ (van der Vaart and van

Zanten, 2009, Lemma 4.7), Theorem 57 guarantees the second order BvM results under a weaker constraint on h^* with a larger t_0. To avoid such a strong condition, we consider the dependent prior as in the next subsection.

Dependent prior

We consider the dependent prior (PD) in section 7.4.2. For partially linear models, the least favorable direction $h^*(\cdot)$ takes the form of $-E[U|V = \cdot]$, which does not involves (θ_0, η_0) and thus can be estimated as $\widehat{h}$ by using the design points $\{(U_i, V_i)\}$ directly. For GP prior, shifting the center is equivalent to the translation of the mean function. Therefore, for partially linear models with the above GP prior, the dependent prior version (PD) is,

$$\begin{aligned} \theta \sim &\Pi_\Theta, \quad A^d \sim Ga(a_0, b_0), \\ &\eta|\theta, A \sim GP(\theta\widehat{h}, K^A). \end{aligned} \tag{7.30}$$

An intuitive explanation of the above prior is the following. If we reparameterize the nuisance parameter as $\xi = \eta - \theta\widehat{h}$, then $\xi|A, \theta \sim GP(0, K^A)$ and the partially linear model becomes

$$Y = \theta\big[U + \widehat{h}(V)\big] + \xi(V) + \epsilon, \tag{7.31}$$

with the truth $\theta = \theta_0$ and $\xi = \xi_0 \triangleq \eta_0 - \theta_0\widehat{h}$. If we consider $U + \widehat{h}(V)$ as a new $\widetilde{U}$, then the least favorable direction of the new model becomes

$$\widetilde{h} = E[\widetilde{U}|V] = E\{U - E[U|V]|V\} + h^* - \widehat{h} = O_{P_0}(\kappa_n) \xrightarrow{P_0} 0,$$

with κ_n being defined in assumption A6. As a result, the semiparametric bias of the new model is negligible.

The following theorem formalizes this observation and provides the seconder order BvM theorem under this dependent prior.

Theorem 58. *Let $X_i = (U_i, V_i, Y_i)$, $i = 1, \ldots, n$, be a sample from the partially linear model. Suppose that $\widehat{h}$ is an estimator of the least favorable direction $h^*(\cdot) = -E[U|V = \cdot]$. Assume the first two conditions in Theorem 57 and A6. Then under the prior* (7.30)*, the following second order asymptotic expansion holds:*

$$\sup_A \left|\Pi(\theta \in A|X^{(n)}) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (nI_0^{-1})\big)(A)\right| = O_P\big(\sqrt{n}\rho_n^2 \log n + \sqrt{n}\kappa_n\rho_n \log n\big), \tag{7.32}$$

where $\widetilde{\Delta}_n = n^{-1/2}\sum_{i=1}^n I_0^{-1} e_i(U_i - E[U|V_i]) \overset{P_0}{\rightsquigarrow} N_k(0, I_0^{-1})$, $I_0 = P(U - E[U|V])^2$ and $\rho_n = n^{-\alpha/(2\alpha+d)}(\log n)^{1+d}$.

By the above theorem, if $\kappa_n = O_P(\rho_n)$, then we can achieve the same adaptive second order efficiency as Theorem 57, but under a much weaker condition A6. For example, when h^* is at least α times differentiable, then the typical construction of $\widehat{h}$ as a kernel type estimator with appropriate choices of the kernel and the bandwidth parameter satisfies A6.

Simulation study

In this part, we conduct a simulation study comparing the dependent prior and the independent prior. In each setting, we generate 100 replicates from the following four models:

M1 $Y_i = 0.5U_i + \exp(V_i) + N(0, 0.5^2)$, with $V_i \overset{iid}{\sim} N(0,1)$ and $U_i|V_i \sim N(0.5|V_i|^3, 1)$;

M2 $Y_i = 0.5U_i + \exp(V_i) + N(0, 0.5^2)$, with $V_i \overset{iid}{\sim} N(0,1)$ and $U_i|V_i \sim N(0.5V_i^3, 1)$.

M3 $Y_i = 0.5U_i + \exp(|V_i|) + N(0, 0.5^2)$, with $V_i \overset{iid}{\sim} N(0,1)$ and $U_i|V_i \sim N(0.5|V_i|^3, 1)$;

M4 $Y_i = 0.5U_i + \exp(|V_i|) + N(0, 0.5^2)$, with $V_i \overset{iid}{\sim} N(0,1)$ and $U_i|V_i \sim N(0.5V_i^3, 1)$.

In M1, the least favorable direction $h^*(v) = 0.5|v|^3$ is twice differentiable in $\mathbb{R}$ but not thrice differentiable at $v = 0$. In contrast, the least favorable direction $h^*(v) = 0.5v^3$

in M2 is infinitely differentiable. M3 and M4 are counterparts of M1 and M2 respectively with non-differentiable nuisance parts at $v = 0$. We consider three procedures to fit these models: P1. the independent prior (7.28); P2. the dependent prior (7.30) with $\widehat{h}(v)$, which estimates $-E[U|V = v]$, produced by the Nadaraya-Watson kernel regression algorithm using the Gaussian kernel with an optimal bandwidth (Bowman and Azzalini, 1997, p.31); P3. the dependent prior (7.30) with $\widehat{h}(v)$ the same as the truth $h(v)$. In all of them, we choose the hyperparameters $a_0 = b_0 = 1$ in the Gamma prior for A and a vague prior for θ as $N(0, 10^2)$. For each replicate, we run the MCMC for $10,000$ iterations and treat the first $5,000$ as the burn-in.

The results for M1 and M2 are displayed in Table 7.1. M1 and M2 have smooth nuisance function models. We vary the sample size n from 20 to 400 for model M1 and M2 and apply the three methods P1, P2 and P3 on each. We record the root mean squared error (RMSE) for θ (under the Euclidean norm) and η (under the empirical norm) respectively. The average estimated standard error based on MCMC (SE) and the empirical coverage of nominal 0.95 credible intervals based on MCMC (CR95) are also reported. From Table 7.1, as n grows, the estimation accuracy of θ with the dependent priors improves. However, the RMSE for θ with the independent prior only significantly decreases as n goes from 20 to 50 and keeps around 0.1 as n further grows. On the other hand, the estimated standard error by P1 decays as that by P2 and P3. As a result of these, the actual coverage by P1 becomes significantly smaller than the nominal level as n grows. This phenomenon occurred with the independent prior empirically justifies the semiparametric bias we discussed after Lemma 53 and illustrates the necessity of compensating the bias by considering the dependent priors. As we expected, the RMSE for θ intimately depends on the RMSE for η: a large RMSE for η usually corresponds to a large RMSE for θ. In most cases, the RMSE for η in the less smooth model M1 is greater than that in

Table 7.1: Simulation results for the partially linear model with a smooth nuisance part based on 100 replicates.

	model	method	RMSE(θ)	SE	RMSE(η)	CR95
$n = 20$	M1	P1	0.277	0.155	0.571	0.91
		P2	0.188	0.176	0.454	0.97
		P3	0.150	0.181	0.392	0.99
	M2	P1	0.311	0.162	0.587	0.85
		P2	0.195	0.182	0.477	0.95
		P3	0.159	0.183	0.390	0.97
$n = 50$	M1	P1	0.115	0.078	0.308	0.92
		P2	0.085	0.082	0.274	0.96
		P3	0.083	0.083	0.270	0.95
	M2	P1	0.104	0.080	0.298	0.84
		P2	0.084	0.085	0.267	0.96
		P3	0.082	0.085	0.268	0.96
$n = 100$	M1	P1	0.103	0.052	0.225	0.83
		P2	0.056	0.056	0.202	0.95
		P3	0.053	0.056	0.204	0.96
	M2	P1	0.096	0.051	0.235	0.85
		P2	0.055	0.054	0.209	0.94
		P3	0.051	0.055	0.206	0.97
$n = 200$	M1	P1	0.106	0.038	0.230	0.62
		P2	0.042	0.038	0.197	0.93
		P3	0.036	0.038	0.187	0.97
	M2	P1	0.094	0.036	0.209	0.72
		P2	0.038	0.038	0.180	0.95
		P3	0.038	0.038	0.183	0.98
$n = 400$	M1	P1	0.115	0.035	0.289	0.38
		P2	0.030	0.028	0.187	0.93
		P3	0.025	0.028	0.187	0.98
	M2	P1	0.107	0.033	0.268	0.45
		P2	0.030	0.027	0.178	0.92
		P3	0.027	0.026	0.179	0.98

the model M2, leading to higher estimation accuracy of θ in M1 than that in M2. Another observation from the displayed results is that as n increases, the difference in the estimation efficiency between P2 and P3 becomes negligible. This might be attributed to the increasing accuracy of the estimation of $\widehat{h}$.

Table 7.2: Simulation results for the partially linear model with a nonsmooth nuisance part based on 100 replicates.

	model	method	RMSE(θ)	SE	RMSE(η)	CR95
$n = 20$	M3	P1	0.404	0.184	0.752	0.76
		P2	0.183	0.181	0.410	0.96
		P3	0.156	0.178	0.377	0.97
	M4	P1	0.207	0.148	0.495	0.92
		P2	0.148	0.180	0.390	0.98
		P3	0.144	0.180	0.409	0.99
$n = 50$	M3	P1	0.243	0.088	0.499	0.74
		P2	0.090	0.085	0.279	0.94
		P3	0.084	0.087	0.280	0.97
	M4	P1	0.194	0.089	0.408	0.80
		P2	0.084	0.087	0.270	0.97
		P3	0.084	0.087	0.265	0.97
$n = 100$	M3	P1	0.217	0.064	0.441	0.67
		P2	0.061	0.056	0.233	0.93
		P3	0.057	0.056	0.231	0.93
	M4	P1	0.122	0.052	0.309	0.84
		P2	0.059	0.055	0.221	0.96
		P3	0.058	0.055	0.219	0.95
$n = 200$	M3	P1	0.189	0.036	0.410	0.53
		P2	0.042	0.039	0.215	0.94
		P3	0.042	0.039	0.212	0.97
	M4	P1	0.106	0.042	0.271	0.77
		P2	0.041	0.038	0.204	0.98
		P3	0.040	0.038	0.203	0.97
$n = 400$	M3	P1	0.194	0.041	0.429	0.21
		P2	0.035	0.029	0.207	0.95
		P3	0.031	0.028	0.205	0.95
	M4	P1	0.115	0.033	0.282	0.65
		P2	0.033	0.028	0.193	0.94
		P3	0.030	0.028	0.193	0.96

Table 7.2 provides the results for M3 and M4, whose nuisance functions are non-smooth. As expected, the overall estimation accuracy in Table 7.2 is worse than that in Table 7.1. However, similar overall trends as those in Table 7.2 can be observed. For example, the estimation performance of P1 are generally worse than that of P2 and P3 and the semiparametric bias in P1 is more salient under M3 and M4 than under M1 and M2. In addition, the RMSE for θ associated with P1 under a non-smooth h^* is significantly worse than that under a smooth h^*. This is consistent with A4 because the semiparametric bias under the independent prior (PI) tends to be larger when h^* is less smooth.

7.5.2 General partially linear models with quasi-likelihood

The corresponding second order semiparametric BvM theorem of GPLM with quasi-likelihood is similar to that of partially linear models and we only provide the version for the dependent prior here. Similar to section 7.5.1, we consider a semiparametric adaptive prior for GPLM. Let $\widehat{h}$ be any estimator of the h given by Lemma 47 that satisfies assumption A6. We still focus on (7.30) based on the GP prior for the nuisance part.

Theorem 59. *Let $X_i = (U_i, V_i, Y_i)$, $i = 1, \ldots, n$, be a sample from the general partially linear model with quasi-likelihood* (7.4)*. Suppose that $\widehat{h}$ is an estimator of h that satisfies A6. Assume Assumption 1. Furthermore, if the following conditions are satisfied:*

1. *η_0 is Hölder α-smooth;*

2. *$P(U - E[U|V])^2 > 0$;*

Then under the prior (7.30)*, the following second order asymptotic expansion holds:*

$$\sup_A \left|\Pi(\theta \in A|X^{(n)}) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (nI_0^{-1})\big)(A)\right| = O_P\big(\sqrt{n}\rho_n^2 \log n + \sqrt{n}\kappa_n\rho_n \log n\big),$$

where

$$\widetilde{\Delta}_n = n^{-1/2} \sum_{i=1}^{n} I_0^{-1} W_i l_0(T_i)(U_i + h^*(V_i)) \overset{P_0}{\rightsquigarrow} N_k(0, I_0^{-1}),$$

$$I_0 = E_0\big[l_0(T) f_0(T)(U + h^*(V))(U + h^*(V))^T\big],$$

and $\rho_n = n^{-\alpha/(2\alpha+d)} (\log n)^{1+d}$.

7.6 Proofs of Theorem 48 and Theorem 56

The following lemma shows that the ILAN (7.15) and assumption Assumption 2 imply the second order semiparametric BvM. Here the notation $a_n \gtrsim b_n$ means that $a \geqslant cb$ for some $c > 0$.

Lemma 60. *Suppose that* $X_1, \ldots, X_n$ *are i.i.d. observations sampled from* $P_0 = P_0$. *If the following conditions hold:*

1. *The ILAN as* (7.15) *holds with a decaying rate* R_n *such that* $R_n(n^{-1/2} \log n) = o(1)$ *as* $n \to \infty$;
2. *The marginal prior for* θ *satisfies S1;*
3. *There exists a sequence* $(\rho_n : n \geqslant 1)$ *satisfying* $\rho_n \to 0$, $n\rho_n^2 \gtrsim -\log R_n(n^{-1/2} \log n)$ *and* $\sup_{|t| \leqslant \rho_n} R_n(t)/t^2 = o(n)$, *such that for* M *sufficiently large, assumption Assumption 2 holds,*

Then the marginal posterior for θ *has the following expansion in total variation,*

$$\sup_A \big|\Pi(\theta \in A | X_1, \ldots, X_n) - N_k\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n, (n\widetilde{I}_{\theta_0,\eta_0})^{-1}\big)(A)\big|$$

$$= O_P[R_n(n^{-1/2} \log n) + \delta_n], \tag{7.33}$$

where $\widetilde{\Delta}_n = n^{-1/2} \sum_{i=1}^n \widetilde{I}_{\theta_0,\eta_0}^{-1} \widetilde{l}_{\theta_0,\eta_0}(X_i) \overset{P_0}{\rightsquigarrow} N_k(0, I_{\theta_0,\eta_0}^{-1})$ *is defined by* (7.2).

The constraint on R_n in Lemma 60 is mild since a typical R_n has the form of $nt^3 + (n\rho_n + \sqrt{n})t^2 + (n\rho_n^2 + n\rho_n\kappa_n + \sqrt{n}\rho_n)t + \sqrt{n}\rho_n^2 + \delta_n$. Under this form of R_n, condition 1 requires that $\rho_n = o(n^{-1/4})$, which is a common condition obtained by various authors for proving the first order semiparametric BvM theorems. The condition $\rho_n \sim n^{-\alpha/(2\alpha+d)} \sim o(n^{-1/4})$ requires that $\alpha > d/2$ when the nuisance part η_0 is a d-variate α-smooth function.

The $\log n$ term appears in the conclusion of Lemma 60 is possible to be suppressed with more efforts. However, since $\log n$ terms commonly appear in the posterior convergence rate ρ_n of Bayesian nonparametric models, we exhibit the current result to avoid more involved conditions and longer proof. Comparing to the results in Bickel and Kleijn (2012), we replace their ILAN with the stronger condition (7.15) in order to exchange for a weaker requirement on the marginal convergence rate of θ as ρ_n instead of a parametric rate $n^{-1/2}$.

Theorem 48 is the direct consequence of Lemma 53 and Lemma 60. Theorem 56 can be proved by applying the arguments in Section 7.4.1, Lemma 55, Lemma 53 and Lemma 60.

Appendix A

Appendix for Chapter 2

A.1 Proofs of technical results in Chapter 2

A.1.1 Proof of Theorem 3

(2.6) describes the posterior convergence rate of the regression model with the constructed prior Π_n. Commonly, posterior convergence statements can be proved by applying the results in Ghosal et al. (2000) (Theorem 2.1 for IID observations) and Ghosal and van der Vaart (2007) (Theorem 1 for non-IID observations), which yields the following for the regression model:

$$E_{f_0,Q}\Pi_n\big(h(P_{f,Q}, P_{f_0,Q}) \geqslant M\epsilon_n \big| X_1, Y_1, \ldots, X_n, Y_n\big) \to 0,$$

where h is the Hellinger distance. However, h is bounded above by $||\cdot||_Q$, but is equivalent to this norm only when the function class Σ is uniformly bounded, which is less interesting. Therefore, we apply the techniques in Ghosal and van der Vaart (2007) that allow extensions of h to any distance d_n that satisfies the following test condition:

(T) There exists a sequence of test functions $\{\phi_n\}$ such that $P_{f_0,Q}\phi_n \leqslant e^{-\frac{1}{2}nd_n^2(f_1,f_0)}$ and $P_{f,Q}(1-\phi_n) \leqslant e^{-\frac{1}{2}nd_n^2(f_1,f_0)}$ for all $f \in \Sigma$ such that $d_n(f_1, f_0) \leqslant \frac{1}{18}d_n(f_1, f_0)$.

Suppose that for a sequence ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, constants C and c, and sets $\mathcal{F}_n \subset \Sigma$, we have

$$\log(\epsilon_n, \mathcal{F}_n, c^{-1}d_n) \leqslant n\epsilon_n^2 c^{-2}, \tag{A.1}$$

$$\Pi_n(\Sigma \backslash \mathcal{F}_n) \leqslant \exp\big(-n\epsilon_n^2 c^{-2}(C+4)\big), \tag{A.2}$$

$$\Pi_n\big(f : K(P_{f_0,Q}, P_{f,Q}) \leqslant \epsilon_n^2 c^{-2},\, V_{2,0}(P_{f_0,Q}, P_{f,Q}) \leqslant \epsilon_n^2 c^{-2}\big) \geqslant \exp(-n\epsilon_n^2 c^{-2} C), \tag{A.3}$$

where $K(P,Q) = P\log(p/g)$ is the Kullback-Leibler divergence between two probability distributions P and Q, and $V_{2,0}(P,Q) = P\big(\log(p/q) - K(P,Q)\big)^2$. Then under (T), by combining the proofs of Theorem 4 in Ghosal and van der Vaart (2007) and Theorem 2.1 in Ghosal et al. (2000), we have

$$E_{f_0,Q}\Pi_n\big(c^{-1}d_n(f,f_0) \geqslant M\epsilon_n \big| X_1, Y_1, \ldots, X_n, Y_n\big) \to 0,$$

where M is a sufficiently large constant independent with c. Note that (A.1)-(A.3) generalize the conditions in Theorem 2.1 in Ghosal et al. (2000) by allowing an arbitrary tuning parameter c. By the results in Birgé (2006), (T) is satisfied with $d_n(f,g) = \sigma||f-g||_Q$ for the regression model with random design $X_i \sim Q$. Therefore, by choosing $c = \sigma$, our generalized conditions allow us to track the impact of σ on the posterior convergence rate ϵ_n.

Return to our regression problem with an ϵ_n satisfying $n\epsilon_n^2 \geqslant \sigma^2 \log N(\epsilon_n, \Sigma, ||\cdot||_Q)$. Assume that $\mathcal{E}_n$ is an ϵ_n-covering set with $N(\epsilon_n, \Sigma, ||\cdot||_Q)$ elements. Let Π_n be the discrete uniform probability measure on the finite set $\mathcal{E}_n$. Let $\mathcal{F}_n = \Sigma$ for all n. We will prove (2.6) by verifying the conditions (A.1)-(A.3) with $c = \sigma$. (A.1) is satisfied by the constraint on ϵ_n and (A.2) is trivially satisfied by the choice of $\mathcal{F}_n$. So we only need to check (A.3).

The Kullback-Leibler divergence between two regression models $P_{f,Q}$ and $P_{f_0,Q}$ indexed by regression functions f and f_0 respectively is given by

$$P_{f_0,Q}\left(\log\frac{P_{f_0,Q}}{P_{f,Q}}\right) = \frac{1}{2\sigma^2}E_X\big(f(X) - f_0(X)\big)^2 = \frac{1}{2\sigma^2}||f-f_0||_Q^2.$$

Similarly, we have,

$$P_{f_0,Q}\bigg(\log\frac{P_{f_0,Q}}{P_{f_0,Q}}-K(P_{f_0,Q},P_{f,Q})\bigg)^2=\frac{1}{2\sigma^2}||f-f_0||_Q^2.$$

Therefore, for some universal constant C,

$$\big\{f:||f-f_0||_Q^2\leqslant\epsilon_n^2\big\}\subset$$
$$\bigg\{f:K(P_{f_0,Q},P_{f,Q})\leqslant C\epsilon_n^2\sigma^{-2},\ V_{2,0}(P_{f_0,Q},P_{f,Q})\leqslant C\epsilon_n^2\sigma^{-2}\bigg\}.$$

Since $\mathcal{E}_n$ forms an ϵ_n-covering set for Σ, there exists an $\tilde{f}$ such that $||\tilde{f}-f_0||_Q\leqslant\epsilon_n$ for $f_0\in\Sigma$. Therefore, we have

$$\Pi_n\big(f:K_{f_0,Q},P_{f,Q})\leqslant\epsilon_n^2\sigma^{-2},\ V_{2,0}(P_{f_0,Q},P_{f,Q})\leqslant\epsilon_n^2\sigma^{-2}\big)$$
$$\geqslant\Pi_n(f=\tilde{f})=\frac{1}{|\mathcal{E}_n|}=\exp\{-\log N(\epsilon_n,\Sigma,||\cdot||_Q)\}\geqslant\exp\{-n\epsilon_n^2\sigma^{-2}\},$$

which proves the condition (A.3). Therefore,

$$E_{f_0,Q}\Pi_n\big(f:||f-f_0||_Q>M\epsilon_n|X_1,Y_1,\ldots,X_n,Y_n\big)\to 0.$$

The second part can be proved similarly as Theorem 2.5 in Ghosal et al. (2000).

A.1.2 Proof of Theorem 4

Unlike the random-design perspective in the proof of Theorem 3, now we treat the regression model to be fixed-design by conditioning on $(X_1,\ldots,X_n)$. As a result, we use $P_f(Y|X)$ instead of $P_{f,Q}(X,Y)$ for the likelihood function in the proof. We first states two lemmas. The first lemma strengthens Lemma 8.1 in Ghosal et al. (2000) under the regression framework.

Lemma 61. *Assume $Y_i|X_i\sim N(f_0(X_i),\sigma^2)$ and Π to be a probability measure on the set $\{f:n^{-1}\sum_{i=1}^n K(P_{f_0,Q}(\cdot|X_i),P_{f,Q}(\cdot|X_i))\leqslant\epsilon^2\sigma^{-2}/2\}$ for a fixed $\epsilon>0$. Then*

for any $C > 0$,

$$P_{f_0}\bigg(\int \prod_{i=1}^{n} \frac{P_f(Y_i|X_i)}{P_{f_0}(Y_i|X_i)} d\Pi(f) \leqslant \exp(-(1+C)n\epsilon^2\sigma^{-2}) \bigg| X_1, \ldots, X_n \bigg) \leqslant \exp(-Cn\epsilon^2\sigma^{-2}).$$

Proof. By Cauchy's inequality,

$$\int \prod_{i=1}^{n} \frac{P_f(Y_i|X_i)}{P_{f_0}(Y_i|X_i)} d\Pi(f) \cdot \int \prod_{i=1}^{n} \frac{P_{f_0}(Y_i|X_i)}{P_f(Y_i|X_i)} d\Pi(f) \geqslant 1.$$

Combining the above with Markov inequality and Fubini's theorem yields

$$\begin{aligned}
&P_{f_0}\bigg(\int \prod_{i=1}^{n} \frac{P_f(Y_i|X_i)}{P_{f_0}(Y_i|X_i)} d\Pi(f) \leqslant \exp(-(1+C)n\epsilon^2\sigma^{-2}) \bigg| X_1, \ldots, X_n \bigg) \\
\leqslant &P_{f_0}\bigg(\int \prod_{i=1}^{n} \frac{P_{f_0}(Y_i|X_i)}{P_f(Y_i|X_i)} d\Pi(f) \geqslant \exp((1+C)n\epsilon^2\sigma^{-2}) \bigg| X_1, \ldots, X_n \bigg) \\
\leqslant &\exp(-(1+C)n\epsilon^2\sigma^{-2}) \int P_{f_0}\bigg(\exp\bigg\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \big\{ (f(X_i) - f_0(X_i))^2 \\
&\qquad - 2\epsilon_i(f(X_i) - f_0(X_i)) \big\} \bigg\} \bigg| X_1, \ldots, X_n \bigg) d\Pi_n(f) \\
= &\exp(-(1+C)n\epsilon^2\sigma^{-2}) \int \exp\left(\frac{n}{\sigma^2} ||f - f_0||_n^2 \right) d\Pi_n(f).
\end{aligned}$$

with $\epsilon_i = Y_i - f(X_i) \sim N(0, \sigma^2)$. In the regression framework, $n^{-1} \sum_{i=1}^{n} K(P_{f_0}(\cdot|X_i), P_f(\cdot|X_i)) = ||f - f_0||_n^2/(2\sigma^2)$, so on the set $\{f : n^{-1} \sum_{i=1}^{n} K(P_{f_0}(\cdot|X_i), P_f(\cdot|X_i)) \leqslant \epsilon^2\sigma^{-2}/2\}$, we have

$$P_{f_0}\bigg(\int \prod_{i=1}^{n} \frac{P_f(Y_i|X_i)}{P_{f_0}(Y_i|X_i)} d\Pi(f) \leqslant \exp(-(1+C)n\epsilon^2\sigma^{-2}) \bigg| X_1, \ldots, X_n \bigg) \leqslant \exp(-Cn\epsilon^2\sigma^{-2}).$$

□

The second lemma compares the $||\cdot||_Q$ norm with the $||\cdot||_n$ norm.

Lemma 62. *Suppose that Σ is uniformly bounded and ϵ_n satisfies*

$$n\epsilon_n^2 \geqslant \sigma^2 \sup_R \log N(\epsilon_n, \Sigma, ||\cdot||_R),$$

then for some c_1, c_2, c_3, c_4, c_5 and any $\eta \in (0,1)$,

$$P\left(\sup_{f\in\Sigma, ||f||_Q \geqslant c_1\epsilon_n/\eta} \left| \frac{||f||_n}{||f||_Q} - 1 \right| \geqslant \eta \right) \leqslant c_2 \exp(-c_3 n\epsilon_n^2),$$

$$P\left(\sup_{f\in\Sigma, ||f||_Q \leqslant c_1\epsilon_n} ||f||_n \geqslant c_5\epsilon_n \right) \leqslant c_2 \exp(-c_3 n\epsilon_n^2).$$

Proof. The second inequality is implied by Lemma 5.4 in van de Geer (2000). The proof of the first inequality is a combination of the proofs of Lemma 5.4 and Lemma 5.16 in van de Geer (2000), with the bracketing entropy condition replaced with the uniform covering entropy condition. □

Return to the proof of Theorem 4. Choose $d_n = \sigma||\cdot||_n$ in the proof of Theorem 3. By the results in Birgé (2006), (T) is satisfied with d_n for the regression model with fixed design. Moreover, we use the results in Ghosal and van der Vaart (2007) for non IID observations since now we have a regression model with fixed design (Ghosal and van der Vaart, 2007, section 7.7). Let $\mathcal{E}_n$ be an ϵ_n-covering set under $||\cdot||_n$, which contains $N(\epsilon_n, \Sigma, ||\cdot||_n)$ elements. Then the first part follows by adapting the proof of Theorem 4 in Ghosal and van der Vaart (2007) to d_n with the help of Lemma 62 and Lemma 61, where the decay rate of $E_{f_0,Q}\Pi_n\big(f : ||f - f_0||_n > M\epsilon_n | X_1, Y_1, \ldots, X_n, Y_n\big)$ is determined by the decay rate of $P_{f_0}\big(\int \prod_{i=1}^n \frac{P_f(Y_i|X_i)}{P_{f_0}(Y_i|X_i)} d\Pi(f) \leqslant \exp(-(1+C)n\epsilon^2\sigma^{-2}) \big| X_1, \ldots, X_n\big)$ for Π the restriction of Π_n on the set $\{f : n^{-1}\sum_{i=1}^n K(P_{f_0}(\cdot|X_i), P_f(\cdot|X_i)) \leqslant \epsilon^2\sigma^{-2}/2\}$ in Lemma 61. The second part is implied by Theorem 2.5 in Ghosal et al. (2000). The third part follows by

noticing the fact that for the posterior expectation $\hat{f} = \int f d\Pi_n(f|X_1, Y_1, \ldots, X_n, Y_n)$,

$$
\begin{aligned}
& E_{f_0,Q}\big(||\hat{f} - f_0||_Q^2\big) \lesssim E_{f_0,Q}\big(||\hat{f} - f_0||_n^2\big) \\
\leqslant & E_{f_0,Q} \int ||f - f_0||_n^2 d\Pi_n(f|X_1, Y_1, \ldots, X_n, Y_n) \\
\leqslant & M^2\epsilon_n^2 + 4C_1^2 E_{f_0,Q}\Pi_n\big(||f - f_0||_n^2 \geqslant M\epsilon_n|X_1, Y_1, \ldots, X_n, Y_n\big) \leqslant D\epsilon_n^2,
\end{aligned}
$$

where the first step follows by the convexity of $||\cdot||_n^2$.

A.1.3 Proof of Theorem 5

The results of Theorem 5 are standard minimax risk lower bounds for regression. For self-containment, we sketch a proof, which follows a standard information-theoretic argument using Fano's inequality (Yu, 1997; Yang and Barron, 1999; Tsybakov, 2009):

Step 1: Reduction to bounds in probability. By the Markov inequality, for any $\tau > 0$,

$$
E_{f,Q}||\hat{f} - f||_Q^2 \geqslant P_{f,Q}\big\{||\hat{f} - f||_Q \geqslant \epsilon_n\big\}\epsilon_n^2. \tag{A.4}
$$

Therefore, in order to prove that ϵ_n is a lower bound of the minimax risk, it suffices to show that $\inf_{\hat{f}} \sup_{f\in\Sigma} P_{f,Q}\big\{||\hat{f} - f||_Q \geqslant \epsilon_n\big)$ is lower bounded by some universal constant $\tau > 0$ independent of n.

Step 2: Reduction to a finite number of hypotheses. Since

$$
\inf_{\hat{f}} \sup_{f\in\Sigma} P_{f,Q}\big\{||\hat{f} - f_0||_Q \geqslant \epsilon_n\big) \geqslant \inf_{\hat{f}} \sup_{f\in\{f_0,\ldots,f_N\}} P_{f,Q}\big\{||\hat{f} - f_0||_Q \geqslant \epsilon_n\big),
$$

for any finite set $\{f_0, \ldots, f_N\}$ contained in Σ, we can reduce the original inf involving infinite number to only $N + 1$ many well chosen model elements. Intuitively, these elements should be well-separated in order to represent the entire model space.

Step 3: Choice of $2\epsilon_n$-separated hypotheses. If

$$
||f_s - f_t||_Q \geqslant 2\epsilon_n, \quad \forall s, t : s \neq t,
$$

then for any estimator $\hat{f}$,

$$P_{f_s,Q}\big(||\hat{f}-f_s||_Q \geqslant \epsilon_n\big) \geqslant P_{f_s,Q}(Z \neq s), \quad s=0,1,\ldots,M,$$

where Z is the value of t for which $||\hat{f}-f_t||_Q$ is minimized.

Therefore, if a random variable S is defined to be uniformly distributed on $\{0,1,\ldots,N\}$ and the observation follows P_{f_s} conditioning on $S=s$, then

$$\inf_{\hat{f}} \sup_{f\in\{f_0,\ldots,f_N\}} P_{f,Q}\big\{||\hat{f}-f_0||_Q$$

$$\geqslant\epsilon_n\big) \geqslant \inf_{\hat{f}} \frac{1}{N+1}\sum_{s=0}^{N} P_{f_s,Q}(Z\neq s) = \inf_{Z} P(Z\neq S).$$

From Fano's inequality (Cover and Thomas (1991)), we have:

$$P(Z\neq S|X^n) \geqslant 1-\frac{I_{X^n}(S;Y^n)+\log 2}{\log N},$$

where X^n and Y^n are the observed covariate and response with sample size n and $I_{X^n}(S;Y^n)$ denotes the conditional mutual information between S and Y^n given X_n. Therefore,

$$\inf_{\hat{f}}\sup_{f\in\Sigma} P_{f,Q}\big\{||\hat{f}-f||_Q \geqslant \epsilon_n\big) \geqslant 1-\frac{E_{X^n}\big[I_{X^n}(S;Y^n)\big]+\log 2}{\log N}. \tag{A.5}$$

By definition, this conditional mutual information

$$\begin{aligned} I_{X^n}(S;Y^n) &= \frac{1}{N+1}\sum_{s=0}^{N} KL(P_{f_s,Q},\bar{P}) \\ =&\frac{1}{N+1}\sum_{s=0}^{N} KL(P_{f_s,Q},P_0) - KL(\bar{P},P_0) \leqslant \frac{1}{N+1}\sum_{s=0}^{N} KL(P_{f_s,Q},P_0). \end{aligned} \tag{A.6}$$

where $KL(P,Q)=P\log(dP/dQ)$ is the Kullback-Leibler divergence between P and Q, $\bar{P}=\frac{1}{N+1}\sum_{s=0}^{N}P_{f_s}$ and P_0 can be an arbitrary model, which is taken to be $P_{f_0,Q}$ in the following. In the regression settings,

$$E_{X^n}KL(P_s,P_{f_0}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n} E_{X^n}\big(f_s(X_i)-f_0(X_i)\big)^2 = \frac{n}{2\sigma^2}||f_s-f_0||_Q^2. \tag{A.7}$$

Therefore, by (A.5), (A.6) and (A.7),

$$\inf_{\hat{f}} \sup_{f\in\Sigma} E_{f,Q}||\hat{f}-f||_Q^2 \geqslant \epsilon_n^2 \inf_{\hat{f}} \sup_{f\in\Sigma} P_{f,Q}\big\{||\hat{f}-f||_Q \geqslant \epsilon_n\big\} \tag{A.8}$$

$$\geqslant \epsilon_n^2\left\{1-\frac{n\max_s ||f_s-f_0||_Q^2}{\sigma^2\log N}\right\}. \tag{A.9}$$

If $\{f_s-f_0\}_{s=1}^N$ forms a modified $2\epsilon_n$-packing set, i.e.

$$||f_s-f_0||_Q^2 \leqslant K\epsilon_n^2, \quad ||f_s-f_t|| \geqslant 2\epsilon_n, \text{ for all } s\neq t,$$

then the theorem can be proved by the choice of ϵ_n and taking $N = C(2\epsilon_n, K, \Sigma, ||\cdot||_Q)$.

A.1.4 Proof of Theorem 6

Let $\{\alpha_1,\ldots,\alpha_k\}$ be some tuning parameters satisfying $\sum_{s=1}^k \alpha_s^2 = 1$ that will be determined later. By F2, any two different functions f_s and g_s in $\mathcal{E}_s(\alpha\epsilon)$ satisfy $||f_s-g_s||_Q \geqslant \alpha_s\epsilon$ and $||f_s||_Q \leqslant K\alpha_s\epsilon$. In the following, we apply a probabilistic argument to construct a desired $\epsilon/2$-packing set $\mathcal{E}(\epsilon/2)$ for $\mathcal{F}$ as a subset of $\Omega = \bigoplus_{s=1}^k \mathcal{E}_s(\alpha_s\epsilon)$ such that the size of the subset is comparable to $\prod_{s=1}^k C_s(\alpha_s\epsilon)$. Then, by F1,

$$||\sum_{s=1}^k f_s||_Q^2 \leqslant B\sum_{s=1}^k ||f_s||_Q^2 \leqslant BK^2\sum_{s=1}^k \alpha_s^2\epsilon^2 = BK^2\epsilon^2.$$

Hence, $\mathcal{E}(\epsilon/2)$ is also a modified $\epsilon/2$-packing set for $\mathcal{F}$.

For notational convenience, $C_s(\alpha_s\epsilon, K)$ will be further abbreviated as C_s when α_s, ϵ and K are fixed. However, $C_s(\alpha_s\epsilon)$ will also be used when we want to emphasize the dependence of C_s on α_s and ϵ.

Consider the probability space (Ω,Σ,P), where the Sigma-field Σ is composed of all subsets of Ω and P is the uniform measure over Ω. If $F=(F_1,\ldots,F_k)$ is a random variable on Ω with distribution P, then it is easy to see that $F_1,\ldots,F_k$ are independent with marginal distributions $P(F_s=f_s)=1/C_s$, for any $f_s\in\mathcal{F}_s$,

$s = 1, \ldots, k$. For M independent copies $\{F^{(m)}\}_{m=1}^{M}$ of F, where M will be determined later, our goal is to estimate

$$P\bigg\{||F^{(i)} - F^{(j)}||_Q \geqslant \frac{\epsilon}{2},\ \forall 1 \leqslant i < j \leqslant M\bigg\}.$$

If this probability is positive, then we can choose $f^{(1)}, \ldots, f^{(M)}$ from the sample space Ω, such that $\{f^{(m)}\}_{m=1}^{M}$ forms an ϵ-packing set of $\mathcal{F}$ with M elements. Since

$$\begin{aligned}
&P\bigg\{||F^{(i)} - F^{(j)}||_Q \geqslant \frac{\epsilon}{2},\ \forall 1 \leqslant i < j \leqslant M\bigg\} \\
\geqslant &1 - \sum_{1\leqslant i<j\leqslant M} P\bigg\{||F^{(i)} - F^{(j)}||_Q < \frac{\epsilon}{2}\bigg\} \\
= &1 - \frac{1}{2}M(M-1)\cdot P\bigg\{||F^{(1)} - F^{(2)}||_Q < \frac{\epsilon}{2}\bigg\},
\end{aligned}$$

we want to find the maximal number M so that

$$P\bigg\{||F^{(1)} - F^{(2)}||_Q < \frac{\epsilon}{2}\bigg\} \lesssim \exp(-2\log M). \tag{A.10}$$

Let $Z_s = I(F_s^{(1)} \neq F_s^{(2)})$ be an indicator variable. By independence of $F^{(1)}, \ldots, F^{(s)}$, $Z_1, \ldots, Z_s$ are also independent with marginal distributions $Z_s \sim \text{Bernoulli}(1-1/C_s)$. By the assumptions on $\mathcal{E}_s(\alpha_s\epsilon)$, we have

$$||F^{(1)} - F^{(2)}||_Q^2 = \sum_{s=1}^{k} ||F_s^{(1)} - F_s^{(2)}||_Q^2 \geqslant \epsilon^2 \sum_{s=1}^{k} \alpha_s^2 Z_s.$$

Therefore

$$P\bigg\{||F^{(1)} - F^{(2)}||_Q < \frac{\epsilon}{2}\bigg\} \leqslant P\bigg\{\sum_{s=1}^{k} \alpha_s^2 Z_s < \frac{1}{4}\bigg\}.$$

For a tuning parameter $\lambda > 0$, by Markov inequality and independence, we have

$$\begin{aligned}
P\bigg\{\sum_{s=1}^{k} \alpha_s^2 Z_s < \frac{1}{4}\bigg\} \leqslant &P\big\{e^{-\lambda\sum_{s=1}^{k}\alpha_s^2 Z_s} > e^{-\frac{\lambda}{4}}\big\} \leqslant e^{\frac{\lambda}{4}} \prod_{s=1}^{k} E\big\{e^{-\lambda\alpha_s^2 Z_s}\big\} \\
\leqslant &e^{\frac{\lambda}{4}} \prod_{s=1}^{k} \bigg\{\frac{1}{C_s} + e^{-\lambda\alpha_s^2}\bigg\} = e^{-\frac{3}{4}\lambda} \prod_{s=1}^{k} \big\{1 + e^{\lambda\alpha_s^2 - \log C_s}\big\},
\end{aligned}$$

where the last step holds since $\sum_{s=1}^{k} \alpha_s^2 = 1$. Since for any $x \in \mathbb{R}$, $1 + e^x \leqslant 2e^{xI(x \geqslant 0)}$, we further have

$$P\Big\{ \sum_{s=1}^{k} \alpha_s^2 Z_s < \frac{1}{4} \Big\} \lesssim \exp\Big\{ -\frac{3}{4}\lambda + \sum_{s:\log C_s/\alpha_s^2 \leqslant \lambda} (\lambda \alpha_s^2 - \log C_s) \Big\}.$$

Let $(\bar{\alpha}_1, \ldots, \bar{\alpha}_k)$ be the solution of

$$\frac{\log C_1(\alpha_1 \epsilon)}{\alpha_1^2} = \cdots = \frac{\log C_s(\alpha_s \epsilon)}{\alpha_s^2} = \cdots = \frac{\log C_k(\alpha_k \epsilon)}{\alpha_k^2} = \sum_{s=1}^{k} \log C_s(\alpha_s \epsilon) \triangleq A.$$

By setting $\lambda = A$, we obtain

$$\begin{aligned} P\Big\{ \sum_{s=1}^{k} \alpha_s^2 Z_s < \frac{1}{4} \Big\} &\lesssim \exp\Big\{ \frac{1}{4}\lambda - \sum_{s=1}^{k} \log C_s(\alpha_s \epsilon) \Big\} \\ &= \exp\Big\{ -\frac{3}{4} \sum_{s=1}^{k} \log C_s(\alpha_s \epsilon) \Big\}. \end{aligned}$$

Therefore, we can choose $M \sim \sum_{s=1}^{k} \log C_s(\alpha_s \epsilon)$ in (A.10), which finishes the proof of the first part.

For the second part, for each $s \in \{1, \ldots, k\}$, let $\mathcal{G}_s(\alpha_s \epsilon)$ be the corresponding $\alpha_s \epsilon$-covering set for $\mathcal{F}_s$ with $N_s(\alpha_s \epsilon)$ elements, i.e. for any function f_s in $\mathcal{F}_s$, there exists some g_s in $\mathcal{G}_s(\alpha_s \epsilon)$ such that $||f_s - g_s||_Q \leqslant \alpha_s \epsilon$. As a result, for any $f = \sum_{s=1}^{k} f_s \in \mathcal{F}$, we can find $g = \sum_{s=1}^{k} g_s \in \Delta = \bigoplus_{s=1}^{k} \mathcal{G}_s(\alpha_s \epsilon)$, such that $||f_s - g_s||_Q \leqslant \alpha_s \epsilon$ holds for each s, which yields

$$||f - g||_Q^2 \leqslant B \sum_{s=1}^{k} ||f_s - g_s||_Q^2 \leqslant B \sum_{s=1}^{k} \alpha_s^2 \epsilon^2 = B\epsilon^2.$$

Therefore, Δ forms an $\sqrt{B}\epsilon$-covering set for $\mathcal{F}$. Moreover, $\log |\Delta| = \sum_{s=1}^{k} \log N_s(\alpha_s \epsilon)$.

A.1.5 Proof of Corollary 7

By Theorem 3, Theorem 5 and Theorem 6, ϵ_n is the solution of $\sum_{s=1}^{k} \log \hat{N}_s(\bar{\alpha}_s\epsilon) \sim n\epsilon^2$, with $(\bar{\alpha}_1, \ldots, \bar{\alpha}_k) \in \mathbb{R}_+^k$ satisfying:

$$\frac{\log \hat{N}_1(\alpha_1\epsilon)}{\alpha_1^2} = \cdots = \frac{\log \hat{N}_s(\alpha_s\epsilon)}{\alpha_s^2} = \cdots = \frac{\log \hat{N}_k(\alpha_k\epsilon)}{\alpha_k^2} \sim n\epsilon^2.$$

If we let $\delta_s = \alpha_s\epsilon$, then the above is equivalent to $\log \hat{N}_s(\delta_s) \sim n\delta_s^2$ for $s = 1, \ldots, k$, with $\epsilon^2 = \sum_{s=1}^{k} \alpha_s^2\epsilon^2 = \sum_{s=1}^{k} \delta_s^2$.

A.1.6 Proof of Lemma 8

We work out a proof though a number of smaller parts.

Modified packing entropy of $\Sigma(\alpha, L, d)$

We first consider the case when Q is the uniform distribution on $[-1, 1]$. Let $K : [-1, 1]^d \to \mathbb{R}$ be a d-variate function satisfying

$$K \in C^\infty(\mathbb{R}^d), \ \int_{-1}^{1} K(u)du_j = 0, \text{ for any } u_{-j} \in [-1, 1]^{d-1} \text{ and } j = 1, \ldots, d. \quad \text{(A.11)}$$

Note that the last requirement is not need for the current proof for $\Sigma(\alpha, L, d)$ but will play a key role for the other two cases. Such functions K exist. For example, we can take

$$K(u) = c_1 K_0(u), \text{ where } K_0(u) = \prod_{j=1}^{d} \left\{ \exp\left(-\frac{1}{1 - |u_j|^2} \right) - c_2 \right\} I(|u_j| \leqslant 1), \quad \text{(A.12)}$$

for any $u \in \mathbb{R}^d$ with $c_2 = \int_{-1}^{1} \exp\{-1/(1 - |t|^2)\}dt$ and $c_1 > 0$ such that K satisfies (2.1) with $L = 1$.

Fix $h > 0$ as a positive tuning parameter to be determined later. Let $m = \lfloor \frac{1}{h} \rfloor$, $M = m^d$ and $\{x_k : 1 \leqslant k \leqslant M\}$ be a sequence of regular grids on $[-1, 1]^d$ with the

form

$$\left(\frac{k_1 - 1/2}{m}, \ldots, \frac{k_d - 1/2}{m}\right), \text{ for all } (k_1, \ldots, k_d) \in \{1, \ldots, m\}^d.$$

The order in x_k can be arbitrary. Define

$$\phi_k(x) = Lh^\alpha K\left(\frac{x - x_k}{h}\right), \quad k = 1, \ldots, M, \quad x \in [-1, 1]^d. \tag{A.13}$$

For any multi-index $|a|$, the mixed partial derivative with respect to a of ϕ_k is

$$D^a \phi_k(x) = Lh^{\alpha - |a|} D^a K\left(\frac{x - x_k}{h}\right).$$

Since the support of ϕ_k is $[x_k - h, x_k + h]$, we have that for $l = \lfloor \alpha \rfloor$,

$$\begin{aligned} &\max_{|a|=l} \sup_{x,y \in [-1,1]^d, x \neq y} \frac{|D^a \phi_k(x) - D^a \phi_k(y)|}{|x - y|^{\alpha - l}} \\ \leqslant & L \max_{|a|=l} \sup_{x,y \in [-1,1]^d, x \neq y} \frac{|D^a K(x) - D^a K(y)|}{|x - y|^{\alpha - l}} \leqslant L, \end{aligned}$$

which implies that $\phi_k(x) \in \Sigma(\alpha, L, d)$ for all k.

Denote the set of all binary sequence of length M by $\Omega = \big\{\omega = (\omega_1, \ldots, \omega_M) : \omega_i \in \{0, 1\}\big\}$. The desired ϵ-packing set will be chosen from the collection of functions $\mathcal{E} = \big\{f_\omega(x) = \sum_{k=1}^M \omega_k \phi_k(x), \omega \in \Omega\big\}$. Since for any $k \neq k'$, ϕ_k and $\phi_{k'}$ have distinct support, $\mathcal{E}$ is a subset of $\Sigma(\alpha, L, d)$. Moreover, for all $\omega, \omega' \in \Omega$, we have

$$\begin{aligned} d(f_\omega, f_{\omega'}) =& \Big[\int_{[-1,1]^d} (f_\omega(x) - f_{\omega'}(x))^2 dx \Big]^{1/2} \\ =& \Big[\sum_{k=1}^M (\omega_k - \omega_k')^2 \int_{\Delta_k} \phi_k^2(x) dx \Big]^{1/2} = Lh^{\alpha + d/2} ||K|| \sqrt{\rho(\omega, \omega')}, \end{aligned} \tag{A.14}$$

where Δ_k is a d-dim square with edge length $1/m$ centered at x_k and $\rho(\omega, \omega') = \sum_{k=1}^M I(\omega_k \neq \omega_k')$ is the Hamming distance between the ω and ω'.

By Lemma 2.9 in Tsybakov (2009), for $M \geqslant 8$, there exists a subset $\Omega_0 = \{\omega^{(0)}, \ldots, \omega^{(N)}\}$ of Ω with $N \geqslant 2^{M/8}$ such that $\omega^{(0)} = (0, \ldots, 0)$ and

$$\rho(\omega^{(k)}, \omega^{(k')}) \geqslant \frac{M}{8}, \quad \forall\ 0 \leqslant k < k' \leqslant N.$$

Therefore for any different $\omega^{(k)}, \omega^{(k')} \in \Omega_0$, by (A.14) and the definition of m, we have

$$d(f_{\omega^{(k)}}, f_{\omega^{(k')}}) \geqslant Lh^{\alpha+d/2}||K||\sqrt{\frac{h^{-d}}{8}} = \frac{L||K||}{2\sqrt{2}}h^{\alpha}.$$

In addition, by (A.14), we have

$$||f_{\omega^{(k)}}||_Q = d(f_{\omega^{(k)}}, 0) \leqslant Lh^{\alpha+d/2}||K||\sqrt{M} = L||K||h^{\alpha}.$$

Therefore by choosing $h = \left(2\sqrt{2}\epsilon/(L||K||)\right)^{\frac{1}{\alpha}}$, the set $\mathcal{E}_0(\epsilon) = \left\{f_\omega : \omega \in \Omega_0\right\}$ forms the desired ϵ-packing set of $\Sigma(\alpha, L, d)$ with

$$\log|\mathcal{E}_0(\epsilon)| \geqslant \frac{M}{8} \geqslant \frac{1}{8}\left(\frac{1}{h}\right)^d = K_1\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}},$$

with $K_1 = \frac{1}{8}\left(\frac{||K||}{2\sqrt{2}}\right)^{d/\alpha}$ and $K_2 = 2\sqrt{2}$. By the construction of K in (A.12), $||K|| = ||K||/||K||_{C^\alpha}$ is independent of L.

For general Q, ϕ_k can still be constructed by (A.13), but with kernel K_k being dependent on k such that $\int_0^1 \phi_k(x)dx_j = 0$ for any $x_{-j} \in [0,1]^{d-1}$ and $j = 1, \ldots, d$. This can be achieved by allowing each product component in (A.12) to have different $c_{2,j}$. By the assumption on q, $||K_k||$ will be both upper and lower bounded by multiples of $||K||$ and (A.14) will still be valid with different multiplicative constant.

Modified packing entropy of $\Sigma_S(\alpha, L, d, p)$

Similar to the proof for $\Sigma(\alpha, L, d)$, for notation simplicity, we assume that Q is the uniform distribution on $[-1, 1]$. To prove the conclusion, we need to construct a set of mutually orthogonal modified ϵ-packing sets $\mathcal{E}_0^b(\epsilon)$'s for all function space

$\Sigma^b(\alpha, L, d) = \{f(X^b) : f \in \Sigma(\alpha, L, d)\}$ with binary inclusion vector b's satisfying $|b| = d$, that is:

(a). For any two inclusion vectors $b \neq b'$ with $|b| = |b'| = d$, $\mathcal{E}_0^b(\epsilon) \perp \mathcal{E}_0^{b'}(\epsilon)$, i.e. if $f \in \mathcal{E}_0^b(\epsilon)$ and $f' \in \mathcal{E}_0^{b'}(\epsilon)$, then

$$\langle f, f' \rangle_Q = \int_{[-1,1]^{|b \cup b'|}} f(x^b) f'(x^{b'}) dx^{b \cup b'} = 0.$$

(b). For each inclusion vector b with $|b| = d$, functions in $\mathcal{E}_0^b(\epsilon)$ satisfies (i) and (ii) in the lemma. Moreover, $|\mathcal{E}_0^b(\epsilon)| \geqslant K(L/\epsilon)^{d/\alpha}$ for some $K > 0$ for each b with size d.

If this construction is possible, then a desired ϵ-packing set for $\Sigma_S(\alpha, L, d, p)$ can be specified as $\mathcal{E}_S(\epsilon) = \bigcup_{b:|b|=d} \mathcal{E}_0^b(\epsilon)$, where b in this union ranges from all possible inclusion vectors with size d. In fact, for any two functions f, f' in $\mathcal{E}_S(\epsilon)$, if they come from the same $\mathcal{E}_0^b$, then by construction of $\mathcal{E}_0^b(\epsilon)$, $||f - f'||_Q \geqslant \epsilon$; If they come from different $\mathcal{E}_0^b(\epsilon)$'s, then by the orthogonality condition (a) and the fact that 0 belongs to $\mathcal{E}_0^b(\epsilon)$, $||f - f'||_Q^2 = ||f - 0||_Q^2 + ||f' - 0||_Q^2 \geqslant 2\epsilon^2$. In both situations, we have $||f - f'||_Q \geqslant \epsilon$. Combining this result with condition (b), $\mathcal{E}_S(\epsilon)$ forms a modified ϵ-packing set for $\Sigma_S(\alpha, L, d, p)$. Moreover, the size of $\mathcal{E}_S(\epsilon)$ satisfies

$$\log |\mathcal{E}_S(\epsilon)| \geqslant K_1 \left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}} + \log \binom{p}{d} \sim K_1 \left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}} + d \log \frac{p}{d}.$$

In the following, we construct such a $\mathcal{E}_0^b(\epsilon)$ satisfying condition (a) and (b). In this construction, we use the crucial property in (A.11) that

$$\int_{\mathbb{R}} K(u) du_j = 0, \text{ for all } u_{-j} \in [-1, 1]^{d-1} \text{ and } \forall j = 1, \ldots, d. \tag{A.15}$$

For each fixed inclusion vector b with $|b| = d$, $\mathcal{E}_0^b(\epsilon)$ is constructed as in the proof for $\Sigma(\alpha, L, d)$. With this construction, we only need to verify condition (a). Under

the same notations, it suffices to prove $\phi_k(x^b) \perp \phi_{k'}(x^{b'})$ for $1 \leqslant k, k' \leqslant M$ and $b \neq b'$. In fact, since $k \neq k'$, we can always find a index j_0 such that $b_{j_0} = 0$ and $b'_{j_0} = 1$. With this j_0, we have

$$\langle \phi_k(x^b), \phi_{k'}(x^{b'}) \rangle_Q = L^2 h^{2\alpha} \int_{[-1,1]^{|b \cup b'|}} K\left(\frac{x^b - x^b_k}{h}\right) K\left(\frac{x^{b'} - x^{b'}_{k'}}{h}\right) dx^{b \cup b'}$$

$$= L^2 h^{2\alpha} \int_{[-1,1]^{|b \cup b'|-1}} K\left(\frac{x^b - x^b_k}{h}\right) \left\{ \int_{-1}^{1} K\left(\frac{x^{b'} - x^{b'}_{k'}}{h}\right) dx_{j_0} \right\} \prod_{j \in b \cup b' \setminus \{j_0\}} dx_j = 0,$$

where the second last step follows by Fubini's theorem and the last step follows from (A.15).

Modified packing entropy of $\Sigma_A\big((\alpha_1, \ldots, \alpha_k), (L_1, \ldots, L_k), (d_1, \ldots, d_k),\, p, \bar{d}\big)$

By definition, the size difference between $\Sigma_A\big((\alpha_1, \ldots, \alpha_k),\, (L_1, \ldots, L_k),\, (d_1, \ldots, d_k),\, p,\, \bar{d}\big)$ and $\bigoplus_{s=1}^{k} \Sigma_S(\alpha_s, L_s, d_s)$ is negligible for large p, so we only need to provide a construction for the latter. As the condition (a) in the proof for $\Sigma_S(\alpha, L, d, p)$ suggests, the modified packing sets for different additive components are orthogonal. Hence, the conclusion is an easy consequence of the second part and Theorem 6.

A.1.7 Proof of Lemma 9

Covering entropy of $\Sigma_S(\alpha, L, d, p)$

By the discussions before Lemma 9, for any inclusion vector b, we can find an ϵ covering set $\mathcal{E}^b$ for the subset under $||\cdot||_Q$ consisted of all functions in $\Sigma_S(\alpha, L, d, p)$ that depend on the d variables selected by the b, such that

$$\log N\big(\epsilon, \mathcal{E}^b, ||\cdot||_Q\big) \leqslant K\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}}.$$

Then an ϵ covering set for $\Sigma_S(\alpha, L, d, p)$ can be chosen as the union of all such $\mathcal{E}^b$'s with b ranging over all inclusion vectors with size d. Since there are $\binom{p}{d}$ such inclusion

vectors b's, we conclude that

$$\log N\big(\epsilon, \Sigma_S(\alpha, L, d, p), ||\cdot||_Q\big) \leqslant \sum_{b:|b|=d} \log N\big(\epsilon, \mathcal{E}^b, ||\cdot||_Q\big)$$

$$= K\left(\frac{L}{\epsilon}\right)^{\frac{d}{\alpha}} + d\log\frac{p}{d}.$$

Covering entropy of $\Sigma_A\big((\alpha_1,\ldots,\alpha_k),(L_1,\ldots,L_k),(d_1,\ldots,d_k),p,\bar{d}\big)$

The conclusion follows by the covering entropy upper bound for $\Sigma_S(\alpha, L, d, p)$ and the second half of Theorem 6.

A.1.8 Proof of Theorem 1

The result follows by applying Theorem 3, Theorem 5, Lemma 8 and Lemma 9.

A.1.9 Proof of Theorem 2

The result follows by applying Corollary 7, Lemma 8 and Lemma 9.

A.1.10 Proof of Lemma 10

The proof is extracted from some key steps in van der Vaart and van Zanten (2009), which help understand how the sieve construction works and how the parameters (M, r, ϵ, δ) balance each other.

By Lemma 4.6, Lemma 4.7 in van der Vaart and van Zanten (2009) and Borell's inequality, for any $a \leqslant r$,

$$P(W_a \notin \mathcal{B}_{M,r,\epsilon,\delta}) \leqslant e^{-M^2/8},$$

for $M^2 > C_0 r^d (\log(r/\epsilon))^{1+d},\ r > 1,\ \epsilon < \epsilon_0$, where ϵ_0 is some fixed positive number. The above inequality provides the complementary probability for a fixed inverse bandwidth parameter a.

Combining the above complementary probability, the Lemma 4.9 in van der Vaart and van Zanten (2009) and the exponential tail for the prior density $g(a)$, we have

the following complementary probability for a random inverse bandwidth A for all r larger than another fixed positive constant a_0:

$$\begin{aligned} P(W_A \notin \mathcal{B}_{M,r,\epsilon,\delta}) \leqslant & P(A > r) + \int_0^r P(W^a \notin \mathcal{B}_{M,r,\epsilon,\delta}) g(a) da \\ \leqslant & C_1 r^{p-d+1} e^{-C_2 r^d} + e^{-M^2/8}. \end{aligned}$$

This proves (2.14).

By Lemma 4.8 in van der Vaart and van Zanten (2009), for $\epsilon > \tau \delta M$, where τ is some positive constant,

$$N\left(3\epsilon, \bigcup_{a<\delta} M\mathbb{H}_1^a + \epsilon \mathbb{B}_1, ||\cdot||_\infty\right) \leqslant \frac{2C_4 M}{\epsilon}.$$

By Lemma 4.5 in van der Vaart and van Zanten (2009), for some constant $K > 0$ and any $\epsilon < 1/2$,

$$\log N(\epsilon, \mathbb{H}_1^a, ||\cdot||_\infty) \leqslant K a^d \left(\log \frac{1}{\epsilon}\right)^{1+d}.$$

Combining the above two and choosing $\delta = \epsilon/(2d^{3/2}M)$ yields

$$\log N(3\epsilon, \mathcal{B}_{M,r,\epsilon,\delta}, ||\cdot||_\infty) \leqslant C_3 r^d \left(\log \frac{M^{3/2}\sqrt{2d^{3/2}r}}{\epsilon^{3/2}}\right)^{1+d} + 2\log \frac{C_4 M}{\epsilon},$$

which proves (2.15).

A.1.11 Proof of Theorem 11

To apply the standard procedure in Ghosal et al. (2000) to determine the posterior convergence rate ϵ_n, we need to construct a sequence of sieves $(\mathcal{F}_n : n \geqslant 1)$ such that $\log N(\epsilon_n, \mathcal{F}_n, ||\cdot||) \leqslant n\epsilon_n^2$ and $P(\mathcal{F}_n^c) \leqslant e^{-Cn\epsilon_n^2}$, which are similar to condition (2.10) and (2.11). However, in the variable selection context, to keep the complementary probability small, the entropy number $\log N(\epsilon_n, \mathcal{F}_n^d, ||\cdot||)$ associated with $\mathcal{F}_n^d$, the sieve corresponding to d variate, often increases exponentially fast in d. This will

deteriorate the contraction rate to $n^{-\alpha_0/(2\alpha_0+d_0)}$ if the new sieve is simply constructed as $\bigcup_{d\leqslant d_0}\mathcal{F}_n^d$ and d_0 is larger than the true number $|b_0|$ of important predictors. So we need to modify $\mathcal{F}_n^d$ so that its entropy number can be of the same order as $\mathcal{F}_n^{|b_0|}$. Generally, this modification might not be possible. However, for GP, the flexibility in choosing $\bar{\epsilon}_n$ as mentioned in (2.16) enables the sieve sequence to adapt the contraction rate to true dimension $|b_0|$.

Let $(\mathcal{B}_n^b : n \geqslant 1)$ be a b-dependent sequence of sieves associated with $\bar{\epsilon}_n$ in (2.16). The sequence $(\bar{\epsilon}_n : n \geqslant 1)$ will be specified later. We construct the new sieves as

$$\mathcal{B}_n = \bigcup_{b\in\{0,1\}^p:|b|\leqslant d_0} \mathcal{B}_n^b. \tag{A.16}$$

$\mathcal{B}_n$ can be viewed as a subset of functions depending on at most d_0 components of $x \in [0,1]^p$. Since there are at most p^{d_0} such b in the union, by (2.16), we have the following bound for the entropy number of this sieve:

$$\begin{aligned}\log N(L\bar{\epsilon}_n, \mathcal{B}_n, ||\cdot||_\infty) \leqslant& d_0 \log p + \max_{b\in\{0,1\}^p:|b|\leqslant d_0}\{\log N(L\bar{\epsilon}_n, \mathcal{B}_n^b, ||\cdot||_\infty)\} \\ \leqslant& d_0 \log p + C_5 n\bar{\epsilon}_n^2 (\log n)^{1+d_0}.\end{aligned} \tag{A.17}$$

By (2.16), we can also bound the complementary probability as:

$$\begin{aligned}P(W_A^B \notin \mathcal{B}_n) =& \sum_{b\in\{0,1\}^p:|b|\leqslant d_0} P(W_A^b \notin \mathcal{B}_n | B=b)P(B=b) \\ \leqslant& \sum_{b\in\{0,1\}^p:|b|\leqslant d_0} P(W_A^b \notin \mathcal{B}_n^b)P(B=b) \\ \leqslant& \sum_{b\in\{0,1\}^p:|b|\leqslant d_0} e^{-n\bar{\epsilon}_n^2}P(B=b) = e^{-n\bar{\epsilon}_n^2}.\end{aligned} \tag{A.18}$$

Finally, we calculate a lower bound for the prior concentration by (2.12):

$$\begin{aligned}P(||W_A^B - \omega_0||_\infty \leqslant \rho_n) \geqslant& P(B=b_0)P(||W_A^{b_0} - \omega_0||_\infty \leqslant \rho_n | B=b_0) \\ \geqslant& \left(\frac{1}{p}\right)^{|b_0|}\left(1-\frac{1}{p}\right)^{p-|b_0|} e^{-n\rho_n^2} \geqslant e^{-n\rho_n^2 - |b_0|\log p - 2},\end{aligned} \tag{A.19}$$

for n sufficiently large, where $\rho_n = L_0^{|b_0|/(2\alpha_0+|b_0|)} n^{-\alpha_0/(2\alpha_0+|b_0|)} (\log n)^{\gamma}$ with $\gamma = (1+|b_0|)/(2+|b_0|/\alpha_0)$ and the last inequality holds because $(1-1/p)^{p-|b_0|} \to e^{-1}$ as $p \to \infty$.

By choosing $\bar{\epsilon}_n$ equal to a large multiple of $\rho_n + \sqrt{d_0 \log p/n}$ in (A.17)-(A.19), (2.9)-(2.11) hold with ϵ_n a large multiple of $\rho_n (\log n)^{(1+d_0)/2} + \sqrt{d_0 \log p/n}\ (\log n)^{(1+d_0)/2}$ or $L_0^{|b_0|/(2\alpha_0+|b_0|)}\ n^{-\alpha_0/(2\alpha_0+|b_0|)} (\log n)^{\beta_1} + \sqrt{d_0 \log p/n} (\log n)^{\beta_2}$, where $\beta_1 = (1+|b_0|)/(2+|b_0|/\alpha_0) + (1+d_0)/2$ and $\beta_2 = (1+d_0)/2$.

A.1.12 Proof of Theorem 12

To study the the posterior contraction rate of AGPVS prior, we again utilize the flexibility in choosing $\bar{\epsilon}_n$ in the sieve constructions in (2.16). Note that conditioning on K, each component has identical GPVS prior. So we can use the proof of Theorem 11.

Let $(\mathcal{B}_n : n \geqslant 1)$ be the sequence of sieves constructed as (A.16) associated with $\bar{\epsilon}_n$, where the sequence $(\bar{\epsilon}_n : n \geqslant 1)$ will be determined later. We construct the sieves $(\mathcal{F}_n : n \geqslant 1)$ for the additive GP models as

$$\mathcal{F}_n = \bigcup_{k \leqslant K_0} \mathcal{F}_n^k, \quad \mathcal{F}_n^k = \bigoplus_{h=1}^{k} \mathcal{B}_n = \{ f = \sum_{h=1}^{k} f_h : \omega_h \in \mathcal{B}_n \}.$$

$\mathcal{F}_n^k$ can be viewed as all functions that can be decomposed into a sum of k functions in $\mathcal{B}_n$ and $\mathcal{F}_n$ functions a sum of at most K_0. Since $N(kL\epsilon_n, \mathcal{F}_n^k, ||\cdot||_\infty) \leqslant N^k(L\epsilon_n, \mathcal{B}_n, ||\cdot||_\infty)$ and $N(K_0 L\epsilon_n, \mathcal{F}_n, ||\cdot||_\infty) \leqslant \sum_{k=0}^{K_0} N(kL\epsilon_n, \mathcal{F}_n^k, ||\cdot||_\infty)$, we have $N(K_0 L\epsilon_n, \mathcal{F}_n, ||\cdot||_\infty) \leqslant \sum_{k=0}^{K_0} N^k(L\epsilon_n, \mathcal{B}_n, ||\cdot||_\infty) \leqslant N^{K_0}(L\epsilon_n, \mathcal{B}_n, ||\cdot||_\infty)$.

Combine this with (A.17) in the proof of Theorem 11, we obtain

$$\log N(K_0 L\epsilon_n, \mathcal{F}_n, ||\cdot||_\infty) \leqslant K_0 d_0 \log p + K_0 C_5 n \bar{\epsilon}_n^2 (\log n)^{1+d_0}. \tag{A.20}$$

By (A.18), we can bound the complementary probability as

$$\begin{aligned} P(W \notin \mathcal{F}_n) &\leqslant \sum_{k=0}^{K_0} P(W \notin \mathcal{F}_n^k | K=k) P(K=k) \\ &\leqslant \sum_{k=0}^{K_0} \sum_{h=1}^{k} P(W_h^{B_h} \notin \mathcal{B}_n) P(K=k) \\ &\leqslant \sum_{k=0}^{K_0} e^{-n\bar{\epsilon}_n^2} k P(K=k) = EK e^{-n\bar{\epsilon}_n^2}, \end{aligned} \tag{A.21}$$

where EK is the expectation of K.

Finally, by (A.19) the prior concentration can be lower bounded as:

$$\begin{aligned} & P\bigg(||W-\omega_0||_\infty \leqslant \sum_{h=1}^{k_0} \rho_{n,h} \bigg) \\ \geqslant\ & P(K=k_0) P\bigg(|| \sum_{h=1}^{k_0} W_h^{B_h} - \sum_{h=1}^{k_0} \omega_{0,h} ||_\infty \leqslant \sum_{h=1}^{k_0} \rho_{n,h} \bigg) \\ \geqslant\ & p_{k_0} \prod_{h_1=1}^{k_0} P(||W_h^{B_h} - \omega_{0,h}||_\infty \leqslant \rho_{n,h}) \\ \geqslant\ & \exp\bigg\{ -n \sum_{h=1}^{k_0} \rho_{n,h}^2 - \sum_{h=1}^{k_0} |b_{0,h}| \log p - 2k_0 + \log p_{k_0} \bigg\}, \end{aligned} \tag{A.22}$$

for sufficiently large n, where

$$\rho_{n,h} = L_{0,h}^{|b_{0,h}|/(2\alpha_0+|b_{0,h}|)} n^{-\alpha_{0,h}/(2\alpha_{0,h}+|b_{0,h}|)} (\log n)^{(1+|b_{0,h}|)/(2+|b_{0,h}|/\alpha_{0,h})}.$$

By choosing $\bar{\epsilon}_n$ equal to a large multiple of $\sum_{h=1}^{k_0} \rho_{n,h} + \sqrt{K_0 d_0 \log p/n}$ in (A.20)-(A.22), (2.9)-(2.11) will hold with ϵ_n a large multiple of

$$\sqrt{K_0} \sum_{h=1}^{k_0} \rho_{n,h} (\log n)^{(1+d_0)/2} + \sqrt{K_0} \sqrt{K_0 d_0 \log p/n} (\log n)^{(1+d_0)/2}$$

or

$$\sqrt{K_0}\sum_{h=1}^{k_0} L_{0,h}^{|b_{0,h}|/(2\alpha_0+|b_{0,h}|)} n^{-\alpha_{0,h}/(2\alpha_{0,h}+|b_{0,h}|)} (\log n)^{\beta_{1,h}} + \sqrt{K_0}\sqrt{K_0 d_0 \log p/n}(\log n)^{\beta_2},$$

where $\beta_{1,h} = (1 + |b_{0,h}|)/(2 + |b_{0,h}|/\alpha_{0,h}) + (1 + d_0)/2$ and $\beta_2 = (1 + d_0)/2$.

Appendix B

Appendix for Chapter 3

B.1 Geometric properties

We introduce some concepts and results in differential and Riemannian geometry, which play an important role in the convergence rate. For detailed definitions and notations, the reader is referred to do Carmo (1992).

B.1.1 Riemannian manifold

A manifold is a topological space that locally resembles Euclidean space. A d-dimensional topological manifold $\mathcal{M}$ can be described using an atlas, where an atlas is defined as a collection $\{(U_s, \phi_s)\}$ such that $\mathcal{M} = \bigcup_s U_s =$ and each chart $\phi_s : V \to U_s$ is a homeomorphism from an open subset V of d-dimensional Euclidean space to an open subset U_s of $\mathcal{M}$. By constructing an atlas whose transition functions $\{\tau_{s,\beta} = \phi_\beta^{-1} \circ \phi_s\}$ are C^γ differentiable, we can further introduce a differentiable structure on $\mathcal{M}$. With this differentiable structure, we are able to define differentiable functions and their smoothness level $s \leqslant \gamma$. Moreover, this additional structure allows us to extend Euclidean differential calculus to the manifold. To

measure distances and angles on a manifold, the notion of Riemannian manifold is introduced. A Riemannian manifold $(\mathcal{M}, g)$ is a differentiable manifold $\mathcal{M}$ in which each tangent space $T_p\mathcal{M}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_p = g_p(\cdot, \cdot)$ that varies smoothly in p. The family g_p of inner products is called a Riemannian metric and is denoted by g. With this Riemannian metric g, a distance $d_{\mathcal{M}}(p, q)$ between any two points $p, q \in \mathcal{M}$ can be defined as the length of the shortest path on $\mathcal{M}$ connecting them. For a given manifold $\mathcal{M}$, such as the set $P(n)$ of all $n \times n$ positive symmetric matrices (Moakher and Zéraï, 2011; Hiai and Petz, 2009), a Riemannian metric g is not uniquely determined and can be constructed in various manners so that certain desirable properties, such as transformation or group action invariability, are valid. Although g is not uniquely determined, the smoothness of a given function f on $\mathcal{M}$ only depends on $\mathcal{M}$'s differential structure instead of its Riemannian metric. Therefore, to study functions on the manifold $\mathcal{M}$, we could endow it with any valid Riemannian metric. Since a low dimensional manifold structure on the $\mathbb{R}^D$-valued predictor X is assumed in this paper, we will focus on the case in which $\mathcal{M}$ is a submanifold of a Euclidean space.

Definition 63. *$\mathcal{M}$ is called a C^γ submanifold of $\mathbb{R}^D$ if there exists an inclusion map $\Phi : \mathcal{M} \mapsto \mathbb{R}^D$, called embedding, such that Φ is a diffeomorphism between $\mathcal{M}$ and $\Phi(\mathcal{M}) \subset \mathbb{R}^D$, which means:*

(1) Φ is injective and γ-differentiable;

(2) The inverse $\Phi^{-1} : \Phi(\mathcal{M}) \to \mathcal{M}$ is also γ-differentiable.

A natural choice of the Riemannian metric g of $\mathcal{M}$ is the one induced by the Euclidean metric e of $\mathbb{R}^D$ through

$$g_p(u, v) = e_{\Phi(p)}(d\Phi_p(u), d\Phi_p(v)) = \langle d\Phi_p(u), d\Phi_p(v) \rangle_{\mathbb{R}^D}, \quad \forall u, v \in T_p\mathcal{M},$$

for any $p \in \mathcal{M}$. Under this Riemannian metric g, $d\Phi_p : T_p\mathcal{M} \mapsto d\Phi_p(T_p\mathcal{M}) \subset T_{\Phi(p)}\mathbb{R}^D$ is an isometric embedding. Nash Embedding Theorem (Nash, 1956) implies that any valid Riemannian metric on $\mathcal{M}$ could be considered as being induced from a Euclidean metric of $\mathbb{R}^m$ with a sufficiently large m. Therefore, we would use this naturally induced g as the Riemannian metric of predictor manifold $\mathcal{M}$ when studying the posterior contraction rate of our proposed GP prior defined on this manifold. Under such choice of g, $\mathcal{M}$ is isometrically embedded in the ambient space $\mathbb{R}^D$. In addition, in the rest of this paper, we will occasionally identify $\mathcal{M}$ with $\Phi(\mathcal{M})$ when no confusion arises.

Tangent spaces and Riemannian metric can be represented in terms of local parameterizations. Let $\phi : U \mapsto \mathcal{M}$ be a chart that maps a neighborhood U of the origin in $\mathbb{R}^d$ to a neighborhood $\phi(U)$ of $p \in \mathcal{M}$. In the case that $\mathcal{M}$ is a C^γ submanifold of $\mathbb{R}^D$, ϕ itself is γ-differentiable as a function from $U \in \mathbb{R}^d$ to $\mathbb{R}^D$. Given $i \in \{1, \ldots, d\}$ and $q = \phi(u)$, where $u = (u_1, \ldots, u_d) \in U$, define $\frac{\partial}{\partial u_i}(q)$ to be the linear functional on $C^\gamma(\mathcal{M})$ such that

$$\frac{\partial}{\partial u_i}(q)(f) = \frac{d(f \circ \phi(u + te_i))}{dt}\bigg|_{t=0}, \ \forall f \in C^\gamma(\mathcal{M}),$$

where the d-dimensional vector e_i has 1 in the i-th component and 0's in others. Then $\frac{\partial}{\partial u_i}(q)$ can be viewed as a tangent vector in the tangent space $T_q\mathcal{M}$. Moreover, $\{\frac{\partial}{\partial u_i}(q) : 1 \leqslant i \leqslant d\}$ forms a basis of $T_q\mathcal{M}$ so that each tangent vector $v \in T_q\mathcal{M}$ can written as

$$v = \sum_{i=1}^{d} v_i \frac{\partial}{\partial u_i}(q).$$

Under this basis, the tangent space of $\mathcal{M}$ can be identified as $\mathbb{R}^d$ and the matrix representation of differential $d\Phi_q$ at q has a (j, i)th element given by

$$\left\{ d\Phi_q\left(\frac{\partial}{\partial u_i}\right)\right\}_j = \frac{d(\Phi_j \circ \phi(u + te_i))}{dt}\bigg|_{t=0}, \ i = 1, \ldots, d, \ j = 1, \ldots, D,$$

where we use the notation F_j to denote the jth component of a vector-valued function F. In addition, under the same basis, the Riemannian metric g_q at q can be expressed as

$$g_q(v, w) = \sum_{i,j=1}^{d} v_i w_j g_{ij}^{\phi}(u_1, \ldots, u_d),$$

where $(v_1, \ldots, v_d)$ and $(w_1, \ldots, w_d)$ are the local coordinates for $v, w \in T_q\mathcal{M}$. By the isometry assumption,

$$g_{ij}^{\phi}(u_1, \ldots, u_d) = \langle d\Phi_q(\frac{\partial}{\partial u_i}), d\Phi_q(\frac{\partial}{\partial u_j})\rangle_{R^D}.$$

Riemannian volume measure (form) of a region R contained in a coordinate neighborhood $\phi(U)$ is defined as

$$\mathrm{Vol}(R) = \int_R dV(q) \triangleq \int_{\phi^{-1}(R)} \sqrt{\det(g_{ij}^{\phi}(u))} du_1 \ldots du_d.$$

The volume of a general compact region R, which is not contained in a coordinate neighborhood, can be defined through partition of unity (do Carmo, 1992). Vol generalizes the Lebesque measure of Euclidean spaces and can be used to define the integral of a function $f \in C(\mathcal{M})$ as $\int_{\mathcal{M}} f(q)dV(q)$. In the special case that f is supported on a coordinate neighborhood $\phi(U)$,

$$\int_{\mathcal{M}} f(q)dV(q) = \int_U f(\phi(u))\sqrt{\det(g_{ij}^{\phi}(u))} du_1 \ldots du_d. \tag{B.1}$$

B.1.2 Exponential map

Geodesic curves, generalizations of straight lines from Euclidean spaces to curved spaces, are defined as those curves whose tangent vectors remain parallel if they are transported and are locally the shortest path between points on the manifold. Formally, for $p \in \mathcal{M}$ and $v \in T_p\mathcal{M}$, the geodesic $\gamma(t, p, v), t > 0$, starting at p

with velocity v, i.e. $\gamma(0,p,v) = p$ and $\gamma'(t,p,v) = v$, can be found as the unique solution of an ordinary differential equation. The exponential map $\mathcal{E}_p : T_p\mathcal{M} \mapsto \mathcal{M}$ is defined by $\mathcal{E}_p(v) = \gamma(1,p,v)$ for any $v \in T_p\mathcal{M}$ and $p \in \mathcal{M}$. Under this special local parameterization, calculations can be considerably simplified since quantities such as $\mathcal{E}_p$'s differential and Riemannian metric would have simple forms.

Although Hopf-Rinow theorem ensures that for compact manifolds the exponential map $\mathcal{E}_p$ at any point p can be defined on the entire tangent space $T_p\mathcal{M}$, generally this map is no longer a global diffeomorphism. Therefore to ensure good properties of this exponential map, the notion of a normal neighborhood is introduced as follows.

Definition 64. *A neighborhood V of $p \in \mathcal{M}$ is called normal if:*

(1) Every point $q \in V$ can be joined to p by a unique geodesic $\gamma(t,p,v), 0 \leqslant t \leqslant 1$, with $\gamma(0,p,v) = p$ and $\gamma(1,p,v) = q$;

(2) $\mathcal{E}_p$ is a diffeomorphism between V and a neighborhood of the origin in $T_p\mathcal{M}$.

Proposition 2.7 and 3.6 in do Carmo (1992) ensure that every point in $\mathcal{M}$ has a normal neighborhood. However, if we want to study some properties that hold uniformly for all exponential maps $\mathcal{E}_q$ with q in a small neighborhood of p, we need a notion stronger than normal neighborhood, whose existence has been established in Theorem 3.7 in do Carmo (1992).

Definition 65. *A neighborhood W of $p \in \mathcal{M}$ is called uniformly normal if there exists some $\delta > 0$ such that:*

(1) For every $q \in W$, $\mathcal{E}_p$ is defined on the δ-ball $B_\delta(0) \subset T_q\mathcal{M}$ around the origin of $T_q\mathcal{M}$. Moreover, $\mathcal{E}_p(B_\delta(0))$ is a normal neighborhood of q;

(2) $W \subset \mathcal{E}_p(B_\delta(0))$, which implies that W is a normal neighborhood of all its points.

Moreover, as pointed out by Gine and Koltchinskii (2005) and Ye and Zhou (2008), by shrinking W and reducing δ at the same time, a special uniformly normal neighborhood can be chosen.

Proposition 66. *For every $p \in \mathcal{M}$ there exists a neighborhood W such that:*

(1) W is a uniformly normal neighborhood of p with some $\delta > 0$;

(2) The closure of W is contained in a strongly convex neighborhood U of p;

(3) The function $F(q, v) = (q, \mathcal{E}_q(v))$ is a diffeomorphism from $W_\delta = W \times B_\delta(0)$ onto its image in $\mathcal{M} \times \mathcal{M}$. Moreover, $|dF|$ is bounded away from zero on W_δ.

Here U is strongly convex if for every two points in U, the minimizing geodesic joining them also lies in U.

Throughout the rest of the paper, we will assume that the uniformly normal neighborhoods also possess the properties in the above proposition. Given a point $p \in \mathcal{M}$, we choose a uniformly normal neighborhood W of p. Let $\{e_1, \ldots, e_d\}$ be an orthonormal basis of $T_p\mathcal{M}$. For each $q \in W$, we can define a set of tangent vectors $\{e_1^q, \ldots, e_d^q\} \subset T_q\mathcal{M}$ by parallel transport (do Carmo, 1992): $e_i \in T_p\mathcal{M} \mapsto e_i^{\gamma(t)} \in T_{\gamma(t)}\mathcal{M}$ from p to q along the unique minimizing geodesic $\gamma(t)$ $(0 \leqslant t \leqslant 1)$ with $\gamma(0) = p, \gamma(1) = q$. Since parallel transport preserves the inner product in the sense that $g_{\gamma(t)}(v^{\gamma(t)}, w^{\gamma(t)}) = g_p(v, w), \forall v, w \in T_p\mathcal{M}$, $\{e_1^q, \ldots, e_d^q\}$ forms an orthonormal basis of $T_q\mathcal{M}$. In addition, the orthonormal frame defined in this way is unique and depends smoothly on q. Therefore, we obtain on W a system of normal coordinates at each $q \in W$, which parameterizes $x \in \mathcal{E}_q(B_\delta(0))$ by

$$x = \mathcal{E}_q\left(\sum_{i=1}^{d} u_i e_i^q\right) = \phi^q(u_1, \ldots, u_d),\ u = (u_1, \ldots, u_d) \in B_\delta(0). \tag{B.2}$$

Such coordinates are called q-normal coordinates. The basis of $T_q\mathcal{M}$ associated with this coordinate chart $(B_\delta(0), \phi^q)$ is given by

$$\frac{\partial}{\partial u_i}(q)(f) = \left.\frac{d(f \circ \mathcal{E}_q(te_i^q))}{dt}\right|_{t=0} = \left.\frac{d(f \circ \gamma(t, q, e_i^q))}{dt}\right|_{t=0} = e_i^q(f),\ i = 1, \ldots, d.$$

Therefore $\{\frac{\partial}{\partial u_i}(q) = e_i^q : 1 \leqslant i \leqslant d\}$ forms an orthonormal basis on $T_q\mathcal{M}$. By Proposition 66, for each $x \in \mathcal{E}_q(B_\delta(0))$, there exists a minimizing geodesic $\gamma(t, q, v), 0 \leqslant t \leqslant 1$, such that $\gamma(0, q, v) = q, \gamma'(0, q, v) = v$ and $\gamma(1, q, v) = x$, where $v = \mathcal{E}_q^{-1}(x) = \sum_{i=1}^d u_i e_i^q \in T_q\mathcal{M}$. Hence $d_\mathcal{M}(q, x) = \int_0^1 |\gamma'(t, q, v)|dt = |v| = ||u||$, i.e.

$$d_\mathcal{M}\left(q, \mathcal{E}_q\left(\sum_{i=1}^d u_i e_i^q\right)\right) = ||u||, \quad \forall u \in B_{\delta_p}(0), \tag{B.3}$$

where $||\cdot||$ is the Euclidean norm on $\mathbb{R}^d$. The components $g_{ij}^q(u)$ of the Riemannian metric in q-normal coordinates satisfy $g_{ij}^q(0) = g_q(e_i^q, e_j^q) = \delta_{ij}$. The following results (Gine and Koltchinskii, 2005, Proposition 2.2) provide local expansions for the Riemannian metric $g_{ij}^q(u)$, the Jacobian $\sqrt{\det(g_{ij}^\phi(u))}$ and the distance $d_\mathcal{M}(q, \sum_{i=1}^d u_i e_i^q)$ in a neighborhood of p.

Proposition 67. *Let $\mathcal{M}$ be a submanifold of $\mathbb{R}^D$ which is isometrically embedded. Given a point $p \in \mathcal{M}$, let W and δ be as in Proposition 66, and consider for each $q \in W$ the q-normal coordinates defined above. Suppose that $x = \sum_{i=1}^d u_i e_i^q \in \mathcal{E}_q(B_\delta(0))$. Then:*

(1) The components $g_{ij}^q(u)$ of the metric tensor in q-normal coordinates admit the following expansion, uniformly in $q \in W$ and $x \in \mathcal{E}_q(B_\delta(0))$:

$$g_{ij}^q(u_1, \ldots, u_d) = \delta_{ij} - \frac{1}{3}\sum_{r,s=1}^d R_{irsj}^q(0)u_r u_s + O(d_\mathcal{M}^3(q, x)), \tag{B.4}$$

where $R_{irsj}^q(0)$ are the components of the curvature tensor at q in q-normal coordinates.

(2) The Jacobian $\sqrt{\det(g_{ij}^q)}(u)$ *in q-normal coordinates has the following expansion, uniformly in* $q \in W$ *and* $x \in \mathcal{E}_q(B_\delta(0))$*:*

$$\sqrt{\det(g_{ij}^q)}(u_1, \ldots, u_d) = 1 - \frac{1}{6} \sum_{r,s=1}^{d} Ric_{rs}^q(0) u_r u_s + O(d_{\mathcal{M}}^3(q, x)), \tag{B.5}$$

where $Ric_{rs}^q(0)$ *are the components of the Ricci tensor at q in q-normal coordinates.*

(3) There exists $C_p < \infty$ *such that*

$$0 \leqslant d_{\mathcal{M}}^2(q, x) - ||q - x||^2 \leqslant C_p d_{\mathcal{M}}^4(q, x) \tag{B.6}$$

holds uniformly in $q \in W$ *and* $x \in \mathcal{E}_q(B_\delta(0))$*.*

Note that in Proposition 67, (3) only provides a comparison of geodesic distance and Euclidean distance in local neighborhoods. Under a stronger compactness assumption on $\mathcal{M}$, the following lemma offers a global comparison of these two distances.

Lemma 68. *Let* $\mathcal{M}$ *be a connected compact submanifold of* $\mathbb{R}^D$ *with a Riemannian metric g that is not necessarily induced from the Euclidean metric. Then there exist positive constants* C_1 *and* C_2 *dependent on g, such that*

$$C_1 ||x - y|| \leqslant d_{\mathcal{M}}(x, y) \leqslant C_2 ||x - y||, \quad \forall x, y \in \mathcal{M}, \tag{B.7}$$

where $||\cdot||$ *is the Euclidean distance in* $\mathbb{R}^D$*. Moreover, if* $\mathcal{M}$ *is further assumed to be isometrically embedded, i.e. g is induced from the Euclidean metric of* $\mathbb{R}^D$*, then* C_1 *could be chosen to be one and* $C_2 \geqslant 1$*.*

Proof. We only prove the first half of the inequality since the second half follows by a similar argument and is omitted here. Assume in the contrary that for any positive

integer k, there exists (x_k, y_k) such that $||x_k - y_k|| \geqslant k d_{\mathcal{M}}(x_k, y_k)$. Let $\Phi : \mathcal{M} \to \mathbb{R}^D$ be the embedding. Since $\mathcal{M}$ is compact, $\{x_k\}$ and $\{y_k\}$ have convergent subsequences, whose notations are abused as $\{x_k\}$ and $\{y_k\}$ for simplicity. Denote the limits of these two sequences as x_0 and y_0. By the compactness of $\mathcal{M}$ and continuity of Φ, we know that $\Phi(\mathcal{M})$ is also compact and therefore $d_{\mathcal{M}}(x_k, y_k) \to 0$, as $k \to \infty$. This implies that $x_0 = y_0 = p$.

For each $j \in \{1, \ldots, p\}$, the jth component $\Phi_j : \mathcal{M} \to \mathbb{R}$ of Φ is differentiable. Let δ_p and W_p be the δ and W specified in Proposition 66. Define $f(q, v) = \Phi\big(\pi_2(F(q, v))\big) = \Phi(\mathcal{E}_p(v))$, where π_2 is the projection of $\mathcal{M} \times \mathcal{M}$ on to its second component. By Proposition 66, f is differentiable on the compact set $\bar{W}_{\delta_p}$ and therefore for each $l \in \{1, \ldots, d\}$, $\frac{\partial f}{\partial v_l}$ is uniformly bounded on $\bar{W}_{\delta_p}$. This implies that for some constant $C > 0$, $||x - y|| = ||f(y, \mathcal{E}_y^{-1}(x)) - f(y, \mathcal{E}_y^{-1}(y))|| \leqslant C||\mathcal{E}_y^{-1}(x) - \mathcal{E}_y^{-1}(y)|| = C d_{\mathcal{M}}(x, y)$ for all $x, y \in W_p$ with $d_{\mathcal{M}}(x, y) \leqslant \delta_p$. Since $x_k \to p$ and $y_k \to p$, there exists an integer k_0 such that for all $k > k_0$, x_k, $y_k \in W_p$ and $d_{\mathcal{M}}(x_k, y_k) \leqslant \delta_p$. Therefore $||x_k - y_k|| \leqslant C d_{\mathcal{M}}(x_k, y_k)$, which contradict our assumption that $||x_k - y_k|| \geqslant k d_{\mathcal{M}}(x_k, y_k)$ for all k.

Consider the case when Φ is an isometric embedding. For any points $x, y \in \mathcal{M}$, we can cover the compact geodesic path $l_{x,y}$ from x to y by $\{W_{p_i} : i = 1, \ldots, n\}$ associated with a finite number of points $\{p_1, \ldots, p_n\} \subset \mathcal{M}$. Therefore we can divide $l_{x,y}$ into $\bigcup_{s=1}^{n} l(x_{s-1}, x_s)$ such that $x_0 = x$, $x_n = y$, and each segment $l(x_{s-1}, x_s)$ lies in one of the W_{p_i}'s. By Proposition 67 (3), for each $s \in \{1, \ldots, n\}$, $d_{\mathcal{M}}(x_{s-1}, x_s) \geqslant ||x_{s-1} - x_s||$. Therefore,

$$d_{\mathcal{M}}(x, y) = \sum_{s=1}^{n} d_{\mathcal{M}}(x_{s-1}, x_s) \geqslant \sum_{s=1}^{n} ||x_{s-1} - x_s|| \geqslant ||x - y||,$$

where the last step follows from the triangle inequality. □

The above lemma also implies that geodesic distances induced by different Rie-

mannian metrics on $\mathcal{M}$ are equivalent to each other.

Fix $p \in \mathcal{M}$ and let W and $\delta > 0$ be specified as in Proposition 66. Since $\mathcal{M}$ is a submanifold of $\mathbb{R}^D$, for any point $q \in \mathcal{M}$, the exponential map $\mathcal{E}_q : B_\delta(0) \to \mathcal{M} \subset R^D$ is a differentiable function between two subsets of Euclidean spaces. Here, we can choose any orthonormal basis of $T_q\mathcal{M}$ since the representations of $\mathcal{E}_q$ under different orthonormal bases are the same up to $d \times d$ rotation matrices. Under the compactness assumption on $\mathcal{M}$, the following lemma, which will be applied in the proof of lemma 73, ensures the existence of a bound on the partial derivatives of $\mathcal{E}_q$'s components $\{\mathcal{E}_{q,i} : i = 1, \ldots, D\}$ uniformly for all q in the δ neighborhood of p:

Lemma 69. *Let $\mathcal{M}$ be a connected C^γ compact submanifold of $\mathbb{R}^D$ with γ being ∞ or any integer greater than two. Let k be an integer such that $k \leqslant \gamma$. Then:*

1. *There exists a universal positive number δ_0, such that for every $p \in \mathcal{M}$, proposition 66 is satisfied with some $\delta > \delta_0$ and W_p;*
2. *With this δ_0, for any $p \in \mathcal{M}$, mixed partial derivatives with order less than or equal to k of each component of $\mathcal{E}_p$ are bounded in $B_{\delta_0}(0) \in T_p\mathcal{M}$ by a universal constant $C > 0$.*

Proof. Note that $\mathcal{M} = \bigcup_{p \in \mathcal{M}} W(p, \delta_p)$, where δ_p and $W(p, \delta_p)$ are the corresponding p dependent δ and open neighborhood W in proposition 66. By the compactness of $\mathcal{M}$, we can choose a finite covering $\{W(p_1, \delta_{p_1}), \ldots, W(p_n, \delta_{p_n})\}$. Let $\delta_0 = \min\{\delta_{p_1}, \ldots, \delta_{p_n}\}$. Then the first condition is satisfied with this δ_0 since for any $p \in \mathcal{M}$, W_p could be chosen as any $W(p_j, \delta_{p_j})$ that contains p.

Next we prove the second condition. For each j, we can define q-normal coordinates on $W(p_j, \delta_{p_j})$ as before such that the exponential map at each point $q \in W(p_j, \delta_{p_j})$ can be parameterized as (B.2). Define $F_j : W(p_j, \delta_{p_j}) \times B_{\delta_{p_j}}(0) \to \mathbb{R}^D$ by $F_j(q, u) = \mathcal{E}_q(\sum_{i=1}^d u_i e_i^q) = \phi^q(u)$. Then any order k mixed partial derivative $\frac{\partial^k \phi_j^q}{\partial u_{i1} \cdots \partial u_{ik}}(u)$ of $F_j(q, u)$ with respect to u is continuous on the compact set

$W(p_j, \delta_{p_j}) \times B_{\delta_{p_j}}(0)$. Therefore these partial derivatives are bounded uniformly in $q \in W(p_j, \delta_{p_j})$ and $u \in B_{\delta_{p_j}}(0)$. Since $\mathcal{M}$ is covered by a finite number of sets $\{W(p_1, \delta_{p_1}), \ldots, W(p_n, \delta_{p_n})\}$, the second conclusion is also true. □

By lemma 69, when a compact submanifold $\mathcal{M}$ has smoothness level greater than or equal to k, we can approximate the exponential map $\mathcal{E}_p : B_{\delta_0}(0) \subset \mathbb{R}^d \to \mathbb{R}^D$ at any point $p \in \mathcal{M}$ by a local Taylor polynomial of order k with error bound $C\delta_0^k$, where C is a universal constant that only depends on k and $\mathcal{M}$.

B.2 Posterior contraction rate of the GP on manifold

In the GP prior (3.4), the covariance function $K^a : \mathcal{M} \times \mathcal{M} \to R$ is essentially defined on the submanifold $\mathcal{M}$. Therefore, (3.4) actually defines a GP on $\mathcal{M}$ and we can study its posterior contraction rate as a prior for functions on the manifold. In this section, we combine geometry properties and Bayesian nonparametric asymptotic theory to prove the theorems in section 2.

B.2.1 Reproducing kernel Hilbert space on manifold

Being viewed as a covariance function defined on $[0,1]^D \times [0,1]^D$, $K^a(\cdot,\cdot)$ corresponds to a reproducing kernel Hilbert space (RKHS) $\mathbb{H}^a$, which is defined as the completion of $\mathcal{H}$, the linear space of all functions on $[0,1]^D$ with the following form

$$x \mapsto \sum_{i=1}^{m} a_i K^a(x_i, x), x \in [0,1]^D,$$

indexed by $a_1, \ldots, a_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in [0,1]^D$, $m \in \mathbb{N}$, relative to the norm induced by the inner product defined through $\langle K^a(x,\cdot), K^a(y,\cdot)\rangle_{\mathbb{H}^a} = K^a(x,y)$. Similarly, $K^a(\cdot,\cdot)$ can also be viewed as a convariance function defined on $\mathcal{M} \times \mathcal{M}$, with the associated RKHS denoted by $\tilde{\mathbb{H}}^a$. Here $\tilde{\mathbb{H}}^a$ is the completion of $\tilde{\mathcal{H}}$, which is the

linear space of all functions on $\mathcal{M}$ with the following form

$$x \mapsto \sum_{i=1}^{m} a_i K^a(x_i, x), x \in \mathcal{M},$$

indexed by $a_1, \ldots, a_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in \mathcal{M},\ m \in \mathbb{N}$.

Many probabilistic properties of GPs are closely related to the RKHS associated with its covariance function. Readers can refer to Aronszajn (1950) and van der Vaart and van Zanten (2008b) for introductions on RKHS theory for GPs on Euclidean spaces. In order to generalize RKHS properties in Euclidean spaces to submanifolds, we need a link to transfer the theory. The next lemma achieves this by characterizing the relationship between $\mathbb{H}^a$ and $\tilde{\mathbb{H}}^a$.

Lemma 70. *For any $f \in \tilde{\mathbb{H}}^a$, there exists $g \in \mathbb{H}^a$ such that $g|_{\mathcal{M}} = f$ and $||g||_{\mathbb{H}^a} = ||f||_{\tilde{\mathbb{H}}^a}$, where $g|_{\mathcal{M}}$ is the restriction of g on $\mathcal{M}$. Moreover, for any other $g' \in \mathbb{H}^a$ with $g'|_{\mathcal{M}} = f$, we have $||g'||_{\mathbb{H}^a} \geqslant ||f||_{\tilde{\mathbb{H}}^a}$, which implies $||f||_{\tilde{\mathbb{H}}^a} = \inf_{g \in \mathbb{H}^a, g|_{\mathcal{M}} = f} ||g||_{\mathbb{H}^a}$.*

Proof. Consider the map $\Phi : \tilde{\mathcal{H}} \to \mathcal{H}$ that maps the function

$$\sum_{i=1}^{m} a_i K^a(x_i, \cdot) \in \tilde{\mathcal{H}},\ a_1, \ldots, a_m \in \mathbb{R}, x_1, \ldots, x_m \in \mathcal{M}, m \in \mathbb{N}$$

on $\mathcal{M}$ to the function of the same form

$$\sum_{i=1}^{m} a_i K^a(x_i, \cdot) \in \mathcal{H},$$

but viewed as a function on $[0,1]^D$. By definitions of RKHS norms, Φ is an isometry between $\tilde{\mathcal{H}}$ and a linear subspace of $\mathcal{H}$. As a result, Φ can be extended to an isometry between $\tilde{\mathbb{H}}^a$ and a complete subspace of $\mathbb{H}^a$. To prove the first part of this lemma, it suffices to justify that for any $f \in \tilde{\mathbb{H}}^a$, $g = \Phi(f)|_{\mathcal{M}} = f$. Assume that the sequence $\{f_n\} \in \tilde{\mathcal{H}}$ satisfies

$$||f_n - f||_{\tilde{\mathbb{H}}^a} \to 0,\ \text{as } n \to \infty,$$

then by the definition of Φ on $\tilde{\mathcal{H}}$, $\Phi(f_n)|_{\mathcal{M}} = f_n$. For any $x \in [0,1]^D$, by the reproducing property and Cauchy-Schwarz inequality,

$$\begin{aligned} |\Phi(f_n)(x) - g(x)| &= |\langle K^a(x,\cdot), \Phi(f_n) - g\rangle_{\mathbb{H}^a}| \\ &\leqslant \sqrt{K^a(x,x)}\ ||\Phi(f_n) - \Phi(f)||_{\mathbb{H}^a} \\ &= ||f_n - f||_{\tilde{\mathbb{H}}^a} \to 0, \text{ as } n \to \infty, \end{aligned}$$

where the last step is by isometry. This indicates that g can be obtained as a point limit of $\Phi(f_n)$ on $[0,1]^D$ and in the special case when $x \in \mathcal{M}$,

$$g(x) = \lim_{n\to\infty} \Phi(f_n)(x) = \lim_{n\to\infty} f_n(x) = f(x).$$

Denote the orthogonal complement of $\Phi(\tilde{\mathbb{H}}^a)$ in $\mathbb{H}^a$ as $\Phi(\tilde{\mathbb{H}}^a)^\perp$. Since $(g' - g)|_{\mathcal{M}} = 0$, which means $\langle K^a(x,\cdot), g - g'\rangle_{\mathbb{H}^a} = 0$ for all $x \in \mathcal{M}$. Therefore by the previous construction, $g - g' \perp \Phi(\tilde{\mathbb{H}}^a)$, i.e. $g' - g \in \Phi(\tilde{\mathbb{H}}^a)^\perp$ and using Pythagorean theorem, we have

$$||g'||_{\mathbb{H}^a} = ||g||_{\mathbb{H}^a} + ||g - g'||_{\mathbb{H}^a} \geqslant ||g||_{\mathbb{H}^a}.$$

□

This lemma implies that any element f in the RKHS $\tilde{\mathbb{H}}^a$ could be considered as the restriction of some element g in the RKHS $\mathbb{H}^a$. Particularly, there exists a unique such element g in $\mathbb{H}^a$ such that the norm is preserved, i.e. $||g||_{\mathbb{H}^a} = ||f||_{\tilde{\mathbb{H}}^a}$.

B.2.2 Background on posterior convergence rate for GP

As shown in Ghosal et al. (2000), in order to characterize the posterior contraction rate in a Bayesian nonparametric problem, such as density estimation, fixed/random design regression or classification, we need to verify some conditions on the prior measure Π. Specifically, we describe the sufficient conditions for randomly rescaled GP prior as (3.3) given in van der Vaart and van Zanten (2009). Let $\mathcal{X}$ be the

predictor space and f_0 be the true function $f_0 : \mathcal{X} \to \mathbb{R}$, which is the log density $\log p(x)$ in density estimation, regression function $E[Y|X]$ in regression and logistic transformed conditional probability $\text{logit}P(Y = 1|X)$ in classification. We will not consider density estimation since to specify the density by log density f_0, we need to know the support $\mathcal{M}$ so that e^{f_0} can be normalized to produce a valid density. Let ϵ_n and $\bar{\epsilon}_n$ be two sequences. If there exist Borel measurable subsets B_n of $C(\mathcal{X})$ and constant $K > 0$ such that for n sufficiently large,

$$\begin{aligned} P(||W^A - f_0||_\infty \leqslant \epsilon_n) &\geqslant e^{-n\epsilon_n^2}, \\ P(W^A \notin B_n) &\leqslant e^{-4n\epsilon_n^2}, \\ \log N(\bar{\epsilon}_n, B_n, ||\cdot||_\infty) &\leqslant n\bar{\epsilon}_n^2, \end{aligned} \tag{B.8}$$

where $W^A \sim \Pi$ and $||\cdot||_\infty$ is the sup-norm on $C(\mathcal{X})$, then the posterior contraction rate would be at least $\epsilon_n \vee \bar{\epsilon}_n$. In our case, $\mathcal{X}$ is the d-dimensioanl submanifold $\mathcal{M}$ in the ambient space $\mathbb{R}^D$. To verify the first concentration condition, we need to give upper bounds to the so-called concentration function (van der Vaart and van Zanten, 2009) $\phi^a_{f_0}(\epsilon)$ of the GP W^a around truth f_0 for given a and ϵ. $\phi^a_{f_0}(\epsilon)$ is composed of two terms. Both terms depend on the RKHS $\tilde{\mathbb{H}}^a$ associated with the covariance function of the GP W^a. The first term is the decentering function $\inf\{||h||^2_{\tilde{\mathbb{H}}^a} : ||h - f_0||_\infty < \epsilon\}$, where $||\cdot||_{\tilde{\mathbb{H}}^a}$ is the RKHS norm. This quantity measures how well the truth f_0 could be approximated by the elements in the RKHS. The second term is the negative log small ball probability $-\log P(||W^a||_\infty < \epsilon)$, which depends on the covering entropy $\log N(\epsilon_n, \tilde{\mathbb{H}}^a_1, ||\cdot||_\infty)$ of the unit ball in the RKHS $\tilde{\mathbb{H}}^a$. As a result of this dependence, by applying Borell's inequality (van der Vaart and van Zanten, 2008b), the second and third conditions can often be proved as byproducts by using the conclusion on the small ball probability .

As pointed out by van der Vaart and van Zanten (2009), the key to ensure the adaptability of the GP prior on Euclidean spaces is a sub-exponential type tail of its

stationary covariance function's spectral density, which is true for squared exponential and Matérn class covariance functions. More specifically, a squared exponential covariance function $K_1(x,y) = \exp\big\{ -||x-y||^2/2\big\}$ on $\mathbb{R}^D$ has a spectral representation as

$$K_1(x,y) = \int_{\mathbb{R}^D} e^{-i(\lambda, x-y)} \mu(d\lambda),$$

where μ is its spectral measure with a sub-Gaussian tail, which is lighter than sub-exponential tail in the sense that: for any $\delta > 0$,

$$\int e^{\delta||\lambda||} \mu(d\lambda) < \infty. \tag{B.9}$$

For convenience, we will focus on squared exponential covariance function, since generalizations to other covariance functions with sub-exponential decaying spectral densities are possible with more elaboration.

B.2.3 Decentering function

To estimate the decentering function, the key step is to construct a function $I_a(f)$ on the manifold $\mathcal{M}$ to approximate a differentiable function f, so that the RKHS norm $||I_a(f)||_{\tilde{\mathbb{H}}^a}$ can be tightly upper bounded. Unlike in Euclidean spaces where functions in the RKHS $\mathbb{H}^a$ can be represented via Fourier transformations (van der Vaart and van Zanten, 2009), there is no general way to represent and calculate RKHS norms of functions in the RKHS $\tilde{\mathbb{H}}^a$ on manifold. Therefore in the next lemma, we provide a direct way to construct the approximation function $I_a(f)$ for any truth f via convolving f with K^a on manifold $\mathcal{M}$:

$$\begin{aligned} I_a(f)(x) &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} K^a(x,y) f(y) dV(y) \\ &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}} \exp\left\{ -\frac{a^2||x-y||^2}{2} \right\} f(y) dV(y), \quad x \in \mathcal{M}, \end{aligned} \tag{B.10}$$

where V is the Riemannian volume form of $\mathcal{M}$. Heuristically, for large a, the above integrand only has non-negligible value in a small neighborhood around x. Therefore we can conduct a change of variable in the above integral with transformation $\phi^x : B_\delta \to W$ defined by (B.2) in a small neighborhood W of x:

$$\begin{aligned} I_a(f)(x) &= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2||\phi^x(u)-\phi^x(0)||^2}{2}\right\} f\big(\phi^x(u)\big)\sqrt{\det(g_{ij}^{\phi}(u))}du, \\ &\approx \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2||u||^2}{2}\right\} f\big(\phi^x(u)\big)du, \\ &\approx f\big(\phi^x(0)\big) = f(x), \quad x \in \mathcal{M}, \end{aligned}$$

where the above approximation holds since: 1. $\phi^x(0) = x$; 2. ϕ^x preserve local distances (Proposition 67 (3)); 3. the Jacobian $\sqrt{\det(g_{ij}^{\phi}(u))}$ is close to one (Proposition 67 (2)). From this heuristic argument, we can see that the approximation error $||I_a(w) - f_0||_\infty$ is determined by two factors: the convolution error $\left|\left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{a^2||u||^2}{2}\right\} f\big(\phi^x(u)\big)du - f(x)\right|$ and the non-flat error caused by the nonzero curvature of $\mathcal{M}$. Moreover, we can expand each of these errors as a polynomial of $1/a$ and call the expansion term related to $1/a^k$ as kth order error.

When $\mathcal{M}$ is Euclidean space $\mathbb{R}^d$, the non-flat error is zero, and by Taylor expansion the convolution error has order s if $f_0 \in C^s(\mathbb{R}^d)$ and $s \leqslant 2$, where $C^s(\mathbb{R}^d)$ is the Holder class of s-smooth functions on $\mathbb{R}^d$. This is because the Gaussian kernel $\exp\{-||(x-y)||^2/2\}$ has a vanishing moment up to first order: $\int x \exp(-||(x-y)||^2/2)dx = 0$. Generally, the convolution error could have order up to $s+1$ if the convolution kernel K has vanishing moments up to order s, i.e. $\int x^t K(x)dx = 0, t = 1, \ldots, s$. However, for general manifold $\mathcal{M}$ with non-vanishing curvature tensor, the non-flat error always has order two (see the proof of Lemma 71). This implies that even though carefully chosen kernels for the covariance function can improve the convolution error to have order higher than two, the overall approximation still tends to

have second order error due to the deterioration caused by the nonzero curvature of the manifold. The following lemma formalizes the above heuristic argument on the order of the approximation error by (B.10) and further provides an upper bound on the decentering function.

Lemma 71. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of R^D. Let $C^s(\mathcal{M})$ be the set of all functions on $\mathcal{M}$ with holder smoothness s. Then for any $f \in C^s(\mathcal{M})$ with $s \leqslant \min\{2, \gamma\}$, there exist constants $a_0 \geqslant 1$, $C > 0$ and $B > 0$ depending only on μ, $\mathcal{M}$ and f such that for all $a \geqslant a_0$,*

$$\inf\{||h||^2_{\tilde{\mathbb{H}}^a} : \sup_{x\in\mathcal{M}} |h(x) - f(x)| \leqslant Ca^{-s}\} \leqslant Ba^d.$$

Proof. The proof consists of two parts. In the first part, we prove that the approximation error of $I_a(f)$ can be decomposed into four terms. The first term T_1 is the convolution error defined in our previous heuristic argument. The second term T_2 is caused by localization of the integration, which is negligible due to the exponential decaying of the squared exponential covariance function. The third and fourth terms T_3, T_4 correspond to the non-flat error, with T_3 caused by approximating the geodesic distance with Euclidean distance $\left|||\phi^q(u) - q||^2 - ||u||^2\right|$, and T_4 by approximating the Jacobian $\left|\sqrt{\det(g^\phi_{ij}(u))} - 1\right|$. Therefore the overall approximation error $|I_a(f)(x) - f(x)|$ has order s in the sense that for some constant $C > 0$ dependent on $\mathcal{M}$ and f:

$$\sup_{x\in\mathcal{M}} |I_a(f)(x) - f(x)| \leqslant Ca^{-s},\ s \leqslant \min\{2, \gamma\}. \tag{B.11}$$

In the second part, we prove that $I_a(f)$ belongs to $\tilde{\mathbb{H}}^a$ and has a squared RKHS norm:

$$||I_a(f)||^2_{\tilde{\mathbb{H}}^a} \leqslant Ba^d,$$

where B is a positive constant not dependent on a.

Step 1 (Estimation of the approximation error): This part follows similar ideas as in the proof of Theorem 1 in Ye and Zhou (2008), where they have shown that (B.11) holds for $s \leqslant 1$. Our proof generalizes their results to $s \leqslant 2$ and therefore needs more careful estimations.

By Proposition 67, for each $p \in \mathcal{M}$, there exists a neighborhood W_p and an associated δ_p satisfying the two conditions in Proposition 66 and equations (B.4)-(B.6). By compactness, $\mathcal{M}$ can be covered by $\cup_{p\in\mathcal{P}} W_p$ for a finite subset $\mathcal{P}$ of $\mathcal{M}$. Then $\sup_{x\in\mathcal{M}} |I_a(f)(x) - f(x)| = \sup_{p\in\mathcal{P}}\{\sup_{x\in W_p} |I_a(f)(x) - f(x)|\}$. Let $\delta^* = \min_{p\in\mathcal{P}}\{\min\{\delta_p, 1/\sqrt{2C_p}\}\} > 0$, where C_p is defined as in equation (B.6). Choose $a_0 \geqslant 1$ sufficiently large such that $C_0\sqrt{(2d+8)\log a_0}/a_0 < \delta^*$, where C_0 is the C_2 in Lemma 68.

Let $q \in W_p$ and $a > a_0$. Define $B_a^q = \left\{x \in \mathcal{M} : d_{\mathcal{M}}(q, x) < C_0\sqrt{(2d+8)\log a}/a\right\}$. Combining equation (B.3) and the fact that $\mathcal{E}_q$ is a diffeomorphism on $B_{\delta^*}(0)$,

$$B_a^q = \left\{\mathcal{E}_q(\sum_{i=1}^d u_i e_i^q) : u \in \tilde{B}_a\right\} \subset \mathcal{E}_q(B_{\delta^*}(0)),$$

where $\tilde{B}_a = \left\{u : ||u|| < C_0\sqrt{(2d+8)\log a}/a\right\} \subset B_{\delta^*}(0)$.

Denote $\phi^q(u) = \mathcal{E}_q(\sum_{i=1}^d u_i e_i^q)$. Then $B_a^q = \phi^q(\tilde{B}_a)$. By definition (B.1),

$$\begin{aligned}
&\left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{B_a^q} K^a(x,y) f(y) dV(y) \\
&= \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2||q - \phi^q(u)||^2}{2}\right\} f(\phi^q(u))\sqrt{\det(g_{ij}^q)}(u) du.
\end{aligned}$$

Therefore, by (B.10) we have the following decomposition:

$$I_a(f)(q) - f(q) = T_1 + T_2 + T_3 + T_4,$$

where

$$T_1 = \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2||u||^2}{2}\right\}\left[f\left(\phi^q(u)\right) - f\left(\phi^q(0)\right)\right]du$$

$$T_2 = \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}\backslash B_a^q} K^a(q,y)f(y)dV(y) - \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d\backslash\tilde{B}_a} \exp\left\{-\frac{a^2||u||^2}{2}\right\}f(q)du,$$

$$T_3 = \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \left\{\exp\left\{-\frac{a^2||q-\phi^q(u)||^2}{2}\right\} - \exp\left\{-\frac{a^2||u||^2}{2}\right\}\right\}f(\phi^q(u))du,$$

$$T_4 = \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2||q-\phi^q(u)||^2}{2}\right\}f(\phi^q(u))(\sqrt{\det(g_{ij}^q)}(u) - 1)du.$$

Step 1.1 (Estimation of T_1): Let $g = f \circ \phi^q$. Since $f \in C^s(\mathcal{M})$ and $(\phi^q, B_{\delta*}(0))$ is a C^γ coordinate chart, we have $g \in C^s(\mathbb{R}^d)$ and therefore

$$g(u) - g(0) = \begin{cases} R(u,s), & \text{if } 0 < s \leqslant \min\{1,\gamma\}, \\ \sum_{i=1}^d \frac{\partial g}{\partial u_i}(0)u_i + R(u,s), & \text{if } 1 < s \leqslant \min\{2,\gamma\}, \end{cases}$$

where the remainder term $|R(u,s)| \leqslant C_1||u||^s$ for all $0 < s \leqslant \min\{2,\gamma\}$. Since $\tilde{B}_a$ is symmetric,

$$\int_{\tilde{B}_a} \exp\left\{-\frac{a^2||u||^2}{2}\right\}u_i du = 0, \quad i = 1,\ldots,d,$$

and therefore

$$|T_1| \leqslant C_1\left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\tilde{B}_a} \exp\left\{-\frac{a^2||u||^2}{2}\right\}||u||^s du = C_2 a^{-s}.$$

Step 1.2 (Estimation of T_2): Denote $T_2 = S_1 + S_2$ where S_1 and S_2 are the first term and second term of T_2, respectively. By Lemma 68, for $y \in \mathcal{M}\backslash B_a^q$, $||q-y|| \geqslant$

$d_{\mathcal{M}}(q,y)/C_0 \geqslant \sqrt{(2d+8)\log a}/a$. Therefore,

$$\begin{aligned}
|S_1| &= \left|\left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathcal{M}\setminus B_a^q} \exp\left\{-\frac{a^2||q-y||^2}{2}\right\} f(y)dV(y)\right| \\
&\leqslant ||f||_\infty \mathrm{Vol}(\mathcal{M})\left(\frac{a}{\sqrt{2\pi}}\right)^d \exp\left\{-\frac{(2d+8)\log a}{2}\right\} \\
&= C_3 a^{-4} \leqslant C_3 a^{-s}.
\end{aligned}$$

As for S_2, we have

$$\begin{aligned}
|S_2| &\leqslant ||f||_\infty \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{||u||\geqslant C_0\sqrt{(2d+8)\log a}/a} \exp\left\{-\frac{a^2||u||^2}{2}\right\} du \\
&\leqslant ||f||_\infty \left(\frac{a}{\sqrt{2\pi}}\right)^d \int_{\mathbb{R}^d} \exp\left\{-\frac{C_0^2(2d+8)\log a}{4}\right\} \exp\left\{-\frac{a^2||u||^2}{4}\right\} du \\
&= C_4 a^{-C_0^2(d/2+2)} \leqslant C_4 a^{-s},
\end{aligned}$$

since $d \geqslant 1$, $C_0 \geqslant 1$ and $a \geqslant a_0 \geqslant 1$.

Combining the above inequalities for S_1 and S_2, we obtain

$$|T_2| \leqslant (C_3 + C_4)a^{-s} = C_5 a^{-s}.$$

Step 1.3 (Estimation of T_3): By equation (B.6) in Proposition 67 and equation (B.3), we have

$$\left|||u||^2 - ||q-\phi^q(u)||^2\right| = \left|d_{\mathcal{M}}^2(q,\phi^q(u)) - ||q-\phi^q(u)||^2\right| \leqslant C_p d_{\mathcal{M}}^4(q,\phi^q(u)) = C_p||u||^4. \tag{B.12}$$

Therefore by using the inequality $|e^{-a} - e^{-b}| \leqslant |a-b|\max\{e^{-a}, e^{-b}\}$ for $a, b > 0$, we have

$$\begin{aligned}
|T_3| \leqslant & ||f||_\infty \left(\frac{a}{\sqrt{2\pi}}\right)^d \\
& \int_{\tilde{B}_a} \max\left\{\exp\left\{-\frac{a^2||q-\phi^q(u)||^2}{2}\right\}, \exp\left\{-\frac{a^2||u||^2}{2}\right\}\right\} \frac{a^2||u||^4}{2} du.
\end{aligned}$$

By equation (B.12) and the fact that $u \in \tilde{B}_a$, $||u||^2 \leqslant (\delta^*)^2 \leqslant 1/(2C_p)$ and hence

$$\big|||u||^2 - ||q - \phi^q(u)||^2\big| \leqslant \frac{1}{2}||u||^2, \quad ||q - \phi^q(u)||^2 \geqslant \frac{1}{2}||u||^2. \tag{B.13}$$

Therefore

$$|T_3| \leqslant ||f||_\infty \bigg(\frac{a}{\sqrt{2\pi}}\bigg)^d \int_{\tilde{B}_a} \exp\bigg\{-\frac{a^2||u||^2}{4}\bigg\}\frac{a^2||u||^4}{2}du = C_6 a^{-2} \leqslant C_6 a^{-s},$$

since $a \geqslant a_0 \geqslant 1$.

Step 1.4 (Estimation of T_4): By equation (B.5) in Proposition 67, there exists a constant C_7 depending on the Ricci tensor of the manifold $\mathcal{M}$, such that

$$\big|\sqrt{\det(g^q_{ij})}(u) - 1\big| \leqslant C_7||u||^2.$$

Therefore, by applying equation (B.13) again, we obtain

$$|T_4| \leqslant C_4||f||_\infty \bigg(\frac{a}{\sqrt{2\pi}}\bigg)^d \int_{\tilde{B}_a} \exp\bigg\{-\frac{a^2||u||^2}{4}\bigg\}||u||^2 du = C_8 a^{-2} \leqslant C_8 a^{-s}.$$

Combining the above estimates for T_1, T_2, T_3 and T_4, we have

$$\sup_{x\in\mathcal{M}} |I_a(f)(q)(x) - f(q)(x)| \leqslant (C_2 + C_3 + C_6 + C_8)a^{-s} = Ca^{-s}.$$

Step 2 (Estimation of the RKHS norm): Since $\langle K^a(x,\cdot), K^a(y,\cdot)\rangle_{\tilde{\mathbb{H}}^a} = K^a(x,y)$, we have

$$\begin{aligned}||I_a(f)||_{\tilde{\mathbb{H}}^a} &= \bigg(\frac{a}{\sqrt{2\pi}}\bigg)^{2d} \int_{\mathcal{M}}\int_{\mathcal{M}} K^a(x,y)f(x)f(y)dV(x)dV(y) \\ &\leqslant ||f||^2_\infty \bigg(\frac{a}{\sqrt{2\pi}}\bigg)^d \int_{\mathcal{M}} dV(x) \bigg(\frac{a}{\sqrt{2\pi}}\bigg)^d \int_{\mathcal{M}} K^a(x,y)dV(y).\end{aligned}$$

Applying the results of the first part to function $f \equiv 1$, we have

$$\bigg|\bigg(\frac{a}{\sqrt{2\pi}}\bigg)^d \int_{\mathcal{M}} K^a(x,y)dV(y) - 1\bigg| \leqslant Ca^{-2} \leqslant C,$$

since $a \geqslant a_0 \geqslant 1$. Therefore,

$$||I_a(f)||_{\tilde{\mathbb{H}}^a} \leqslant (1+C)||f||_\infty^2 \left(\frac{a}{\sqrt{2\pi}}\right)^d \mathrm{Vol}(\mathcal{M}) = Ba^d.$$

□

B.2.4 Centered small ball probability

As indicated by the proof of Lemma 4.6 in van der Vaart and van Zanten (2009), to obtain an upper bound on $-\log P(||W^a||_\infty < \epsilon)$, we need to provide an upper bound for the covering entropy $\log N(\epsilon, \tilde{\mathbb{H}}_1^a, ||\cdot||_\infty)$ of the unit ball in the RKHS $\tilde{\mathbb{H}}^a$ on the submanifold $\mathcal{M}$. Following the discussion in section 4.1, we want to link $\tilde{\mathbb{H}}^a$ to $\mathbb{H}^a$, the associated RKHS defined on the ambient space $\mathbb{R}^D$. Therefore, we need a lemma to characterize the space $\mathbb{H}^a$ (van der Vaart and van Zanten, 2009, Lemma 4.1).

Lemma 72. *$\mathbb{H}^a$ is the set of real parts of the functions*

$$x \mapsto \int e^{i(\lambda,x)}\psi(\lambda)\mu_a(d\lambda),$$

when ψ runs through the complex Hilbert space $L_2(\mu_a)$. Moreover, the RKHS norm of the above function is $||\psi||_{L_2(\mu_a)}$, where μ_a is the spectral measure of the covariance function K^a.

Based on this representation of $\mathbb{H}^a$ on $\mathbb{R}^D$, van der Vaart and van Zanten (2009) proved an upper bound $Ka^D\big(\log\frac{1}{\epsilon}\big)^{D+1}$ for $\log N(\epsilon, \tilde{\mathbb{H}}_1^a, ||\cdot||_\infty)$ through constructing an ϵ-covering set composed of piecewise polynomials. However, there is no straightforward generalization of their scheme from Euclidean spaces to manifolds. The following lemma provides an upper bound for the covering entropy of $\tilde{\mathbb{H}}_1^a$, where the D in the upper bounds for $\mathbb{H}_1^a$ is reduced to d. The main novelty in our proof is the construction of an ϵ-covering set composed of piecewise transformed polynomials

(B.19) via analytically extending the truncated Taylor polynomial approximations (B.16) of the elements in $\tilde{\mathbb{H}}_1^a$. As the proof indicates, the d in a^d relates to the covering dimension d of $\mathcal{M}$, i.e. the ϵ-covering number $N(\epsilon, \mathcal{M}, \epsilon)$ of $\mathcal{M}$ is proportional to $1/\epsilon^d$. The d in $\left(\log \frac{1}{\epsilon}\right)^{d+1}$ relates to the order of the number k^d of coefficients in piecewise transformed polynomials of degree k in d variables.

Lemma 73. *Assume that $\mathcal{M}$ is a d-dimensional C^γ compact submanifold of R^D with $\gamma \geqslant 2$. Then for squared exponential covariance function K^a, there exists a constant K depending only on d, D and $\mathcal{M}$, such that for $\epsilon < 1/2$ and $a > \max\{a_0, \epsilon^{-1/(\gamma-1)}\}$, where δ_0 is defined in Lemma 69 and a_0 is a universal constant,*

$$\log N(\epsilon, \tilde{\mathbb{H}}_1^a, ||\cdot||_\infty) \leqslant K a^d \left(\log \frac{1}{\epsilon}\right)^{d+1}.$$

Proof. By Lemma 70 and Lemma 72, a typical element of $\tilde{\mathbb{H}}^a$ can be written as the real part of the function

$$h_\psi(x) = \int e^{i(\lambda, x)} \psi(\lambda) \mu_a(d\lambda), \text{ for } x \in \mathcal{M}$$

for $\psi : \mathbb{R}^D \to \mathbb{C}$ a function with $\int |\psi|^2 \mu_a(d\lambda) \leqslant 1$. This function can be extended to $\mathbb{R}^D$ by allowing $x \in \mathbb{R}^D$. For any given point $p \in \mathcal{M}$, by (B.10), we have a local coordinate $\phi^p : B_{\delta_0}(0) \subset \mathbb{R}^d \to \mathbb{R}^D$ induced by the exponential map $\mathcal{E}_p$. Therefore, for $x \in \phi_p(B_{\delta_0}(0))$, $h_\psi(x)$ can be written in local q-normal coordinates as

$$h_{\psi,p}(u) = h_\psi\big(\phi^p(u)\big) = \int e^{i(\lambda, \phi^p(u))} \psi(\lambda) \mu_a(d\lambda), \ u \in B_{\delta_0}(0). \tag{B.14}$$

Similar to the idea in the proof of Lemma 4.5 in van der Vaart and van Zanten (2009), we want to extend the function $h_{\psi,p}$ to an analytical function $z \mapsto \int e^{i(\lambda, \phi^p(z))} \psi(\lambda) \mu_a(d\lambda)$ on the set $\Omega = \{z \in \mathbb{C}^d : ||\mathrm{Re} z|| < \delta_0, ||\mathrm{Im} z|| < \rho/a\}$ for some $\rho > 0$. Then we can obtain upper bounds on the mixed partial derivatives of

the analytic extension $h_{\psi,p}$ via Cauchy formula, and finally construct an ϵ-covering set of $\tilde{\mathbb{H}}_1^a$ by piecewise polynomials defined on $\mathcal{M}$. Unfortunately, this analytical extension is impossible unless $\phi^p(u)$ is a polynomial. This motivates us to approximate $\phi^p(u)$ by its γth order Taylor polynomial $P_{p,\gamma}(u)$. More specifically, by Lemma 76 and the discussion after Lemma 69, the error caused by approximating $\phi^p(u)$ by $P_{p,\gamma}(u)$ is

$$\left|h_\psi\big(\phi^p(u)\big) - h_\psi\big(P_{p,\gamma}(u)\big)\right| \leqslant a||\phi^p(u) - P_{p,\gamma}(u)|| \leqslant Ca||u||^\gamma. \tag{B.15}$$

For notation simplicity, fix p as a center and denote the function $h_\psi\big(P_{p,\gamma}(u)\big)$ by $r(u)$ for $u \in B_{\delta_0}$. Since $P_{p,\gamma}(u)$ is a polynomial of degree γ, view the function r as a function of argument u ranging over the product of the imaginary axes in $\mathbb{C}^d$, we can extend

$$r(u) = \int e^{i\langle\lambda, P_{p,\gamma}(u)\rangle}\psi(\lambda)\mu_a(d\lambda),\ u \in B_{\delta_0}(0) \tag{B.16}$$

to an analytical function $z \mapsto \int e^{i\langle\lambda, P_{p,\gamma}(z)\rangle}\psi(\lambda)\mu_a(d\lambda)$ on the set $\Omega = \{z \in \mathbb{C}^d : ||\mathrm{Re}z|| < \delta_0, ||\mathrm{Im}z|| < \rho/a\}$ for some $\rho > 0$ sufficiently small determined by the $\delta < 1/2$ in (B.9). Moreover, by Cauchy-Schwarz inequality, $|r(z)| \leqslant C$ for $z \in \Omega$ and $C^2 = \int e^{\delta||\lambda||}\mu(d\lambda)$. Therefore, by Cauchy formula, with D^n denoting the partial derivative of orders $n = (n_1, \ldots, n_d)$ and $n! = n_1! \cdots n_d!$, we have the following bound for partial derivatives of r at any $u \in B_{\delta_0}(0)$,

$$\left|\frac{D^n r(u)}{n!}\right| \leqslant \frac{C}{R^n}, \tag{B.17}$$

where $R = \rho/(a\sqrt{d})$. Based on the inequalities (B.15) and (B.17), we can construct an ϵ-covering set of $\tilde{\mathbb{H}}_1^a$ as follows.

Set $a_0 = \rho/(2\delta_0\sqrt{d})$, then $R < 2\delta_0$. Since $\mathcal{M} \subset [0,1]^D$, with C_2 defined in Lemma 68, let $\{p_1, \ldots, p_m\}$ be an $R/(2C_2)$-net in $\mathcal{M}$ for the Euclidean distance, and let $\mathcal{M} = \bigcup_i B_i$ be a partition of $\mathcal{M}$ in sets $B_1, \ldots, B_m$ obtaining by assigning every

$x \in \mathcal{M}$ to the closest $p_i \in \{p_1, \ldots, p_m\}$. By (B.3) and Lemma 68

$$|(\phi^{p_i})^{-1}(x)| < C_2 \frac{R}{2C_2} = \frac{R}{2} < \delta_0, \tag{B.18}$$

where ϕ_{p_i} is the local normal coordinate chart at p_i. Therefore, we can consider the piecewise transformed polynomials $P = \sum_{i=1}^m P_{i,a_i} 1_{B_i}$, with

$$P_{i,a_i}(x) = \sum_{n_\cdot \leqslant k} a_{i,n} [(\phi^{p_i})^{-1}(x)]^n, \ x \in \phi^{p_i}\big(B_{\delta_0}(0)\big). \tag{B.19}$$

Here the sum ranges over all multi-index vectors $n = (n_1, \ldots, n_d) \in (\mathbb{N} \cup \{0\})^d$ with $n_\cdot = n_1 + \cdots + n_d \leqslant k$. Moreover, for $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, the notation y^n used above is short for $y_1^{n_1} y_2^{n_2} \cdots y_d^{n_d}$. We obtain a finite set of functions by discretizing the coefficients $a_{i,n}$ for each i and n over a grid of meshwidth ϵ/R^n-net in the interval $[-C/R^n, C/R^n]$ (by (B.17)). The log cardinality of this set is bounded by

$$\log\left(\prod_i \prod_{n:n_\cdot \leqslant k} \#a_{i,n}\right) \leqslant m \log\left(\prod_{n:n_\cdot \leqslant k} \frac{2C/R^n}{\epsilon/R^n}\right) \leqslant m k^d \log\left(\frac{2C}{\epsilon}\right).$$

Since $R = \rho/(a\sqrt{d})$, we can choose $m = N\big(\mathcal{M}, ||\cdot||, \rho/(2C_0 a d^{1/2})\big) \simeq a^d$. To complete the proof, it suffices to show that for k of order $\log(1/\epsilon)$, the resulting set of functions is a $K\epsilon$-net for constant K depending only on μ.

For any function $f \in \tilde{\mathbb{H}}_1^a$, by Lemma 70, we can find a $g \in \tilde{\mathbb{H}}_1^a$ such that $g|_{\mathcal{M}} = f$. Assume that r_g (the subcript g indicates the dependence on g) is the local polynomial approximation for g defined as (B.16). Then we have a partial derivative bound on r_g as:

$$\left|\frac{D^n r_g(p_i)}{n!}\right| \leqslant \frac{C}{R^n}.$$

Therefore there exists a universal constant K and appropriately chosen a_i in (B.19), such that for any $z \in B_i \subset \mathcal{M}$,

$$\left|\sum_{n_\cdot > k} \frac{D^n r_g(p_i)}{n!}(z - p_i)^n\right| \leqslant \sum_{n_\cdot > k} \frac{C}{R^n}(R/2)^n \leqslant C \sum_{l=k+1}^{\infty} \frac{l^{d-1}}{2^l} \leqslant KC\left(\frac{2}{3}\right)^k,$$

$$\left| \sum_{n_\cdot \leqslant k} \frac{D^n r_g(p_i)}{n!}(z-p_i)^n - P_{i,n_i}(z) \right| \leqslant \sum_{n_\cdot \leqslant k} \frac{\epsilon}{R^n}(R/2)^n \leqslant \sum_{l=1}^{k} \frac{l^{d-1}}{2^l}\epsilon \leqslant K\epsilon.$$

Moreover, by (B.15) and (B.18),

$$|g(z) - r_g(z)| \leqslant Ca||(\phi^{p_i})^{-1}(z)||^\gamma \leqslant aR^\gamma \leqslant Ka^{-(\gamma-1)} < K\epsilon,$$

where the last step follows by the condition on a.

Consequently, we obtain

$$|f(z)-P_{i,n_i}(z)| = |g(z)-P_{i,n_i}(z)| \leqslant |g(z)-r_g(z)|+|r_g(z)-P_{i,n_i}(z)| \leqslant KC\left(\frac{2}{3}\right)^k + 2K\epsilon.$$

This suggests that the piecewise polynomials form a $3K\epsilon$-net for k sufficiently large so that $(2/3)^k$ is smaller than $K\epsilon$. □

Similar to Lemma 4.6 in van der Vaart and van Zanten (2009), Lemma 73 implies an upper bound on $-\log P(||W^a||_\infty < \epsilon)$.

Lemma 74. *Assume that $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of $\mathbb{R}^D$ with $\gamma \geqslant 2$. If K^a is the squared exponential covariance function with inverse bandwidth a, then for some $a_0 > 0$, there exist constants C and ϵ_0 that only depend on a_0, μ, d, D and $\mathcal{M}$, such that, for $a \geqslant \max\{a_0, \epsilon^{-1/(\gamma-1)}\}$ and $\epsilon < \epsilon_0$,*

$$-\log P\Big(\sup_{x\in\mathcal{M}} |W_x^a| \leqslant \epsilon\Big) \leqslant Ca^d\left(\log\frac{a}{\epsilon}\right)^{d+1}.$$

Before proving Theorem 13, we need another two technical lemmas for preparations, which are the analogues of Lemma 4.7 and 4.8 in van der Vaart and van Zanten (2009) for RKHS on Euclidean spaces.

Lemma 75. *For squared exponential covariance function, if $a \leqslant b$, then $\sqrt{a}\tilde{\mathbb{H}}_1^a \subset \sqrt{b}\tilde{\mathbb{H}}_1^b$.*

Proof. For any $f \in \sqrt{a}\tilde{\mathbb{H}}_1^a$, by Lemma 70, there exists $g \in \sqrt{a}\mathbb{H}_1^a$ such that $g|_{\mathcal{M}} = f$. By Lemma 4.7 in van der Vaart and van Zanten (2009), $\sqrt{a}\mathbb{H}_1^a \subset \sqrt{b}\mathbb{H}_1^b$, so $g \in \sqrt{b}\mathbb{H}_1^b$. Again by Lemma 70, since $g|_{\mathcal{M}} = f$, $||f||_{\tilde{\mathbb{H}}^b} \leqslant ||g||_{\mathbb{H}^b} \leqslant \sqrt{b}$, implying that $f \in \sqrt{b}\tilde{\mathbb{H}}_1^b$. □

Lemma 76. *Any $h \in \tilde{\mathbb{H}}_1^a$ satisfies $|h(x)| \leqslant 1$ and $|h(x) - h(x')| \leqslant a||x - x'||\tau$ for any $x, x' \in \mathcal{M}$, where $\tau^2 = \int ||\lambda||^2 d\mu(\lambda)$.*

Proof. By the reproducing property and Cauchy-Schwarz inequality

$$\begin{aligned} |h(x)| &= |\langle h, K^a(x,\cdot)\rangle_{\tilde{\mathbb{H}}^a}| \leqslant ||K^a(x,\cdot)||_{\tilde{\mathbb{H}}^a} = 1 \\ |h(x) - h(x')| &= |\langle h, K^a(x,\cdot) - K^a(x',\cdot)\rangle_{\tilde{\mathbb{H}}^a}| \\ &\leqslant ||K^a(x,\cdot) - K^a(x',\cdot)||_{\tilde{\mathbb{H}}^a} \\ &= \sqrt{2(1 - K^a(x,x'))}. \end{aligned}$$

By the spectral representation $K(x,x') = \int e^{i(\lambda,t)}\mu_a(d\lambda)$ and the fact that μ_a is symmetric,

$$\begin{aligned} 2(1 - K^a(x,x')) &= 2\int (1 + i(\lambda, x - x') - e^{i(\lambda, x-x')})\mu_a(d\lambda) \\ &\leqslant ||x - x'||^2 \int ||\lambda||^2 \mu_a(d\lambda) \\ &= a^2||x - x'||^2 \int ||\lambda||^2 \mu(d\lambda). \end{aligned}$$

□

B.2.5 Posterior contraction rate of GP on manifold

We provide proofs for Theorem 13 and Theorem 14.

Proof of Theorem 13. Define centered and decentered concentration functions of the

process $W^a = (W_{ax} : x \in \mathcal{M})$ by

$$\phi_0^a(\epsilon) = -\log P(|W^a|_\infty \leqslant \epsilon),$$
$$\phi_{f_0}^a(\epsilon) = \inf_{h \in \tilde{\mathbb{H}}^a : |h - f_0|_\infty \leqslant \epsilon} ||h||_{\tilde{\mathbb{H}}^a}^2 - \log P(|W^a|_\infty \leqslant \epsilon),$$

where $|h|_\infty = \sup_{x \in \mathcal{M}} |f(x)|$ is the sup norm on the manifold $\mathcal{M}$. Then $P(|W^a|_\infty \leqslant \epsilon) = \exp(-\phi_0^a(\epsilon))$ by definition. Moreover, by the results in Kuelbs and Linde (1994),

$$P(||W^a - f_0||_\infty \leqslant 2\epsilon) \geqslant e^{-\phi_{f_0}^a(\epsilon)}. \tag{B.20}$$

Suppose that $f_0 \in C^s(\mathcal{M})$ for some $s \leqslant \min\{2, \gamma - 1\}$. By Lemma 74 and Lemma 71, for $a > a_0$ and $\epsilon > C \max\{a^{-(\gamma-1)}, a^{-s}\} = Ca^{-s}$,

$$\phi_{f_0}^s(\epsilon) \leqslant Da^d + C_4 a^d \left(\log \frac{a}{\epsilon} \right)^{1+d} \leqslant K_1 a^d \left(\log \frac{a}{\epsilon} \right)^{1+d}.$$

Since A^d has a Gamma prior, there exists $p, C_1, C_2 > 0$, such that $C_1 a^p \exp(-D_2 a^d) \leqslant g(a) \leqslant C_2 a^p \exp(-D_2 a^d)$. Therefore by equation (B.20),

$$\begin{aligned} P(||W^A - f_0||_\infty \leqslant 2\epsilon) &\geqslant P(||W^A - f_0||_\infty \leqslant 2\epsilon, A \in [(C/\epsilon)^{1/s}, 2(C/\epsilon)^{1/s}]) \\ &\geqslant \int_{(C/\epsilon)^{1/s}}^{2(C/\epsilon)^{1/s}} e^{-\phi_{f_0}^s(\epsilon)} g(a) da \\ &\geqslant C_1 e^{-K_2 (1/\epsilon)^{d/s} (\log(1/\epsilon))^{1+d}} \left(\frac{C}{\epsilon} \right)^{p/s} \left(\frac{C}{\epsilon} \right)^{1/s}. \end{aligned}$$

Therefore,

$$P(||W^A - f_0||_\infty \leqslant \epsilon_n) \geqslant \exp(-n\epsilon_n^2),$$

for ϵ_n a large multiple of $n^{-s/(2s+d)} (\log n)^{\kappa_1}$ with $\kappa_1 = (1+d)/(2+d/s)$ and sufficiently large n.

Similar to the proof of Theorem 3.1 of van der Vaart and van Zanten (2009), by Lemma 75,

$$B_{M,r,\delta,\epsilon} = \left(M \sqrt{\frac{r}{\delta}} \tilde{\mathbb{H}}_1^r + \epsilon \mathbb{B}_1 \right) \cup \left(\bigcup_{a < \delta} (M \tilde{\mathbb{H}}_1^a) + \epsilon \mathbb{B}_1 \right),$$

with $\mathbb{B}_1$ the unit ball of $C(\mathcal{M})$, contains the set $M\tilde{\mathbb{H}}_1^a + \epsilon\mathbb{B}_1$ for any $a \in [\delta, r]$. Furthermore, if

$$M \geqslant 4\sqrt{\phi_0^r(\epsilon)} \quad \text{and} \quad e^{-\phi_0^r(\epsilon)} < 1/4, \tag{B.21}$$

then

$$P(W^A \notin B) \leqslant \frac{2C_2 r^{p-d+1} e^{-D_2 r^d}}{D_2 d} + e^{-M^2/8}. \tag{B.22}$$

By Lemma 74, equation (B.21) is satisfied if

$$M^2 \geqslant 16C_4 r^d (\log(r/\epsilon))^{1+d}, \quad r > 1, \quad \epsilon < \epsilon_1,$$

for some fixed $\epsilon_1 > 0$. Therefore

$$P(W^A \notin B) \leqslant \exp(-C_0 n\epsilon_n^2),$$

for r and M satisfying

$$r^d = \frac{2C_0}{D_2} n\epsilon_n^2, \quad M^2 = \max\{8C_0, 16C_4\} n\epsilon_n^2 (\log(r/\epsilon_n))^{1+d}. \tag{B.23}$$

Denote the solution of the above equation as r_n and M_n.

By Lemma 73, for $M\sqrt{r/\delta} > 2\epsilon$ and $r > a_0$,

$$\begin{aligned}\log N\left(2\epsilon, M\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}_1^r + \epsilon\tilde{\mathbb{B}}_1, ||\cdot||_\infty\right) &\leqslant \log N\left(\epsilon, M\sqrt{\frac{r}{\delta}}\tilde{\mathbb{H}}_1^r, ||\cdot||_\infty\right) \\ &\leqslant Kr^d\left(\log\left(\frac{M\sqrt{r/\delta}}{\epsilon}\right)\right)^{1+d}.\end{aligned}$$

By Lemma 76, every element of $M\tilde{\mathbb{H}}_1^a$ for $a < \delta$ is uniformly at most $\delta\sqrt{D}\tau M$ distant from a constant function for a constant in the interval $[-M, M]$. Therefore for $\epsilon > \delta\sqrt{D}\tau M$,

$$\log N\left(3\epsilon, \bigcup_{a<\delta}(M\tilde{\mathbb{H}}_1^a) + \epsilon\tilde{\mathbb{B}}_1, ||\cdot||_\infty\right) \leqslant N(\epsilon, [-M, M], |\cdot|) \leqslant \frac{2M}{\epsilon}.$$

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

$$M \geqslant \epsilon,\ M^{3/2}\sqrt{2\tau r}D^{1/4} \geqslant 2\epsilon^{3/2},\ r > a_0,$$

which is satisfied when $r = r_n$ and $M = M_n$, we have

$$\log N\big(3\epsilon, B, ||\cdot||_\infty\big) \leqslant Kr^d\bigg(\log\bigg(\frac{M^{3/2}\sqrt{2\tau r}D^{1/4}}{\epsilon^{3/2}}\bigg)\bigg)^{1+d} + \log\frac{2M}{\epsilon}. \tag{B.24}$$

Therefore, for $r = r_n$ and $M = M_n$,

$$\log N\big(3\bar{\epsilon}_n, B, ||\cdot||_\infty\big) \leqslant n\bar{\epsilon}_n^2,$$

for $\bar{\epsilon}_n$ a large multiple of $\epsilon_n(\log n)^{\kappa_2}$ with $\kappa_2 = (1+d)/2$. □

Proof of Theorem 14. Under d', the prior concentration inequality becomes:

$$\begin{aligned} P(||W^A - f_0||_\infty \leqslant 2\epsilon) &\geqslant P(||W^A - f_0||_\infty \leqslant 2\epsilon, A \in [(C/\epsilon)^{1/s}, 2(C/\epsilon)^{1/s}]) \\ &\geqslant \int_{(C/\epsilon)^{1/s}}^{2(C/\epsilon)^{1/s}} e^{-\phi_{f_0}^s(\epsilon)} g(a)da \\ &\geqslant C_1 e^{-K_2(1/\epsilon)^{d\vee d'/s}(\log(1/\epsilon))^{1+d}}\bigg(\frac{C}{\epsilon}\bigg)^{p/s}\bigg(\frac{C}{\epsilon}\bigg)^{1/s}. \end{aligned} \tag{B.25}$$

The complementary probability becomes:

$$P(W^A \notin B) \leqslant \frac{2C_2 r^{p-d'+1}e^{-D_2 r^{d'}}}{D_2} + e^{-M^2/8}, \tag{B.26}$$

with $M^2 \geqslant 16C_4 r^d(\log(r/\epsilon))^{1+d}$, $r > 1$ and $\epsilon < \epsilon_1$, where $\epsilon_1 > 0$ is a fixed constant.

An upper bound for the covering entropy is unchanged and still given by (B.24).

1. $d' > d$: With ϵ_n a multiple of $n^{-s/(2s+d')}(\log n)^{\kappa_1}$ with $\kappa_1 = (1+d)/(2+d'/s)$, $\bar{\epsilon}_n < \epsilon_n$,

$$r^{d'} = \frac{2C_0}{D_2}n\epsilon_n^2, \text{ and } M^2 = \max\{8C_0, 16C_4\}n\epsilon_n^2(\log(r/\epsilon_n))^{1+d},$$

inequalities (B.25), (B.26) and (B.24) become

$$\begin{aligned} P(||W^A - f_0||_\infty \leqslant \epsilon_n) \geqslant & \exp(-n\epsilon_n^2), \\ P(W^A \notin B) \leqslant & \exp(-C_0 n\epsilon_n^2), \\ \log N\big(3\bar{\epsilon}_n, B, ||\cdot||_\infty\big) \leqslant & n\bar{\epsilon}_n^2. \end{aligned}$$

Comparing the above with (B.8), we arrive at the conclusion that under $d' > d$, the posterior contraction rate will be at least a multiple of $n^{-s/(2s+d')}(\log n)^\kappa$ with $\kappa = (1+d)/(2+d'/s)$.

2. $\frac{d^2}{2s+d} < d' < d$*:* With ϵ_n a multiple of $n^{-s/(2s+d)}(\log n)^{\kappa_1}$ with $\kappa_1 = (1+d)/(2+d/s)$, $\bar{\epsilon}_n$ a multiple of $n^{d/(2d')-1}\epsilon_n^{d/d'}(\log n)^{(d+1)/2} = n^{-\frac{(2s+d)d'-d^2}{2(2s+d)d'}}(\log n)^{\kappa_2}$ with $\kappa_2 = (d+d^2)/(2d'+dd'/s) + (1+d)/2$,

$$r^{d'} = \frac{2C_0}{D_2} n\epsilon_n^2, \text{ and } M^2 = \max\{8C_0, 16C_4\} n\epsilon_n^2 (\log(r/\epsilon_n))^{1+d},$$

inequalities (B.25), (B.26) and (B.24) become

$$\begin{aligned} P(||W^A - f_0||_\infty \leqslant \epsilon_n) \geqslant & \exp(-n\epsilon_n^2), \\ P(W^A \notin B) \leqslant & \exp(-C_0 n\epsilon_n^2), \\ \log N\big(3\bar{\epsilon}_n, B, ||\cdot||_\infty\big) \leqslant & n\bar{\epsilon}_n^2. \end{aligned}$$

Comparing the above with (B.8), we arrive at the conclusion that under $d' < d$, the posterior contraction rate will be at least a multiple of $n^{-\frac{(2s+d)d'-d^2}{2(2s+d)d'}}(\log n)^\kappa$ with $\kappa = (d+d^2)/(2d'+dd'/s) + (1+d)/2$. To make this rate meaningful, we need $(2s+d)d' - d^2 > 0$, i.e. $d' > d^2/(2s+d)$. □

Appendix C

Appendix for Chapter 4

C.1 Proofs of technical results in Chapter 4

C.1.1 Proof of Theorem 17

First reshape $P(y|x_1,\ldots,x_p)$ according to x_1 as a matrix $A^{(1)}$ of size $d_1 \times d_0 d_2 d_3 \ldots d_p$, with the h^{th} row a long vector,

$$\big\{P(1|h,1,\ldots,1,1), P(1|h,1,\ldots,1,2),\ldots,P(1|h,1,\ldots,1,d_p),$$
$$P(1|h,1,\ldots,2,1),\ldots,P(1|h,1,\ldots,2,d_j),\ldots,P(d_0|h,d_2,\ldots,d_{p-1},d_p)\big\},$$

denoted $A^{(1)}\{h,(y,x_2,\ldots,x_p)\}$. Apply nonnegative matrix decomposition for $A^{(1)}$, we obtain

$$P(y|x_1,\ldots,x_p) = A^{(1)}\{x_1,(y,x_2,\ldots,x_p)\} = \sum_{h_1=1}^{k_1} \lambda^{(1)}_{h_1 x_2 \ldots x_p}(y)\pi^{(1)}_{h_1}(x_1), \qquad \text{(C.1)}$$

where $k_1 \leqslant d_1$ corresponds to the nonnegative rank of the matrix $A^{(1)}$ (Cohen and Rothblum, 1993). Without loss of generality, we can assume that the parameters satisfy the constraints $\sum_{y=1}^{d_0} \lambda^{(1)}_{h_1 x_2 \ldots x_p}(y) = 1$ for each $(h_1, x_2, \ldots, x_p)$, $\sum_{h_1=1}^{k_1} \pi^{(1)}_{h_1}(x_1) = 1$

for each x_1, $\lambda^{(1)}_{h_1 x_2 \ldots x_p}(y) \geqslant 0$, and $\pi^{(1)}_{h_1}(x_1) \geqslant 0$. Otherwise, we can always define new $\tilde{\lambda}$'s and $\tilde{\pi}$'s satisfying the above constraints with the same k_1 through the original λ's and π's as following:

$$\tilde{\lambda}^{(1)}_{h_1 x_2 \ldots x_p}(y) = \frac{\lambda^{(1)}_{h_1 x_2 \ldots x_p}(y)}{s_{h_1 x_2 \ldots x_p}},$$

$$\tilde{\pi}^{(1)}_{h_1}(x_1) = s_{h_1 x_2 \ldots x_p} \pi^{(1)}_{h_1}(x_1),$$

where $s_{h_1 x_2 \ldots x_p} = \sum_{y=1}^{d_0} \lambda^{(1)}_{h_1 x_2 \ldots x_p}(y)$. With this definition, the decomposition (C.1) for the new $(\tilde{\lambda}, \tilde{\pi})$'s and the normalizing constraint $\sum_{y=1}^{d_0} \tilde{\lambda}^{(1)}_{h_1 x_2 \ldots x_p}(y) = 1$ are easy to verify. We only need to check the normalizing constraint for $\tilde{\pi}$:

$$\begin{aligned}\sum_{h_1=1}^{k_1} \tilde{\pi}^{(1)}_{h_1}(x_1) &= \sum_{h_1=1}^{k_1} \sum_{y=1}^{d_0} \lambda^{(1)}_{h_1 x_2 \ldots x_p}(y) \pi^{(1)}_{h_1}(x_1) \\ &= \sum_{y=1}^{d_0} P(y | x_1, \ldots, x_p) = 1,\end{aligned}$$

where we have applied (C.1) and the fact that P is a conditional probability.

Taking $\lambda^{(1)}_{h_1 x_2 \ldots x_p}(y)$ from (C.1) with argument x_2, we can apply the same type of decomposition to obtain

$$\lambda^{(1)}_{h_1 x_2 \ldots x_p}(y) = \sum_{h_2=1}^{k_2} \lambda^{(2)}_{h_1 h_2 x_3 \ldots x_p}(y) \pi^{(2)}_{h_2}(x_2),$$

subject to $\sum_{y=1}^{d_0} \lambda^{(2)}_{h_1 h_2 \ldots x_p}(y) = 1$, for each $(h_1, h_2, \ldots, x_p)$, $\sum_{h_2=1}^{k_2} \pi^{(2)}_{h_2}(x_2) = 1$, for each x_2, $\lambda^{(2)}_{h_1 h_2 \ldots x_p}(c) \geqslant 0$, and $\pi^{(2)}_{h_2}(x_2) \geqslant 0$. Plugging back into equation (C.1),

$$P(y | x_1, \ldots, x_p) = \sum_{h_1=1}^{k_1} \sum_{h_2=1}^{k_2} \lambda^{(2)}_{h_1 h_2 x_3 \ldots x_p}(y) \pi^{(1)}_{h_1}(x_1) \pi^{(2)}_{h_2}(x_2).$$

Repeating this procedure another $(p-2)$ times, we obtain equation (C.12) with $\lambda_{h_1 h_2 \ldots h_p}(y) = \lambda^{(p)}_{h_1 h_2 \ldots h_p}(y)$ and constraints (C.10).

Remark: As we can seen from the proof, k_j can be considered as the nonnegative matrix rank corresponds to certain transformation of the jth mode matrix of the tensor P.

C.1.2 Proof of Lemma 18

Given the degeneracy of π, the bias square term can be written as

$$\text{Bias}^2 = \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1 \ldots h_p}} \left(E\tilde{\lambda}_{h_1 \ldots h_p}(y) - P_0(y|x_1,\ldots,x_p)\right)^2 G(dx_1,\ldots,dx_p),$$

where $A_{h_1 \ldots h_p} = \{(x_1,\ldots,x_p) : h_j(x_j) = h_j, j = 1,\ldots,p\}$ and $\tilde{\lambda}$'s are arbitrary estimators of λ's. It can be verified that the above expression is minimized if and only if:

$$E\tilde{\lambda}_{h_1 \ldots h_p}(y) = \frac{\int_{A_{h_1 \ldots h_p}} P_0(y|x_1,\ldots,x_p) G(dx_1,\ldots,dx_p)}{\int_{A_{h_1 \ldots h_p}} G(dx_1,\ldots,dx_p)} \tag{C.2}$$

holds for all possible $(h_1,\ldots,h_p)$. So we only need to check the the MLE $\hat{\lambda}$'s satisfy this condition.

Let $N_{x_1,\ldots,x_p} = \sum_{i=1}^{n} I(X_{i1} = x_1,\ldots,X_{ip} = x_p)$, $\bar{N}_{h_1,\ldots,h_p} = \sum_{A_{h_1 \ldots h_p}} N_{x_1,\ldots,x_p}$, $X = \{X_1,\ldots,X_p\}$ and $Y = \{Y_1,\ldots,Y_p\}$. By the iterative expectation formula:

$$E_{X,Y}\hat{\lambda}_{h_1 \ldots h_p}(y) = \sum_{A_{h_1 \ldots h_p}} E_X \frac{N_{x_1,\ldots,x_p}}{\bar{N}_{h_1,\ldots,h_p}} P_0(y|x_1,\ldots,x_p). \tag{C.3}$$

Note that

$$N_{x_1,\ldots,x_p} \big| \bar{N}_{h_1,\ldots,h_p} \sim \text{Bin}\left(\bar{N}_{h_1,\ldots,h_p}, \frac{G(x_1,\ldots,x_p)}{\int_{A_{h_1 \ldots h_p}} G(dx_1,\ldots,dx_p)}\right). \tag{C.4}$$

Combining this and the iterative expectation formula:

$$E_X \frac{N_{x_1,\ldots,x_p}}{\bar{N}_{h_1,\ldots,h_p}} P_0(y|x_1,\ldots,x_p) = \frac{G(x_1,\ldots,x_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)} P_0(y|x_1,\ldots,x_p). \tag{C.5}$$

Combining (C.3) and (C.5) together, we can prove that (C.2) holds for the MLE $\hat{\lambda}$.

C.1.3 Proof of Lemma 19

Under the same notation as in Lemma 18,

$$\begin{aligned}
\text{Var} =& \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1\ldots h_p}} E_{X,Y}\big(\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2 G(dx_1,\ldots,dx_p) \\
=& \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1\ldots h_p}} E_X Var_{Y|X}\big(\hat{\lambda}_{h_1\ldots h_p} - E_{Y|X}\hat{\lambda}_{h_1\ldots h_p}\big)^2 G(dx_1,\ldots,dx_p) \\
&+ \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} \int_{A_{h_1\ldots h_p}} E_X\big(E_{Y|X}\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2 G(dx_1,\ldots,dx_p) \\
\triangleq& S_1 + S_2,
\end{aligned}$$

where $E_{Y|X}$ and $Var_{Y|X}$ stand for taking conditional expectation and variance given X, respectively.

Estimation of S_1: First, we estimate the integrand in S_1 similar to (C.3):

$$\begin{aligned}
& E_X Var_{Y|X}\big(\hat{\lambda}_{h_1\ldots h_p} - E_{Y|X}\hat{\lambda}_{h_1\ldots h_p}\big)^2 \\
=& \sum_{A_{h_1\ldots h_p}} E_X \frac{N_{x_1,\ldots,x_p}}{\bar{N}^2_{h_1,\ldots,h_p}} P_0(y|x_1,\ldots,x_p)\big(1 - P_0(y|x_1,\ldots,x_p)\big) \\
=& \frac{\int_{A_{h_1\ldots h_p}} P_0(y|x_1,\ldots,x_p)\big(1 - P_0(y|x_1,\ldots,x_p)\big) G(dx_1,\ldots,dx_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)} E_X \frac{I(\bar{N}_{h_1,\ldots,h_p} > 0)}{\bar{N}_{h_1,\ldots,h_p}},
\end{aligned}$$

where the last step is by (C.4) and the iterative expectation formula. Since $\bar{N}_{h_1,\ldots,h_p} \sim \text{Bin}\big(n, \int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)\big)$, by the asymptotic expansion for the expectation of

reciprocal of Binomial random variables in Stephan (1945),

$$E_X \frac{I(\bar{N}_{h_1,\ldots,h_p} > 0)}{\bar{N}_{h_1,\ldots,h_p}} = \frac{1}{n \int_{A_{h_1\ldots h_p}} G(dx_1, \ldots, dx_p)} + O(n^{-2}), \tag{C.6}$$

we obtain

$$S_1 = C_1 \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} (1/n + o(n^{-2})) = 2C_1|k|/n + O(|k|/n^2),$$

where C_1 is some constant with lower and upper bounds independent of n.

Estimation of S_2: By (C.5), the integrand in S_2 is:

$$\begin{aligned}
&E_X \big(E_{Y|X}\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2 \\
=&E_X\bigg(\sum_{A_{h_1\ldots h_p}} \bigg(\frac{N_{x_1,\ldots,x_p}}{\bar{N}_{h_1,\ldots,h_p}} - \frac{G(x_1,\ldots,x_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)}\bigg) P_0(y|x_1,\ldots,x_p)\bigg)^2.
\end{aligned}$$

Similar to (C.4), the joint conditional distribution of $N_{x_1,\ldots,x_p}$ given $\bar{N}_{h_1,\ldots,h_p}$ follows a multinomial distribution:

$$\begin{aligned}
&\{N_{x_1,\ldots,x_p} : (x_1,\ldots,x_p) \in A_{h_1\ldots h_p}\} \big| \bar{N}_{h_1,\ldots,h_p} \\
&\sim \text{Multi}\bigg(\bar{N}_{h_1,\ldots,h_p}, \bigg\{ \frac{G(x_1,\ldots,x_p)}{\int_{A_{h_1\ldots h_p}} G(dx_1,\ldots,dx_p)} : (x_1,\ldots,x_p) \in A_{h_1\ldots h_p} \bigg\}\bigg).
\end{aligned}$$

As a result, by the iterative expectation formula, $E_X \big(E_{Y|X}\hat{\lambda}_{h_1\ldots h_p} - E_{X,Y}\hat{\lambda}_{h_1\ldots h_p}\big)^2$ is also proportional to $E_X \frac{I(\bar{N}_{h_1,\ldots,h_p}>0)}{\bar{N}_{h_1,\ldots,h_p}}$. Therefore, by (C.6)

$$S_2 = C_2 \sum_{y=1}^{2} \sum_{h_1,\ldots,h_p} (1/n + o(n^{-2})) = 2C_2|k|/n + O(|k|/n^2),$$

where C_2 is some constant with lower and upper bounds independent of n.

Combining the estimation of S_1 and S_2, we obtain the desired results with $C = 2C_1 + 2C_2$.

C.1.4 Proof of Theorem 20

To prove Theorem 20 we need some preliminaries. The following theorem is a minor modification of Theorem 2.1 in Ghosal et al. (2000) and the proof is provided in Appendix C.1.5. For simplicity in notation, we denote the observed data for subject i as X_i with $X_i \overset{iid}{\sim} P \in \mathcal{P}$, $P \sim \Pi$, and the true model P_0.

Theorem 77. *Let ϵ_n be a sequence with $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$, $\sum_n \exp(-n\epsilon_n^2) < \infty$. Let d be the total variance distance, $C > 0$ be a constant and sets $\mathcal{P}_n \subset \mathcal{P}$. Define the following conditions:*

1. *$logN(\epsilon_n, \mathcal{P}_n, d) \leqslant n\epsilon_n^2$;*
2. *$\Pi_n(\mathcal{P} \backslash \mathcal{P}_n) \leqslant \exp\{-(2+C)n\epsilon_n^2\}$;*
3. *$\Pi_n(P : ||\log \frac{P}{P_0}||_\infty < \epsilon_n^2) > \exp(-Cn\epsilon_n^2)$.*

If the above conditions hold for all n large enough, then for $M > \sqrt{16+8C}$,

$$\Pi_n\{P : d(P, P_0) \geqslant M\epsilon_n | X_1, \ldots, X_n\} \to 0 \ a.s. P_0^n.$$

In our case, X_i include the response y_i and predictors x_i, P is the random measure characterizing the unknown joint distribution of (y_i, x_i) and P_0 is the measure characterizing the true joint distribution. As our focus is on the conditional probability, $P(y|x)$, we fix the marginal distribution of X at it's true value $P_0(x)$ and model the unknown conditional $P(y|x)$ independently of the marginal of X. By doing so, it is straightforward to show that we can ignore the marginal of X in using Theorem 2 to study posterior convergence. We simply restrict $\mathcal{P}$ to the set of joint probabilities such that $P(x) \equiv P_0(x)$. The total variation distance between the joint probabilities P and P_0 is equivalent to the distance between the conditionals defined in Theorem

2 by the identity

$$\int \sum_{y=1}^{d_0} \big|P(y, x_1, \ldots, x_p) - P_0(y, x_1, \ldots, x_p)\big| dx_1 \cdots dx_p =$$

$$\int \sum_{y=1}^{d_0} \big|P(y|x_1, \ldots, x_p) - P_0(y|x_1, \ldots, x_p)\big| dG_n(dx_1, \cdots, dx_p).$$

Therefore, we will not distinguish the joint probability and the conditional probability and use P to denote both of them henceforth.

To prove Theorem 2, we also need upper bounds on the distance between two models specified by (C.12) when the models are the same size and when they are nested.

Lemma 78. *Let P and $\tilde{P}$ be two models specified by (C.10) with parameter (k, λ, π) and $(\tilde{k}, \tilde{\lambda}, \tilde{\pi})$, respectively. Assume that P and $\tilde{P}$ have the same multirank $\tilde{k} = k = (k_1, \ldots, k_p)$. Then*

$$d(P, \tilde{P}) \leqslant \sum_{y=1}^{d_0} \max_{h_1, \ldots, h_p} |\lambda_{h_1 h_2 \ldots h_p}(y) - \tilde{\lambda}_{h_1 h_2 \ldots h_p}(y)| + d_0 \sum_{j=1}^{p} \max_{x_j} \sum_{h_j=1}^{k_j} |\pi_{h_j}^{(j)}(x_j) - \tilde{\pi}_{h_j}^{(j)}(x_j)|.$$

Proof. By definition of $d(P, \tilde{P})$, we only need to prove that for any $y = 1, \ldots, d_0$ and any combination of $(x_1, \ldots, x_p)$,

$$\begin{aligned} |P(y|x_1, \ldots, x_p) - \tilde{P}(y|x_1, \ldots, x_p)| \qu\leqslant \quad & \max_{h_1, \ldots, h_p} |\lambda_{h_1 h_2 \ldots h_p}(y) - \tilde{\lambda}_{h_1 h_2 \ldots h_p}(y)| \\ & + \sum_{j=1}^{p} \sum_{h_j=1}^{k_j} |\pi_{h_j}^{(j)}(x_j) - \tilde{\pi}_{h_j}^{(j)}(x_j)|. \end{aligned} \tag{C.7}$$

Actually,

$$|P(y|x_1, \ldots, x_p) - \tilde{P}(y|x_1, \ldots, x_p)| \leqslant A + \sum_{s=1}^{p} B_s,$$

where

$$A = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} |\lambda_{h_1 h_2 \ldots h_p}(y) - \tilde{\lambda}_{h_1 h_2 \ldots h_p}(y)| \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j)$$

$$\leqslant \max_{h_1,\ldots,h_p} |\lambda_{h_1 h_2 \ldots h_p}(y) - \tilde{\lambda}_{h_1 h_2 \ldots h_p}(y)| \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \prod_{j=1}^{p} \pi_{h_j}^{(j)}(x_j)$$

$$= \max_{h_1,\ldots,h_p} |\lambda_{h_1 h_2 \ldots h_p}(y) - \tilde{\lambda}_{h_1 h_2 \ldots h_p}(y)|,$$

where the last step is by using the second equation in (C.9), and

$$B_s = \sum_{h_1=1}^{k_1} \cdots \sum_{h_p=1}^{k_p} \tilde{\lambda}_{h_1 h_2 \ldots h_p}(y) \; |\pi_{h_s}^{(s)}(x_s) - \pi_{h_s}^{(s)}(x_s)| \; \prod_{j=1}^{s-1} \tilde{\pi}_{h_j}^{(j)}(x_j) \prod_{j=s+1}^{p} \pi_{h_j}^{(j)}(x_j)$$

$$\leqslant \sum_{h_s=1}^{k_s} |\pi_{h_s}^{(s)}(x_s) - \tilde{\pi}_{h_s}^{(s)}(x_s)|,$$

where the last step is again by using the second equation in (C.10) and the fact that $\lambda_{h_1 h_2 \ldots h_p}(y) \leqslant 1$. Combining the above inequalities we can obtain (C.7). □

Lemma 79. *Let P and $\tilde{P}$ be two models as in (C.10) with parameters (k, λ, π) and $(\tilde{k}, \tilde{\lambda}, \tilde{\pi})$, respectively. Suppose P is nested in $\tilde{P}$, i.e. satisfying:*

1. *$k_j \leqslant \tilde{k}_j$, for $j = 1, \ldots, p$, ;*
2. *$\lambda_{h_1 \cdots h_p} = \tilde{\lambda}_{h_1 \cdots h_p}$, for $h_j \leqslant k_j, j = 1, \ldots, p$;*
3. *$\pi_{h_j}^{(j)}(x_j) = \tilde{\pi}_{h_j}^{(j)}(x_j)$, for $h_j < k_j$, and $\pi_{k_j}^{(j)}(x_j) = \sum_{h_j \geqslant k_j} \tilde{\pi}_{h_j}^{(j)}(x_j)$.*

Then

$$d(P, \tilde{P}) \leqslant d_0 \sum_{j=1}^{p} \max_{x_j} \sum_{h_j=k_j}^{\tilde{k}_j} \tilde{\pi}_{h_j}^{(j)}(x_j).$$

Proof. By condition (c), P can be considered as model P' of size $\tilde{k}_j$ with $\pi' = \tilde{\pi}$ and λ' satisfying:

$$\lambda'_{h_1 h_2 \cdots h_p}(y) = \lambda_{\min(h_1,k_1),\min(h_2,k_2),\cdots,\min(h_p,k_p)}(y),$$

for $y = 1, \ldots, d_0$ and $h_j \leqslant \tilde{k_j}$, $j = 1, \ldots, p$.

As a result, by condition (b)

$$
\begin{aligned}
&|P(y|x_1, \ldots, x_p) - \tilde{P}(y|x_1, \ldots, x_p)| \\
\leqslant &\sum_{h_1=1}^{\tilde{k}_1} \cdots \sum_{h_p=1}^{\tilde{k}_p} |\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| \prod_{j=1}^{p} \tilde{\pi}_{h_j}^{(j)}(x_j) \\
= &\left\{ \sum_{h_1=1}^{k_1} + \sum_{h_1=k_1+1}^{\tilde{k}_1} \right\} \sum_{h_2=1}^{\tilde{k}_2} \cdots \sum_{h_p=1}^{\tilde{k}_p} |\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| \prod_{j=1}^{p} \tilde{\pi}_{h_j}^{(j)}(x_j) \\
\leqslant &\sum_{h_1=k_1+1}^{\tilde{k}_1} \sum_{h_2=1}^{\tilde{k}_2} \cdots \sum_{h_p=1}^{\tilde{k}_p} |\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| \prod_{j=1}^{p} \tilde{\pi}_{h_j}^{(j)}(x_j) \\
&+ \sum_{h_1=1}^{k_1} \left\{ \sum_{h_2=1}^{k_2} + \sum_{h_2=k_2+1}^{\tilde{k}_2} \right\} \cdots \sum_{h_p=1}^{\tilde{k}_p} |\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| \prod_{j=1}^{p} \tilde{\pi}_{h_j}^{(j)}(x_j) \\
\leqslant &\quad \cdots \\
\leqslant &\sum_{h_1=k_1+1}^{\tilde{k}_1} \sum_{h_2=1}^{\tilde{k}_1} \cdots \sum_{h_p=1}^{\tilde{k}_p} |\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| \prod_{j=1}^{p} \tilde{\pi}_{h_j}^{(j)}(x_j) \\
&+ \cdots + \sum_{h_1=1}^{k_1} \cdots \sum_{h_{p-1}=1}^{k_{p-1}} \sum_{h_p=k_p+1}^{\tilde{k}_p} |\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| \prod_{j=1}^{p} \tilde{\pi}_{h_j}^{(j)}(x_j).
\end{aligned}
$$

Here the last inequality holds because $|\tilde{\lambda}_{\min(h_1,k_1)\cdots\min(h_p,k_p)}(y) - \tilde{\lambda}_{h_1\ldots h_p}(y)| = 0$ if $h_j \leqslant k_j$ for all j. Hence, the lemma can be proved by noticing the constraints (C.10) and the fact that $\tilde{\lambda}_{h_1\ldots h_p}(y) \in [0, 1]$. □

Proof of Theorem 20. We verify conditions (a)-(c) in Theorem 77. As we described previously, we do not need to distinguish the joint probability and the conditional probability under our prior specification. Each model one-to-one corresponds to a triplet (k, λ, π), where $k = (k_1, \ldots, k_{p_n})$ is the multirank, $\lambda = \{\lambda_{h_1,\ldots,h_{p_n}}(y) : y =$

$1, \ldots, d_0, h_j \leqslant k_j, j = 1, \ldots, p_n\}$ is the core tensor and $\pi = \{\pi_{h_j}^{(j)}(x_j) : h_j \leqslant k_j, x_j = 1, \ldots, d_j, j = 1, \ldots, p_n\}$ is the mode matrices. Note that the dimension of λ and π depend on k. Let the sieve $\mathcal{P}_n$ be all conditional probability tensors with multirank satisfying $\prod_{j=1}^{p_n} k_j \leqslant D_n$. Since the inclusion of the jth predictor is equivalent to $k_j > 1$, models in $\mathcal{P}_n$ will depends on at most $\bar{r}_n = \log_2 D_n$ predictors.

Condition (a): By the conclusion of lemma 78, we know that an ϵ_n-net E_n of $\mathcal{P}_n$ can be chosen so that for each $(k, \lambda, \pi) \in \mathcal{P}_n$ that satisfies constraints (C.10), there exists $(\tilde{k}, \tilde{\lambda}, \tilde{\pi}) \in E_n$ such that $\tilde{k} = k$, $\max_{y,h_1,\ldots,h_{p_n}} |\lambda_{h_1 h_2 \ldots h_{p_n}}(y) - \tilde{\lambda}_{h_1 h_2 \ldots h_{p_n}}(y)| < \frac{\epsilon_n}{d_0(\bar{r}_n+1)}$ and $\max_{x_j,h_j} |\pi_{h_j}^{(j)}(x_j) - \tilde{\pi}_{h_j}^{(j)}(x_j)| < \frac{\epsilon_n}{dd_0(\bar{r}_n+1)}$ for j satisfying $k_j > 1$. Hence, for fixed k, we can pick ϵ_n d-balls of the form

$$\prod_{h_1,\ldots,h_{p_n},y} \left(\lambda_{h_1 h_2 \ldots h_{p_n}}(y) \pm \frac{\epsilon_n}{d_0(\bar{r}_n + 1)} \right) \times \prod_{j:k_j>1} \prod_{h_j=1}^{k_j} \prod_{x_j=1}^{d_j} \left(\pi_{h_j}^{(j)}(x_j) \pm \frac{\epsilon_n}{dd_0(\bar{r}_n + 1)} \right),$$

where the first product is taken for all integer vector $(h_1, \ldots, h_{p_n}, y)$ satisfying $1 \leqslant y \leqslant d_0$ and $1 \leqslant h_j \leqslant k_j$. For fixed k with $\prod_{j=1}^{p_n} k_j \leqslant D_n$ in $\mathcal{P}_n$, there are at most $d_0 D_n$ such $\lambda_{h_1 h_2 \ldots h_{p_n}}(y)$'s and $\bar{r}_n d^2$ $\pi_{h_j}^{(j)}(x_j)$'s. Equally spaced grids for λ and π can be chosen so that the union of ϵ_n d-balls centering on the grids covers the set of all models in $\mathcal{P}_n$ with multirank k. Note that there are at most $d\bar{r}_n p_n^{\bar{r}_n}$ different multirank k in $\mathcal{P}_n$. This count follows by first choosing at most $\bar{r}_n$ important predictors with $k_j > 1$, then choosing at most $d\bar{r}_n$ for these k_j's. Hence, the log of the minimal number of size-ϵ_n balls needed to cover $\mathcal{P}_n$ is at most

$$\log \left\{ d\bar{r}_n p_n^{\bar{r}_n} \right\} + d_0 D_n \log \frac{d_0(\bar{r}_n + 1)}{2\epsilon_n} + \bar{r}_n d^2 \log \frac{dd_0(\bar{r}_n + 1)}{2\epsilon_n}.$$

By the conditions in the theorem, each term will be bounded by $n\epsilon_n^2/3$ for n sufficiently large.

Condition (b): Because $\Pi_n(\mathcal{P}_n^c) = 0$ in our case, this condition is trivially satisfied. Actually, this condition will still be satisfied as long as $\Pi_n(\prod_j^{p_n} k_j > D_n) \leqslant$

$\exp\{-(2+C)n\epsilon_n^2\}$, which implies that the prior probability assigned to complex models is exponentially small.

Condition (c): As P_0 is lower bounded away from zero by ϵ_0, $||\log\frac{P}{P_0}||_\infty < \epsilon_n^2$ is implied by $||P-P_0||_\infty < \epsilon_0\epsilon_n^2$ for n large enough ($\epsilon_n \to 0$ as n increases). Let $(\tilde{\lambda}, \tilde{\pi})$ denote parameters for the true model P_0. Consider approximating P_0 by model P with $(k^{(n)}, \lambda, \pi)$, where $k^{(n)}$ is specified in the theorem. Applying lemma 79 to bound $d(\bar{P}, P_0)$, where $\bar{P}$ (regard as the P) with $(k^{(n)}, \bar{\lambda}, \bar{\pi})$ is nested in P_0 (regard as the $\tilde{P}$), and then estimating the difference between P and $\bar{P}$ by lemma 78, we have

$$\begin{aligned} d(P,P_0) &\leqslant \sum_{y=1}^{d_0} \max_{h_1\leqslant k_1^{(n)},\ldots,h_{p_n}\leqslant k_{p_n}^{(n)}} |\lambda_{h_1h_2\ldots h_{p_n}}(y) - \bar{\lambda}_{h_1\ldots\ldots h_{p_n}}(y)| \\ &+ d_0 \sum_{j:k_j^{(n)}>1} \max_{x_j} \sum_{h_j=1}^{k_j^{(n)}} \left|\pi_{h_j}^{(j)}(x_j) - \bar{\pi}_{h_j}^{(j)}(x_j)\right| + d_0 \sum_{j=1}^{p_n} \max_{x_j} \sum_{h_j>k_j^{(n)}} \tilde{\pi}_{h_j}^{(j)}(x_j). \end{aligned} \tag{C.8}$$

Applying (C.7) in lemma 78 and combining (C.8) and condition (iv) in Theorem 20, $||\log\frac{P}{P_0}||_\infty < \epsilon_n^2$ is implied by

$$\max_{h_1\leqslant k_1^{(n)},\ldots,h_{p_n}\leqslant k_{p_n}^{(n)},y} |\lambda_{h_1\ldots h_p}(y) - \bar{\lambda}_{h_1\ldots h_{p_n}}(y)| \prec \frac{\epsilon_n^2}{\bar{r}_n+1},$$

$$\max_{h_j\leqslant k_j^{(n)},x_j} |\pi_{h_j}^{(j)}(x_j) - \bar{\pi}_{h_j}^{(j)}(x_j)| \prec \frac{\epsilon_n^2}{(\bar{r}_n+1)d}.$$

Note that the prior probability $P(k=k^{(n)})$ is at least $(r_n/p_n)^{\bar{r}_n}(r_n/(p_nd))^{\bar{r}_n}(1-r_n/p_n)^{p_n-\bar{r}_n}$. Here $(1-r_n/p_n)^{p_n-\bar{r}_n}$ is defined to be 1 if $r_n=p_n$. As $r_n/p_n \to 0$, $\log\Pi_n(k=k^{(n)})$ is bounded below by $2\bar{r}_n\log(r_n/p_n) \geqslant -2\bar{r}_n\log p_n$.

Moreover, since the $\mathrm{Dir}(1/d_j,\ldots,1/d_j)$ and $\mathrm{Dir}(1/d_0,\ldots,1/d_0)$ priors for $\lambda_{h_1h_2\ldots h_{p_n}}(\cdot)$ and $\pi_\cdot^{(j)}(x_j)$ have density lower bounded away from zero by a constant not involving

n,

$$\log \Pi_n\bigg(P : ||\log \frac{P}{P_0}||_\infty < \epsilon_n^2\bigg)$$

$$> -d_0 D_n \log \frac{\bar{r}_n + 1}{\epsilon_n^2} - \bar{r}_n d^2 \log \frac{(\bar{r}_n + 1)d}{\epsilon_n^2} - 2\bar{r}_n \log p_n.$$

By the assumptions in the theorem, for any $C > 0$, for n sufficiently large, $\log \Pi_n(P : ||\log \frac{P}{P_0}||_\infty < \epsilon_n^2) > -Cn\epsilon_n^2$. □

C.1.5 Proof of Theorem 77

The following two lemmas are needed to prove this theorem. The proof of Lemma 80 can be found in Jiang (2006), and the proof of Lemma 81 follows the line of Ghosal et al. (2000) and is given here.

Lemma 80. *Let $\mathcal{P}$ be a subset of all probability measures of X, $P_0 \in \mathcal{P}$ and d be the total variance distance, then for each $\epsilon > 0$ and $n > 0$, there exists a test ϕ_n such that*

$$P_0^n \phi_n \leqslant N\bigg(\frac{\epsilon}{4}, \mathcal{P}, d\bigg) \exp\bigg(-\frac{n}{8}\epsilon^2\bigg),$$

$$\sup_{P \in \mathcal{P} \cap \{P : d(P, P_0) \geqslant \epsilon\}} P^n(1 - \phi_n) \leqslant \exp\bigg(-\frac{n}{8}\epsilon^2\bigg),$$

where P^n is the $n-$fold of P.

Lemma 81. *If $\Pi_n(P : ||\log \frac{P}{P_0}||_\infty < \epsilon_n^2) > \exp(-Cn\epsilon_n^2)$, then for any test ϕ_n, the following inequality holds:*

$$E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n) \leqslant$$

$$P_0^n \phi_n + \exp((1 + C)n\epsilon_n^2)\Pi_n(\mathcal{P}_n^c) + \exp((1 + C)n\epsilon_n^2) \sup_{\mathcal{P}_n \cap \{P : d(P, P_0) \geqslant \epsilon_n\}} P^n(1 - \phi_n).$$

Proof. We can divide the l.h.s. into two pieces

$$\begin{aligned} & E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n) = \\ & \quad E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n)\phi_n \\ & \quad + E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n)(1 - \phi_n). \end{aligned} \tag{C.9}$$

The first term satisfies

$$E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n)\phi_n \leqslant P_0^n \phi_n. \tag{C.10}$$

Next we will estimate the second term. By definition, we have

$$\begin{aligned} & E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n)(1 - \phi_n) = \\ & \quad E_{P_0} \frac{\int_{d(P,P_0)\geqslant\epsilon_n} \prod_{i=1}^n \frac{P}{P_0}(X_i) d\Pi_n(P)(1 - \phi_n)}{\int \prod_{i=1}^n \frac{P}{P_0}(X_i) d\Pi_n(P)}. \end{aligned} \tag{C.11}$$

Let $K_n = \{P : ||\log \frac{P}{P_0}||_\infty < \epsilon_n^2\}$. Using the condition $\Pi_n(K_n) > \exp(-Cn\epsilon_n^2)$, we have

$$\begin{aligned} \int \prod_{i=1}^n \frac{P}{P_0}(X_i) d\Pi_n(P) &\geqslant \int_{K_n} \prod_{i=1}^n \frac{P}{P_0}(X_i) d\Pi_n(P) \\ &\geqslant \Pi_n(K_n) \exp(-n\epsilon_n^2) \geqslant \exp(-(1 + C)n\epsilon_n^2) \ a.s. P_0^n. \end{aligned}$$

By Fubini's theorem and the fact $0 \leqslant \phi_n \leqslant 1$

$$\begin{aligned} & E_{P_0} \int_{d(P,P_0)\geqslant\epsilon_n} \prod_{i=1}^n \frac{P}{P_0}(X_i) d\Pi_n(P)(1 - \phi_n) \\ & \leqslant \Pi_n(\mathcal{P}_n^c) + \int_{\mathcal{P}_n \cap \{P : d(P,P_0)\geqslant\epsilon_n\}} P^n(1 - \phi_n) d\Pi_n(P) \\ & \leqslant \Pi_n(\mathcal{P}_n^c) + \sup_{\mathcal{P}_n \cap \{P : d(P,P_0)\geqslant\epsilon_n\}} P^n(1 - \phi_n). \end{aligned}$$

Combining the above assertions and equation (C.11), we can see that

$$
\begin{aligned}
&E_{P_0}\Pi_n(P : d(P, P_0) \geqslant \epsilon_n | X_1, \ldots, X_n)(1 - \phi_n) \\
&\leqslant \exp((1 + C)n\epsilon_n^2) E_{P_0} \int_{d(P,P_0)\geqslant\epsilon_n} \prod_{i=1}^{n} \frac{P}{P_0}(X_i) d\Pi_n(P)(1 - \phi_n) \qquad \text{(C.12)} \\
&\leqslant \exp((1 + C)n\epsilon_n^2)\Pi_n(\mathcal{P}_n^c) + \exp((1 + C)n\epsilon_n^2) \sup_{\mathcal{P}_n \cap \{P : d(P,P_0)\geqslant\epsilon_n\}} P^n(1 - \phi_n).
\end{aligned}
$$

Combining (C.9), (C.10) and (C.12) will lead to the conclusion. □

Proof of Theorem 77. Let the test in the Lemma 81 to be the test ϕ_n defined in Lemma 80 with $\epsilon = M\epsilon_n$ and $M^2 > 16 + 8C$. Using the condition (a), (b) in the Theorem 77, we have

$$
\begin{aligned}
&E_{P_0}\Pi_n(P : d(P, P_0) \geqslant M\epsilon_n | X_1, \ldots, X_n) \leqslant \\
&\qquad \exp(-n\epsilon_n^2) + \exp(-n\epsilon_n^2) + \exp(-n\epsilon_n^2) = 3\exp(-n\epsilon_n^2).
\end{aligned}
$$

So

$$
E_{P_0} \sum_n \Pi_n(P : d(P, P_0) \geqslant M\epsilon_n | X_1, \ldots, X_n) \leqslant 3 \sum_n \exp(-n\epsilon_n^2) < \infty.
$$

Thus we have

$$
\sum_n \Pi_n(P : d(P, P_0) \geqslant M\epsilon_n | X_1, \ldots, X_n) < \infty \ a.s. P_0^n,
$$

and

$$
\Pi_n(P : d(P, P_0) \geqslant M\epsilon_n | X_1, \ldots, X_n) \to 0 \ a.s. P_0^n.
$$

□

Appendix D

Appendix for Chapter 5

D.1 Posterior computation

In this appendix, we provide details of the MCMC implementation for CA and LA. The key idea is to augment the weight vector $\lambda = (\lambda_1, \ldots, \lambda_M) \sim \text{Diri}(\rho, \ldots, \rho)$ by $\lambda_j = T_j/(T_1 + \cdots + T_M)$ with $T_j \overset{iid}{\sim} Ga(\rho, 1)$ for $j = 1, \ldots, M$ and conduct Metropolis Hastings updating for $\log T_j$'s. Recall that $F = (F_j(X_i))$ is the $n \times M$ prediction matrix.

D.1.1 Convex aggregation

By augmenting the Dirichlet distribution in the prior for CA, we have the following Bayesian convex aggregation model:

$$Y_i = \sum_{j=1}^{M} \lambda_j F_{ij} + \epsilon_i,\ \epsilon_i \sim N(0, 1/\phi),$$

$$\lambda_j = \frac{T_j}{T_1 + \cdots + T_M},\ T_j \sim Ga(\rho, 1),\ \phi \sim Ga(a_0, b_0).$$

We apply a block MCMC algorithm that iteratively sweeps through the following steps, where superscripts "O", "P" and "N" stand for "old", "proposal" and "new"

respectively:

1. Gibbs updating for ϕ: Updating ϕ by sampling from $[\phi|-] \sim Ga(a_n, b_n)$ with

$$a_n = a_0 + \frac{n}{2}, \quad b_n = b_0 + \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j F_{ij}\right)^2.$$

2. MH updating for $T(\lambda)$: For $j = 1$ to M, propose $T_j^P = T_j^O e^{\beta U_j}$, where $U_j \sim U(-0.5, 0.5)$. Calculate $\lambda_j^P = T_j^P/(\sum_{j=1}^{M} T_j^P)$ and the log acceptance ratio

$$\begin{aligned}\log R =& \frac{\phi}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j^P F_{ij}\right)^2 - \frac{\phi}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j^O F_{ij}\right)^2 \quad \text{(log-likelihood)} \\ &+ \sum_{j=1}^{M}\left((\rho-1)\log T_j^P - T_j^P\right) - \sum_{j=1}^{M}\left((\rho-1)\log T_j^O - T_j^O\right) \quad \text{(log-prior)} \\ &+ \sum_{j=1}^{M}\log T_j^P - \sum_{j=1}^{M}\log T_j^O \quad \text{(log-transition probability)}.\end{aligned}$$

With probability $\min\{1, R\}$, set $T_j^N = T_j^P$, $j = 1, \ldots, M$ and with probability $1 - \min\{1, R\}$, set $T_j^N = T_j^O$, $j = 1, \ldots, M$. Set $\lambda_j^N = T_j^N/(\sum_{j=1}^{M} T_j^N)$, $j = 1, \ldots, M$.

In the above algorithm, β serves as a tuning parameter to make the acceptance rate of T around 40%.

D.1.2 Linear aggregation

By augmenting the double Dirichlet distribution in the prior for LA, we have the following Bayesian linear aggregation model:

$$Y_i = \sum_{j=1}^{M}\theta_j F_{ij} + \epsilon_i,\ \epsilon_i \sim N(0, 1/\phi),\ \theta_j = A z_j \lambda_j,\ \lambda_j = \frac{T_j}{T_1 + \cdots + T_M},$$

$$A \sim Ga(c_0, d_0),\ z_j \sim Bernoulli(0.5),\ T_j \sim Ga(\rho, 1),\ \phi \sim Ga(a_0, b_0).$$

The MCMC updating of T (or equivalently λ) and ϕ is the similar as those in the convex aggregation. In each iteration of the block MCMC algorithm, we add two additional steps for updating z and A:

3. MH updating for A: Propose $A^P = A^O e^{\beta U}$, where $U_j \sim U(-0.5, 0.5)$. Calculate $\lambda_j^P = \lambda_j^O e^{\beta U}$ and the log acceptance ratio

$$\begin{aligned}\log R =& \frac{\phi}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j^P F_{ij}\right)^2 - \frac{\phi}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j^O F_{ij}\right)^2 \quad \text{(log-likelihood)} \\ &+ \left((c_1-1)\log A^P - d_1 A^P\right) - \left((c_1-1)\log A^O - d_1 A^O\right) \quad \text{(log-prior)} \\ &+ \log A^P - \log A^O \quad \text{(log-transition probability)}.\end{aligned}$$

With probability $\min\{1, R\}$, set $A^N = A^P$ and with probability $1 - \min\{1, R\}$, set $A^N = A^O$. Set $\lambda_j^N = \lambda_j^O A^N / A^O$, $j = 1, \ldots, M$.

4. MH updating for z: For $j = 1$ to M, propose $z_j^P = z_j^O V_j$, where $P(V_j = \pm 1) = 0.5$. Calculate $\lambda_j^P = \lambda_j^O V_j$ and the log acceptance ratio

$$\log R = \frac{\phi}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j^P F_{ij}\right)^2 - \frac{\phi}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M}\lambda_j^O F_{ij}\right)^2 \quad \text{(log-likelihood)}.$$

With probability $\min\{1, R\}$, set $z_j^N = z_j^P$, $j = 1, \ldots, M$ and with probability $1 - \min\{1, R\}$, set $z_j^N = z_j^O$, $j = 1, \ldots, M$. Set $\lambda_j^N = \lambda_j^O z_j^P / z_j^O$, $j = 1, \ldots, M$.

D.2 Proofs of technical results in Chapter 5

D.2.1 Proof of Lemma 26

The following lemma suggests that for any $m > 0$, each point in Λ or D_{M-1} can be approximated by an m-sparse point in the same space with error at most $\sqrt{2\kappa/m}$.

Lemma 82. *Fix an integer $m \geqslant 1$. Assume (A1) and (B1).*

a. For any $\lambda^* \in \Lambda$, there exists a $\bar{\lambda} \in \Lambda$, such that $||\bar{\lambda}||_0 \leqslant m$ and $d_\Sigma(\bar{\lambda}, \lambda^*) \leqslant \sqrt{\frac{2\kappa}{m}}$.

b. For any $\eta^* \in D_{M-1}$, there exists an $\bar{\eta} \in D_{M-1}$, such that $||\bar{\eta}||_0 \leqslant m$ and $d_F(\bar{\eta}, \eta^*) \leqslant \sqrt{\frac{2\kappa}{m}}$.

Proof. (Proof of a) Consider a random variable $J \in \{1, \ldots, M\}$ with probability distribution $P(J = j) = \lambda_j^*$, $j = 1, \ldots, M$. Let $J_1, \ldots, J_m$ be m iid copies of J and n_j be the number of $i \in \{1, \ldots, n\}$ such that $(J_i = j)$. Then $(n_1, \ldots, n_M) \sim \text{MN}(m, (\lambda_1^*, \ldots, \lambda_M^*)$, where MN denotes the multinomial distribution. Let $V = (n_1/m, \ldots, n_M/m) \in \Lambda$. Then the expectation $E[V]$ of the vector V is λ^*. Therefore, we have

$$\begin{aligned} Ed_\Sigma^2(V, \lambda^*) &= \sum_{j,k=1}^{M} \Sigma_{jk} E\left(\frac{n_j}{m} - \lambda_j^*\right)\left(\frac{n_k}{m} - \lambda_k^*\right) \\ &= \frac{1}{m}\sum_{j=1}^{M} \Sigma_{jj}\lambda_j^*(1 - \lambda_j^*) - \frac{2}{m}\sum_{1 \leqslant j < k \leqslant M} \Sigma_{jk}\lambda_j^*\lambda_k^* \\ &\leqslant \frac{\kappa}{m}\sum_{j=1}^{M} \lambda_j^*(1 - \lambda_j^*) + \frac{2\kappa}{m}\sum_{1 \leqslant j < k \leqslant M} \lambda_j^*\lambda_k^* \;\leqslant\; \frac{2\kappa}{m}, \end{aligned}$$

where we have used (A1), the fact that $|\Sigma_{jk}| \leqslant \Sigma_{jj}^{1/2}\Sigma_{kk}^{1/2}$ and $\sum_{j=1}^{M} \lambda_j^* = 1$. Since the expectation of $d_\Sigma^2(V, \lambda^*)$ is less than or equal to $2\kappa/m$, there always exists a $\bar{\lambda} \in \Lambda$ such that $d_\Sigma(\bar{\lambda}, \lambda^*) \leqslant \sqrt{2\kappa/m}$.

(Proof of b) The proof is similar to that of a. Now we define $J \in \{1, \ldots, M\}$ as a random variable with probability distribution $P(J = j) = |\eta_j^*|$, $j = 1, \ldots, M$ and let $V = (\text{sgn}(\eta_1^*)n_1/m, \ldots, \text{sgn}(\eta_M^*)n_M/m) \in D_{M-1}$. The rest follows the same line as part a. under assumption (B1). □

Now, we can proceed to prove Lemma 26.

(Proof of a) Without loss of generality, we may assume that the index set of all nonzero components of λ^* is $S_0 = \{1, 2, \ldots, s-1, M\}$. Since $\sup_j \Sigma_{jj} \leqslant \kappa$ and

$|\Sigma_{jk}| \leqslant \Sigma_{jj}^{1/2}\Sigma_{kk}^{1/2} \leqslant \kappa$,

$$d_\Sigma(\lambda, \lambda^*) = \sum_{j,k=1}^{M} \Sigma_{jk}(\lambda_j - \lambda_j^*)(\lambda_k - \lambda_k^*) \leqslant \kappa||\lambda - \lambda^*||_1^2.$$

Therefore, for any $\epsilon > 0$, $\{||\lambda - \lambda^*||_1 \leqslant \kappa^{-1/2}\epsilon\} \subset \{d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon\}$. Since $|\lambda_M - \lambda_M^*| \leqslant \sum_{j=1}^{M-1} |\lambda_j - \lambda_j^*|$, for $\delta_1 = \kappa^{-1/2}\epsilon/(4M - 4s)$ and $\delta_0 = \kappa^{-1/2}\epsilon/(4s)$, we have

$$\begin{aligned}\Lambda_\epsilon = \big\{\lambda \in \Lambda : \lambda_j \in (0, \delta_1], j \in S_0^c;\ |\lambda_j - \lambda_j^*| \leqslant \delta_0, j \in S_0 - \{M\}\big\}& \\ \subset \{||\lambda - \lambda^*||_1 \leqslant \kappa^{-1/2}\epsilon\}.&\end{aligned}$$

Combining the above conclusions yields

$$\begin{aligned}&\Pi(d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon) \ \geqslant\ \Pi(\Lambda_\epsilon) \\ &= \int_{\Lambda_\epsilon} \frac{\Gamma(\alpha/M^{\gamma-1})}{\Gamma^M(\alpha/M^\gamma)} \prod_{j=1}^{M-1} \lambda_j^{\alpha/M^\gamma - 1}\bigg(1 - \sum_{j=1}^{M-1} \lambda_j\bigg)^{\alpha/M^\gamma - 1} d\lambda_1 \cdots d\lambda_{M-1},\end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function. By the facts that $\Gamma(x)\Gamma(1-x) = \pi/\sin(\pi x)$ for $x \in (0,1)$ and $c \triangleq \Gamma'(1)$ is finite, we have $\{x\Gamma(x)\}^{-1} = 1 - cx + O(x^2)$ for $x \in (0, 1/2)$. Combining this with the fact that $\lambda_j \leqslant 1$, we have

$$\begin{aligned}\Pi(d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon) &\geqslant \frac{\Gamma(\alpha/M^{\gamma-1})}{\Gamma^M(\alpha/M^\gamma)} \bigg\{ \prod_{j \in S_0 - \{p\}} \int_{\min\{0, \lambda_j - \delta_0\}}^{\max\{1, \lambda_j + \delta_0\}} \lambda_j^{\alpha/M^\gamma - 1} d\lambda_j \bigg\} \\ &\qquad \bigg\{ \prod_{j \in S_0^c} \int_0^{\delta_1} \lambda_j^{\alpha/M^\gamma - 1} d\lambda_j \bigg\} \\ &\gtrsim \alpha^{-1} M^{\gamma-1} \alpha^M M^{-\gamma M} \delta_0^{s-1} \big(\alpha^{-1} M^\gamma \delta_1^{\alpha/M^\gamma}\big)^{M-s} \\ &\gtrsim \alpha^{s-1} M^{-\gamma(s-1)-1} \bigg(\frac{\epsilon}{s}\bigg)^{s-1} \bigg(\frac{\epsilon}{M-s}\bigg)^{\alpha M^{-(\gamma-1)}(1-s/M)} \\ &\gtrsim\ \exp\bigg\{ -C\gamma s \log M - Cs \log \frac{s}{\epsilon} \bigg\}\ \gtrsim\ \exp\bigg\{ -C\gamma s \log \frac{M}{\epsilon} \bigg\},\end{aligned}$$

where we have used the assumption $\gamma \geqslant 1$ and the fact $s \leqslant M$.

(Proof of b) For any integer $m > 0$, let $\bar{\lambda}$ be the m-sparse approximation of λ^* provided in Lemma 82 part a. Then $d_\Sigma(\bar{\lambda}, \lambda^*) \leqslant Cm^{1/2}$. By the conclusion of Lemma 26 part a, we have

$$\Pi(d_\Sigma(\lambda, \bar{\lambda}) \leqslant \epsilon) \gtrsim \exp\left\{ -C\gamma m \log \frac{M}{\epsilon} \right\}.$$

Therefore, by the triangle inequality, we have

$$\Pi\left(d_\Sigma(\lambda, \lambda^*) \leqslant \epsilon + \frac{C}{\sqrt{m}} \right) \gtrsim \exp\left\{ -C\gamma m \log \frac{M}{\epsilon} \right\}.$$

(Proof of c) For the double Dirichlet distribution, the prior mass allocated to each orthant of $\mathbb{R}^M$ is 2^{-M}. A direct application of part a will result a lower bound of order e^{-CM}, which is too small compare to our conclusion. Therefore, we need to adapt the proof of part a.

Let $S_0 = \{1, 2, \ldots, s-1, M\}$ be the index set of all nonzero components of η^*. Similar to the proof of part a, with the same δ_1 and δ_0 we define

$$\Omega_\epsilon = \big\{ \eta \in D_{M-1} : |\eta_j| \leqslant \delta_1, j \in S_0^c;\ |\eta_j - \eta_j^*| \leqslant \delta_0, j \in S_0 - \{M\} \big\}.$$

Similarly, it can be shown that $\Omega_\epsilon \subset \{d_F(\eta, \eta^*) \leqslant \epsilon\}$. So by the fact that $|\eta_j| \leqslant 1$, we have

$$\begin{aligned}
\Pi(d_F(\eta, \eta^*) \leqslant \epsilon) &\geqslant \frac{1}{2^M} \frac{\Gamma(\alpha/M^{\gamma-1})}{\Gamma^M(\alpha/M^\gamma)} \left\{ \prod_{j \in S_0 - \{p\}} \int_{\eta_j - \delta_0}^{\eta_j + \delta_0} |\eta_j|^{\alpha/M^\gamma - 1} d\eta_j \right\} \\
&\qquad \left\{ \prod_{j \in S_0^c} \int_{-\delta_1}^{\delta_1} |\eta_j|^{\alpha/M^\gamma - 1} d\eta_j \right\} \\
&\gtrsim \frac{1}{2^M} \alpha^{-1} M^{\gamma-1} \alpha^M M^{-\gamma M} (2\delta_0)^{s-1} \big(2\alpha^{-1} M^\gamma \delta_1^{\alpha/M^\gamma} \big)^{M-s} \\
&\gtrsim \alpha^{s-1} M^{-\gamma(s-1)-1} \left(\frac{\epsilon}{s} \right)^{s-1} \left(\frac{\epsilon}{M-s} \right)^{\alpha M^{-(\gamma-1)}(1-s/M)} \\
&\gtrsim \exp\left\{ -C\gamma s \log \frac{M}{\epsilon} \right\}.
\end{aligned}$$

As we can seen, now each η_j contributes an additional factor of 2 to the prior concentration probability comparing to that of λ_j in the proof of part a. This additional factor compensates for the 2^{-M} factor in the normalizing constant of the double Dirichlet distribution.

(Proof of d) The proof is similar to that of part b by instead combining the proof of part c and Lemma 82 part b. Therefore, we omit the proof here.

D.2.2 Proof of Corollary 27

By the triangle inequality and assumption (B1), we have

$$d_F(\lambda, \lambda^*) \leqslant d_F(A\eta, A^*\eta) + d_F(A^*\eta, A^*\eta^*) \leqslant \kappa|A - A^*| + A^* d_F(\eta, \eta^*).$$

As a result, $\{|A - A^*| \leqslant \kappa^{-1}\epsilon;\ d_F(\eta, \eta^*) \leqslant (A^*)^{-1}\epsilon\} \subset \{d_F(\lambda, \lambda^*) \leqslant 2\epsilon\}$ and

$$\Pi(d_F(\lambda, \lambda^*) \leqslant \epsilon) \geqslant \Pi(|A - A^*| \leqslant C\epsilon) \cdot \Pi(d_F(\eta, \eta^*) \leqslant C\epsilon).$$

Since $\log \Pi(|A - A^*| \leqslant C\epsilon) \asymp \log \epsilon$, the conclusions can be proved by applying part c and part d in Lemma 26.

D.2.3 Proof of Lemma 28

(Proof of a) For any $\lambda \in \mathcal{F}^{\Lambda}_{s,\epsilon}$, let $S(\lambda)$ be the index set of the s largest λ_j's. For any $\lambda \in \mathcal{F}^{\Lambda}_{s,\epsilon}$, if $\lambda' \in \Lambda$ satisfies $\lambda'_j = 0$, for $j \in S^c(\lambda)$ and $|\lambda'_j - \lambda_j| \leqslant \epsilon/s$, for $j \in S(\lambda)$, then $d_\Sigma(\lambda, \lambda') \leqslant \kappa||\lambda' - \lambda||_1 \leqslant 2\kappa\epsilon$. Therefore, for a fixed index set $S \subset \{1, \ldots, M\}$ with size s, the set of all grid points in $[0, 1]^s$ with mesh size ϵ/s forms an $2\kappa\epsilon$-covering set for all λ such that $S(\lambda) = S$. Since there are at most $\binom{M}{s}$ such an S, the minimal $2\kappa\epsilon$-covering set for $\mathcal{F}^{\Lambda}_{s,\epsilon}$ has at most $\binom{M}{s} \times \left(\frac{s}{\epsilon}\right)^s$ elements, which implies that

$$\log N(2\kappa\epsilon, \mathcal{F}^{\Lambda}_{s,\epsilon}, ||\cdot||_1) \leqslant \log \binom{M}{s} + s \log \frac{s}{\epsilon} \lesssim s \log \frac{M}{\epsilon}.$$

This proves the first conclusion.

For any $\eta \in \mathcal{F}^{\eta}_{B,s,\epsilon}$, let $S(\eta)$ be the index set of the s largest $|\eta_j|$'s. Similarly, for any $\lambda = A\eta \in \mathcal{F}^{\eta}_{B,s,\epsilon}$, if $\eta' \in D_{M-1}$ satisfies $\eta'_j = 0$, for $j \in S^c(\eta)$ and $|\eta'_j - \eta_j| \leqslant \epsilon/(Bs)$,

for $j \in S(\eta)$, and $A' \leqslant B$ satisfies $|A' - A| \leqslant \epsilon$, then $d_F(A'\eta', A\eta) \leqslant \kappa||A'\eta' - A\eta||_1 \leqslant \kappa|A - A'| + B\kappa||\eta' - \eta||_1 \leqslant 3\kappa\epsilon$. Similar to the arguments for $\mathcal{F}^{\Lambda}_{s,\epsilon}$, we have

$$\log N(3\kappa\epsilon, \mathcal{F}^{\eta}_{B,s,\epsilon}, ||\cdot||_1) \leqslant \log \binom{M}{s} + s \log \frac{Bs}{\epsilon} + \log \frac{B}{\epsilon} \lesssim s \log \frac{M}{\epsilon} + s \log B.$$

(Proof of b) By Lemma 82, any $\lambda \in \Lambda$ and $\eta \in BD_{M-1}$ can be approximated by an m-sparse vector in the same space with error $Cm^{-1/2}$ and $CBm^{-1/2}$ respectively. Moreover, by the proof of Lemma 82, all components of such m-sparse vectors are multiples of $1/m$. Therefore, a minimal $C/\sqrt{m}$-covering set of Λ has at most $\binom{M+m-1}{m-1}$ elements, which is the total number of nonnegative integer solutions $(n_1, \ldots, n_M)$ of the equation: $n_1 + \cdots + n_M = m$. Therefore,

$$\log N(C/\sqrt{m}, \Lambda, d_\Sigma) \leqslant \log \binom{M+m-1}{m-1} \lesssim m \log M,$$

$$\log N(CB/\sqrt{m}, \Lambda, d_\Sigma) \leqslant \log \binom{M+m-1}{m-1} + \log \frac{B}{B/\sqrt{m}} \lesssim m \log M.$$

D.2.4 Proof of Lemma 29

(Proof of a) Consider a random probability P drawn from the Dirichlet process (DP) $DP\big((\alpha/M^{\gamma-1})U\big)$ with concentration parameter $\alpha/M^{\gamma-1}$ and the uniform distribution U on the unit interval $[0, 1]$. Then by the relationship between the DP and the Dirichlet distribution, we have

$$(\lambda_1, \ldots, \lambda_M) \sim \big(P(A_1), \ldots, P(A_M)\big),$$

with $A_k = [(k-1)/M, k/M)$ for $k = 1, \ldots, M$. The stick-breaking representation for DP (Sethuraman, 1994) gives $Q = \sum_{k=1}^{\infty} w_k \delta_{\xi_k}$, a.s. where $\xi_k \overset{iid}{\sim} U$ and

$$w_k = w'_k \prod_{i=1}^{k-1}(1 - w'_i), \text{ with } w'_k \overset{iid}{\sim} \text{Beta}(1, \alpha/M^{\gamma-1}).$$

For each k, let $i(k)$ be the unique index such that $\xi_k \in A_{i(k)}$. Let $\lambda_{(1)} \geqslant \cdots \geqslant \lambda_{(M)}$ be an ordering of $\lambda_1, \ldots, \lambda_M$, then

$$\sum_{j=1}^{s} \lambda_{(j)} \geqslant Q\Big(\bigcup_{j=1}^{s} A_{i(j)}\Big) = \sum_{k:\xi_k \in \bigcup_{j=1}^{s} A_{i(j)}} w_k \geqslant \sum_{k=1}^{s} w_k.$$

Combining the above with the definition of w_k provides

$$\sum_{j=s+1}^{M} \lambda_{(j)} \leqslant 1 - \sum_{k=1}^{s} w_k' \prod_{i=1}^{k-1}(1-w_i') = \prod_{k=1}^{s}(1-w_k') \triangleq \prod_{k=1}^{s} v_k,$$

where $v_k = 1 - w_k' \overset{iid}{\sim} \text{Beta}(\alpha/M^{\gamma-1}, 1)$. Since $v_k \in (0,1)$, we have $(\mathcal{F}_{s,\epsilon}^{\Lambda})^c = \big\{\sum_{j=s+1}^{M} \lambda_{(j)} \geqslant \epsilon\big\} \subset \big\{\prod_{k=1}^{s} v_k \geqslant \epsilon\big\}$. Because

$$Ev_k^s = \int_0^1 \frac{\alpha}{M^{\gamma-1}} t^{\alpha/M^{\gamma-1}+s-1} dt = \frac{\alpha}{\alpha + M^{\gamma-1}s} \leqslant \alpha M^{-(\gamma-1)} s^{-1},$$

an application of Markov's inequality yields

$$\Pi\Big\{\prod_{k=1}^{s} v_k \geqslant \epsilon\Big\} \leqslant \epsilon^{-s} \prod_{k=1}^{s} Ev_k^s \lesssim M^{-s(\gamma-1)} s^{-s} \epsilon^{-s}.$$

As a result,

$$\Pi(\lambda \notin \mathcal{F}_{s,\epsilon}^{\lambda}) \leqslant \Pi\Big\{\prod_{k=1}^{s} v_k \geqslant \epsilon\Big\} \leqslant \exp\Big(-Cs(\gamma-1)\log\frac{M}{\epsilon}\Big).$$

(Proof of b) The proof is similar to that of a since $(|\eta_1|, \ldots, |\eta_M|) \sim (\lambda_1, \ldots, \lambda_M)$ and $\Pi(A > B) \leqslant e^{-CB}$ for $A \sim \text{Ga}(a_0, b_0)$.

D.2.5 Proof of Lemma 30

(Proof of a) Under (A3), the conclusion can be proved by applying Lemma 2.1 and Lemma 4.1 in Kleijn and van der Vaart (2006) by noticing the fact that $||\sum_{j=1}^{M} \lambda_j f_j - f^*||_Q = d_\Sigma(\lambda, \lambda^*)$.

(Proof of b) Let $\psi(\lambda, Y) = \frac{1}{2\sigma^2}||Y - F\lambda||_2^2$. We construct the test function as $\phi_n(Y) = I(\psi(\lambda^*, Y) - \psi(\lambda_2, Y) \geqslant 0)$. By the choice of λ^*, under P_0 we can decomposition the response Y as $Y = F\lambda^* + \zeta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$ and $\zeta = F_0 - F\lambda^* \in \mathbb{R}^d$ satisfying $F^T\zeta = 0$. By Markov's inequality, for any $t < 0$, we have

$$\begin{aligned}
P_{\lambda^*}\phi_n(Y) &= P_{\lambda^*}(e^{t\{\psi(\lambda^*,Y)-\psi(\lambda_2,Y)\}} \geqslant 1) \\
&\leqslant E_{\lambda^*} \exp\left\{\frac{t}{2\sigma^2}\big(||\zeta+\epsilon||_2^2 - ||F(\lambda^*-\lambda_2)+\zeta+\epsilon||_2^2\big)\right\} \\
&= E_{\lambda^*} \exp\left\{\frac{t}{\sigma^2}(\lambda_2-\lambda^*)^T F^T\epsilon\right\} \exp\left\{-\frac{t}{2\sigma^2} n d_F^2(\lambda_2,\lambda^*)\right\} \\
&= \exp\big\{-t(2\sigma^2)^{-1} n d_F^2(\lambda_2,\lambda^*) + t^2\sigma^{-2} n d_F^2(\lambda_2,\lambda^*)\big\}, \\
&= \exp\big\{-(16\sigma^2)^{-1} n d_F^2(\lambda_2,\lambda^*)\big\}, \qquad \text{(D.1)}
\end{aligned}$$

with $t = \frac{1}{4} > 0$, where we have used the fact that $\epsilon \sim N(0, \sigma^2 I_n)$ under P_{λ^*} and $F^T\zeta = 0$. Similarly, for any $\lambda \in \mathbb{R}^M$, under P_λ we have $Y = F\lambda + \epsilon$ with $\epsilon \sim N(0, \sigma^2 I_n)$. Therefore, for any $t > 0$ we have

$$\begin{aligned}
P_\lambda(1-\phi_n(Y)) &= P_\lambda(e^{t\{\psi(\lambda_2,Y)-\psi(\lambda^*,Y)\}} \geqslant 1) \\
&\leqslant E_\lambda \exp\left\{\frac{t}{2\sigma^2}\big(||\epsilon - F(\lambda_2-\lambda)||_2^2 - ||\epsilon - F(\lambda^*-\lambda)||_2^2\big)\right\} \\
&= E_\lambda \exp\left\{-\frac{t}{\sigma^2}(\lambda_2-\lambda^*)^T F^T\epsilon\right\} \qquad \text{(D.2)} \\
&\quad \exp\left\{-\frac{t}{2\sigma^2} n(d_F^2(\lambda,\lambda^*) - d_F^2(\lambda,\lambda_2))\right\} \\
&= \exp\big\{-t(2\sigma^2)^{-1} n\big(d_F^2(\lambda,\lambda^*) - d_F^2(\lambda,\lambda_2)\big) + t^2\sigma^{-2} n d_F^2(\lambda_2,\lambda^*)\big\}, \\
&= \exp\left\{-(16\sigma^2)^{-1} n \frac{\big(d_F^2(\lambda,\lambda^*) - d_F^2(\lambda,\lambda_2)\big)^2}{d_F^2(\lambda_2,\lambda^*)}\right\}, \qquad \text{(D.3)}
\end{aligned}$$

with $t = \frac{1}{4}\big(d_F^2(\lambda,\lambda^*) - d_F^2(\lambda,\lambda_2)\big)/d_F^2(\lambda_2,\lambda^*) > 0$ if $d_F(\lambda,\lambda^*) > d_F(\lambda,\lambda_2)$.

Combining (D.1) and (D.3) yields

$$P_{\lambda^*}\phi_n(Y) \leqslant \exp\big\{-(16\sigma^2)^{-1} n d_F^2(\lambda_2, \lambda^*)\big\}$$

$$\sup_{\lambda \in \mathbb{R}^M:\ d_F(\lambda,\lambda_2) < \frac{1}{4} d_F(\lambda^*,\lambda_2)} P_\lambda(1 - \phi_n(Y)) \leqslant \exp\big\{-(64\sigma^2)^{-1} n d_F^2(\lambda_2, \lambda^*)\big\}.$$

Appendix E

Appendix for Chapter 6

E.1 Proofs of technical results in Chapter 6

E.1.1 Proof of Lemma 31

Consider a fixed s. Let $A_s = \{x : |x - \mu_s| \leqslant d_0\}$. Then by the definition of total variation norm and $||\hat{\pi}_t - \pi_t|| \to 0$, we have $|\hat{\pi}_t(A_s) - \pi_t(A_s)| \to 0$ as $t \to \infty$. Because $\int h^{(s)}(x)\lambda(dx) = 1$ and $\delta_t \to 0$, $h_t^{(s)}(A_s^c) \to 0$ uniformly for $s \in \{1, \ldots, S\}$ as $t \to \infty$. As a result, $\lim_{t\to\infty} |\hat{\pi}_t(A_s) - w_s^{(s)}| = 0$. By the weak law of large numbers, $|\hat{w}_t^{(s)} - \hat{\pi}_t(A_s)| \to 0$ in probability as $L \to \infty$. Combining the above, we reach the conclusion.

E.1.2 Proof of Lemma 42

There exists a short proof for this lemma and Lemma 43, which is based on the operator theory and can be considered as a special case of the proof for Lemma 39 with $V \equiv 1$. But for illustration and possible future generalization, we also provide the following proof based on coupling technique.

Denote $\delta = \frac{1}{2}||p_0 - \pi||$. Let $\{X_t : t \geqslant 0\}$ and $\{X_t' : t \geqslant 0\}$ be two Markov chains

defined as follows:

1. $X_0 \sim p_0$;

2. Given $X_0 = x$, with probability $\min\{1, \frac{\pi(x)}{p_0(x)}\}$, set $X_0' = x$; with probability $1 - \min\{1, \frac{\pi(x)}{p_0(x)}\}$, draw

$$X_0' \sim \frac{\pi(\cdot) - \min\{\pi(\cdot), p_0(\cdot)\}}{\delta};$$

3. For $t \geqslant 1$, if $X_0 = X_0'$, draw $X_t = X_t' \sim T(X_{t-1}, \cdot)$, else draw X_t and X_t' independently from $X_t \sim T(X_{t-1}, \cdot)$ and $X_t' \sim T(X_{t-1}', \cdot)$ respectively.

Note that $\frac{\pi(\cdot) - \min\{\pi(\cdot), p_0(\cdot)\}}{\delta}$ is a valid probability density since: 1. it is nonnegative; 2. its integral on E is equal to one by the definition of δ.

From the above construction, it is easy to see that the marginal distribution of X_t is $T^t \circ p_0$. Next we will prove that the marginal distribution of X_t' is π for all t. Since the stationary distribution of T is π, we only need to show that the marginal distribution of X_0' is π. First,

$$\begin{aligned} P(X_0 = X_0') &= \int \min\{1, \frac{\pi(x)}{p_0(x)}\} p_0(x) \lambda(dx) \\ &= \int \min\{p_0(x), \pi(x)\} \lambda(dx) \\ &= 1 - \delta. \end{aligned} \tag{E.1}$$

Then, for any $A \in \mathcal{B}(\mathcal{E})$,

$$\begin{aligned}
P(X_0' \in A) =& P(X_0' \in A, X_0 \neq X_0') + P(X_0' \in A, X_0 = X_0') \\
=& P(X_0' \in A | X_0 \neq X_0') P(X_0 \neq X_0') \\
&+ \int_A P(X_0 = X_0' | X_0' = x) P(X_0' = x) \lambda(dx) \\
=& \delta \int_A \frac{\pi(x) - \min\{\pi(x), p_0(x)\}}{\delta} \lambda(dx) \ + \int_A \min\left\{1, \frac{\pi(x)}{p_0(x)}\right\} p_0(x) \lambda(dx) \\
=& \int_A \pi(x) \lambda(dx).
\end{aligned}$$

By uniform ergodicity, for any probability measure p, we have

$$\begin{aligned}
||T^t \circ p - \pi|| =& \int \left| \int T^t(z, x) p(z) \lambda(dz) - \int \pi(x) p(z) \lambda(dz) \right| \lambda(dx) \\
\leqslant& \int ||T^t(z, \cdot) - \pi(\cdot)|| p(z) \lambda(dz) \\
\leqslant& r(t).
\end{aligned} \tag{E.2}$$

By the above inequality, (E.1) and our construction of X_t and X_t', for any $A \in \mathcal{B}(\mathcal{E})$, we have

$$\begin{aligned}
|T^t \circ p_0(A) - \pi(A)| =& |P(X_t \in A) - P(X_t' \in A)| \\
=& |P(X_0 \neq X_0', X_t \in A) - P(X_0 \neq X_0', X_t' \in A)| \\
\leqslant& P(X_0 \neq X_0') \left\{ |P(X_t \in A | X_0 \neq X_0') - \pi(A)| \right. \\
&\left. + |P(X_t' \in A | X_0 \neq X_0') - \pi(A)| \right\} \\
\leqslant& \delta r(t),
\end{aligned}$$

where the last line follows by the fact that $||p - q|| = 2 \sup_A |p(A) - q(A)|$ and (E.2) with $p(\cdot) = P(X_0 = \cdot | X_0 \neq X_0')$ and $p(\cdot) = P(X_0' = \cdot | X_0 \neq X_0')$. Therefore,

$$||T^t \circ p_0 - \pi|| = 2 \sup_A |T^t \circ p_0(A) - \pi(A)| \leqslant r(t) ||p_0 - \pi||.$$

E.1.3 Proof of Lemma 43

Let $\delta(x) = \frac{1}{2}||T(x,\cdot) - \pi|| \leqslant \rho$. Given an initial point x, we can construct two Markov chains $\{X_t : t \geqslant 0\}$ and $\{X'_t : t \geqslant 0\}$ as follows:

1. $X_0 = x$, $X'_0 \sim \pi$;
2. For $t \geqslant 1$, given $X_{t-1} = x$ and $X'_{t-1} = x'$,
   (a) if $x = x'$, choose $X_t = X'_t \sim T(x,\cdot)$;
   (b) else, first choose $X'_t = y \sim T(x',\cdot)$, then with probability $\min\{1, \frac{T(x,y)}{\pi(y)}\}$, set $X_t = y$, with probability $1 - \min\{1, \frac{T(x,y)}{\pi(y)}\}$, draw

$$X_t \sim \frac{T(x,\cdot) - \min\{T(x,\cdot), \pi(\cdot)\}}{\delta(x)};$$

Then similar to the proof of Lemma 42, the above procedure is valid and the two Markov chains X_t and X'_t have the same transition kernel T, but have initial distribution δ_x and π, respectively. Moreover,

$$\begin{aligned} P(X_t \neq X'_t | X_1, X'_1, \ldots, X_{t-1}, X'_{t-1}) &\leqslant \sup_x \left\{ 1 - \int \min\{1, \frac{T(x,y)}{\pi(y)}\} \pi(y) \lambda(dy) \right\} \\ &= \sup_x \delta(x) \leqslant \rho. \end{aligned}$$

Therefore, we have

$$||T^t(x,\cdot) - \pi|| \leqslant P(X_1 \neq X'_1, \ldots, X_t \neq X'_t) \leqslant \rho^t.$$

E.1.4 Proof of Lemma 44

By Lemma 42, T is uniformly ergodic. Therefore by Theorem 1.3 in Mengersen and Tweedie (1996), (6.15) holds for some $\rho \in (0,1)$ and probability measure ν. Then by the arguments after Lemma 43, (6.16) holds with the same ρ.

E.1.5 Proof of Lemma 45

We construct a new Markov chain $\{\tilde{X}_t : t \geqslant 0\}$ as follows:

1. The state space of $\tilde{X}_t$ is $\tilde{E} = E \cup \{c\}$, where c is an extended "coffin" state.
2. For $t > 0$: if $\tilde{X}_{t-1} \neq c$, then with probability $\delta(x)$, $\tilde{X}_t = X'_t$ and with probability $1 - \delta(x)$, $\tilde{X}_t = c$; if $\tilde{X}_{t-1} = c$, then $\tilde{X}_t = c$. Therefore, c is an absorbing state.
3. $\tilde{X}_0$ is distributed according to p_0.

Then by identifying the coupling $(X_t = X'_t)$ in the proof of Lemma 43 as going to the absorbing state c, we have

$$||T^t \circ p_0 - \pi|| \leqslant P(\tilde{X}_t \neq c) = E\Big\{ \prod_{s=1}^{t} \delta(X'_s) \Big\},$$

since before being coupled, X'_t in the proof of Lemma 43 is a Markov chain with transition kernel T'.

E.1.6 Proof of Theorem 32

We will construct two time inhomogeneous Markov chains $\{X_{t,s} : s = 1, \ldots, m_t,\ t \geqslant 0\}$ and $\{X'_{t,s} : s = 1, \ldots, m_t,\ t \geqslant 0\}$, where a double index is used as the step indicator under the following order $(0,1) \to \cdots \to (0, m_0) \to (1,1) \to \cdots \to (1, m_1) \to (2,1) \to \cdots \to (2, m_2) \to \cdots$. Let $\delta_t(x) = \frac{1}{2}||T_t(x, \cdot) - \pi_t||$. The two chains are constructed as follows: (note that $m_0 = 1$)

1. $X_{0,1} \sim \pi_0$, $X'_{0,1} \sim \pi_0$;
2. For $t \geqslant 1$,
   (a) $s = 1$. Let $X_{t-1,m_{t-1}} = x$ and $X'_{t-1,m_{t-1}} = x'$. Set $X_{t,1} = x$. With probability $\min\{1, \frac{\pi_t(x)}{\pi_{t-1}(x)}\}$, set $X'_{t,1} = x$; with probability $1 - \min\{1, \frac{\pi_t(x)}{\pi_{t-1}(x)}\}$,

draw

$$X'_{t,1} \sim \frac{\pi_t(\cdot) - \min\{\pi_t(\cdot), \pi_{t-1}(\cdot)\}}{\alpha_t};$$

(b) $1 < s \leqslant m_t$. Let $X_{t,s-1} = x$ and $X'_{t,s-1} = x'$.

i. if $x = x'$, choose $X_{t,s} = X'_{t,s} \sim T_t(x, \cdot)$;

ii. else, first choose $X'_{t,s} = y \sim T_t(x', \cdot)$, then with probability $\min\{1, \frac{T_t(x,y)}{\pi_t(y)}\}$, set $X_{t,s} = y$, with probability $1 - \min\{1, \frac{T_t(x,y)}{\pi_t(y)}\}$, draw

$$X_{t,s} \sim \frac{T_t(x, \cdot) - \min\{T_t(x, \cdot), \pi_t(\cdot)\}}{\delta_t(x)};$$

The above construction combines those in Lemma 42 and 43. By the argument in the proof of Lemma 42, the construction for $s = 1$ is valid. Moreover, if $(X_{t-1,m_{t-1}} = X'_{t-1,m_{t-1}})$, then the probability of $(X_{t,1} \neq X'_{t,1})$ is α_t. Similarly, by the argument in the proof of Lemma 43, the construction for $s > 1$ is valid. Moreover, conditioning on $X_{t,s-1}$ and $X'_{t,s-1}$, the conditional probability of $(X_{t,1} \neq X'_{t,1})$ does not exceed ρ_t.

It can been seen that the marginal distribution of $X_{t,s}$ is $T_t^s \circ Q_{t-1} \circ \cdots \circ Q_1 \circ \pi_0$, while the marginal distribution of $X'_{t,s}$ is π_t, for $s = 1, \ldots, m_t$. Therefore,

$$||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t|| \leqslant P(X_{t,m_t} \neq X'_{t,m_t}).$$

Furthermore, we have

$$\begin{aligned}
P(X_{t,m_t} \neq X'_{t,m_t}) &= P(X_{t-1,m_{t-1}} \neq X'_{t-1,m_{t-1}}, X_{t,m_t} \neq X'_{t,m_t}) \\
&\quad + P(X_{t-1,m_{t-1}} = X'_{t-1,m_{t-1}}, X_{t,m_t} \neq X'_{t,m_t}) \\
&\leqslant P(X_{t-1,m_{t-1}} \neq X'_{t-1,m_{t-1}})\rho_t^{m_t} \\
&\quad + \big[1 - P(X_{t-1,m_{t-1}} \neq X'_{t-1,m_{t-1}})\big]\alpha_t\rho_t^{m_t} \\
&= \alpha_t\epsilon_t + \epsilon_t(1 - \alpha_t)P(X_{t-1,m_{t-1}} \neq X'_{t-1,m_{t-1}}) \\
&\leqslant \cdots \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s+1}^{t} \epsilon_u(1 - \alpha_u) \right\} \epsilon_s\alpha_s.
\end{aligned}$$

Combining the above two inequalities, the theorem can be proved.

E.1.7 Proof of Lemma 33

Under the regularization conditions on f_θ, the following second order Berstein Von-Mises theorem holds (for a proof, see, for example, Datta and Mukerjee (2004)):

$$\left|\left|\pi_t - N\left(\hat{\theta}_{n_t}, \frac{1}{n_t} I^{-1}\right)\right|\right| = O_P\left(\frac{1}{\sqrt{n_t}}\right), \tag{E.3}$$

where $N(\mu, \Sigma)$ is the multivariate normal distribution with mean μ and covariance Σ, $\hat{\theta}_{n_t}$ is the maximum likelihood estimator, π_t is the posterior distribution with n_t observations $Y_1, \ldots, Y_{n_t}$ and I is the Fisher information matrix.

Next, we show that under the same regularity conditions, $|\hat{\theta}_{n_t} - \hat{\theta}_{n_{t-1}}| = O_P\big(\frac{\sqrt{\Delta_t}}{n_t}\big)$. In fact, an application of Taylor's expansion for $\sum_{i=1}^{n_t} \dot{l}(Y_i, \theta)$ around $\hat{\theta}_{n_{t-1}}$ yields that for θ in a small neighborhood around $\hat{\theta}_{n_{t-1}}$,

$$\sum_{i=1}^{n_t} \dot{l}(Y_i, \theta) - \sum_{i=1}^{n_t} \dot{l}(Y_i, \hat{\theta}_{n_{t-1}}) = \sum_{i=1}^{n_t} \ddot{l}(Y_i, \hat{\theta}_{n_{t-1}})(\theta - \hat{\theta}_{n_{t-1}}) + O\big(n_t(\theta - \hat{\theta}_{n_{t-1}})^2\big)$$

Plugging in θ with $\hat{\theta}_{n_t}$ and using the facts that $\sum_{i=1}^{n_t} \dot{l}(Y_i, \hat{\theta}_{n_t}) = 0$, $\sum_{i=1}^{n_{t-1}} \dot{l}(Y_i, \hat{\theta}_{n_{t-1}}) = 0$ and $\sum_{i=1}^{n_t} \ddot{l}(Y_i, \hat{\theta}_{n_{t-1}}) \to n_t I$ in probability, we obtain

$$-\sum_{i=0}^{\Delta_t - 1} \dot{l}(Y_{n_t - i}, \hat{\theta}_{n_{t-1}}) = nI(\hat{\theta}_{n_t} - \hat{\theta}_{n_{t-1}}) + o_P\big(n_t|\hat{\theta}_{n_t} - \hat{\theta}_{n_{t-1}}|\big). \tag{E.4}$$

Finally we reach

$$|\hat{\theta}_{n_t} - \hat{\theta}_{n_{t-1}}| = -[1 + o_P(1)]\ (n_t I)^{-1} \sum_{i=0}^{\Delta_t - 1} \dot{l}(Y_{n_t - i}, \hat{\theta}_{n_{t-1}}) = O_P\left(\frac{\sqrt{\Delta_t}}{n_t}\right).$$

Return to the proof of the theorem. Note that the L_1 distance $||p - q||$ between any two densities p and q is bounded by $H(p, q)/\sqrt{2}$, where $H^2(p, q) = \int |\sqrt{p} - \sqrt{q}|^2$ is the square of the Hellinger distance. Moreover, for two normal distributions, $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, we have

$$H^2\big(N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)\big) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}} e^{-\frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}}.$$

Therefore, by combining (E.3) and (E.4), we have

$$\begin{aligned}
||\pi_t - \pi_{t-1}|| &\leqslant \left|\left|\pi_t - N\left(\hat{\theta}_{n_t}, \frac{1}{n_t} I^{-1}\right)\right|\right| + \left|\left|\pi_{t-1} - N\left(\hat{\theta}_{n_{t-1}}, \frac{1}{n_{t-1}} I^{-1}\right)\right|\right| \\
&\quad + \left|\left|N\left(\hat{\theta}_{n_t}, \frac{1}{n_t} I^{-1}\right) - N\left(\hat{\theta}_{n_{t-1}}, \frac{1}{n_{t-1}} I^{-1}\right)\right|\right| \\
&= O_P\left(\frac{1}{\sqrt{n_t}}\right) + O_P\left(\frac{1}{\sqrt{n_t - \Delta_t}}\right) \\
&\quad + \left(1 - \frac{{n_t}^{1/4} n_{t-1}^{1/4}}{(n_t - \Delta_t/2)^{1/2}} e^{-n_t O_P(\Delta_t / n_t^2)}\right)^{1/2} \\
&= O_P\left(\frac{1}{\sqrt{n_t}}\right) + O_P\left(\sqrt{\frac{\Delta_t}{n_t}}\right) = O_P\left(\sqrt{\frac{\Delta_t}{n_t}}\right).
\end{aligned}$$

E.1.8 Proof of Lemma 34

Without loss of generality, we consider one dimensional case because otherwise, we can estimate each component of $\tilde{\theta}_j$ by considering the marginal distribution of $\hat{\pi}$ along the jth dimension. Combining (E.3) in the proof of Lemma 33 (under the same notation) and the assumption $||\hat{\pi}_t - \pi_t|| \leqslant \varepsilon$, we have

$$\left|\left|\hat{\pi}_t - N\left(\hat{\theta}_{n_t}, \frac{1}{n_t} I^{-1}\right)\right|\right| = \iota + \varepsilon, \tag{E.5}$$

where $\iota = O_P(n_t^{-1/2})$. Let $\tilde{\theta}_t$ be the median of $\hat{\pi}_t$. By the definition of the total variation norm $||\cdot||$ and $\hat{\pi}_t\big(\theta \in (-\infty, \tilde{\theta}_t]\big) = 0.5$, we obtain

$$\left|\Phi\big(\sqrt{n_t} I(\tilde{\theta}_t - \hat{\theta}_{n_t})\big) - \frac{1}{2}\right| = \iota + \varepsilon = \Phi(z_{0.5+\iota+\varepsilon}) - \frac{1}{2} = \frac{1}{2} - \Phi(z_{0.5-\iota-\varepsilon}),$$

where Φ is the cdf of the standard normal distribution. Therefore $|\tilde{\theta}_t - \hat{\theta}_{n_t}| = O_P\big(z_{0.5+\iota+\varepsilon} n_t^{-1/2}\big)$. Because f_θ is regular, the MLE satisfies $|\hat{\theta}_{n_t} - \theta_0| = O_P\big(n_t^{-1/2}\big)$. As a result, $|\hat{\theta}_t - \theta_0| = O_P\big(z_{0.5+\iota+\varepsilon} n_t^{-1/2}\big)$, which completes the first part.

For the second part, we do not restrict $\theta \in \mathbb{R}^d$ to be one dimensional. An application of E.5 with A_α, we obtain

$$\left|P_0\big(\theta_0 + n_t^{-1/2}\hat{\Delta}_{n_t} + n_t^{-1/2} I^{-1/2} N \in A_\alpha\big) - \alpha\right| = \iota + \varepsilon,$$

where N is a random vector that follows $N(0, I_d)$ with I_d the d dimensional identity matrix and $\hat{\Delta}_{n_t} = n_t^{1/2}(\hat{\theta}_{n_t} - \theta_0) \to N(0, I^{-1})$ in distribution. Therefore, for

$$B_t = n_t^{1/2} I^{1/2}(A_\alpha - \theta_0 - n_t^{-1/2}\hat{\Delta}_{n_t}),$$

we have $P(N \in B_t) = \alpha + O_P(\iota) + O_P(\varepsilon)$. Therefore,

$$\begin{aligned} P_0(\theta_0 \in A_\alpha) &= P_0(\theta_0 \in \theta_0 + n_t^{-1/2}\hat{\Delta}_{n_t} + n_t^{-1/2} I^{-1/2} B_t) \\ &= P_0(I^{1/2}\hat{\Delta}_{n_t} \in -B_t) \\ &= P(N \in -B_t) + O({n_t}^{-1/2}) \\ &= \alpha + O_P(\iota) + O_P(\varepsilon), \end{aligned}$$

where the third step follows by the fact that $I^{1/2}\hat{\Delta}_{n_t} \to N(0, I_d)$ in distribution and the Edgeworth expansion, and the last step follows by the symmetry of the distribution of $N(0, I_k)$.

E.1.9 Proof of Lemma 35

By factorization of joint probability, we have

$$\begin{aligned} ||\pi_t - J_t \circ \hat{\pi}_{t-1}|| &= \int |\pi_t(\theta^{(t-1)})\pi_t(\eta_t|\theta^{(t-1)}) - p(\theta^{(t-1)})J_t(\theta^{(t-1)}, \eta_t)| d\theta^{(t-1)} \lambda(d\eta_t) \\ &\leqslant \int \pi_t(\theta^{(t-1)})|\pi_t(\eta_t|\theta^{(t-1)}) - J_t(\theta^{(t-1)}, \eta_t)| \lambda(d\eta_t)\lambda(d\theta^{(t-1)}) \\ &\quad + \int |\pi_t(\theta^{(t-1)}) - p(\theta^{(t-1)})| J_t(\theta^{(t-1)}, \eta_t) \lambda(d\theta^{(t-1)})\lambda(d\eta_t) \\ &\leqslant \sup_{\theta^{(t-1)} \in \mathbb{R}^{d_{t-1}}} ||\pi_t(\cdot|\theta^{(t-1)}) - J_t(\theta^{(t-1)}, \cdot)|| + ||\pi_t - p||. \end{aligned}$$

E.1.10 Proof of Theorem 36

The proof is almost the same as that of Theorem 32. The only difference occurs in the constructions of $X_{t,s}$ and $X'_{t,s}$ for $s = 1$, which is provided in the following.

When $t \geqslant 1$ and $s = 1$, let $X_{t-1,m_{t-1}} = x$ and $X'_{t-1,m_{t-1}} = x'$. Draw $X_{t,1} \sim J_t(x, \cdot)$. With probability $\min\{1, \frac{\pi_t(x)}{J_t \circ \pi_{t-1}(x)}\}$, set $X'_{t,1} = x$; with probability $1 - \min\{1, \frac{\pi_t(x)}{J_t \circ \pi_{t-1}(x)}\}$, draw

$$X'_{t,1} \sim \frac{\pi_t(\cdot) - \min\{\pi_t(\cdot), J_t \circ \pi_{t-1}(x)(\cdot)\}}{\tilde{\alpha}_t},$$

where $\tilde{\alpha}_t = \frac{1}{2}||\pi_t - J_t \circ \pi_{t-1}||$ is the probability of $(X_{t,1} \neq X'_{t,1})$ conditioning on $(X_{t-1,m_{t-1}} = X'_{t-1,m_{t-1}})$. Moreover, by Lemma 35, we have $\tilde{\alpha}_t \leqslant \alpha_t + \tau_t$.

E.1.11 Proof of Lemma 37

Recall that the Kullback-Leibler (KL) divergence is defined as

$$K(p, q) = \int p(\theta) \log \frac{p(\theta)}{q(\theta)} m(d\theta),$$

where f and g are two pdfs on Θ. We will use the following relationship between KL divergence and L_1 norm:

$$||p - q|| \leqslant 2\sqrt{K(p, q)}. \tag{E.6}$$

Use the shorthand π_n for the posterior density $\pi(\cdot|Y_1, \ldots, Y_n)$ for θ. By definition,

$$\pi_n(\theta) = \frac{\exp\{\sum_{i=1}^n l_i(\theta)\}\pi(\theta)}{\int_\Theta \exp\{\sum_{i=1}^n l_i(\theta)\}\pi(\theta)m(d\theta)},$$

where $l_i(\theta) = \log p_\theta(Y_i)$ is the log likelihood for the ith observation and π is the prior for θ. Moreover,

$$\begin{aligned}
\log \frac{\pi_{n-1}(\theta)}{\pi_n(\theta)} &= -l_n(\theta) + \log \left\{ \int_\Theta \frac{\exp\{\sum_{i=1}^{n-1} l_i(\theta)\}\pi(\theta)}{\int_\Theta \exp\{\sum_{i=1}^{n-1} l_i(\theta)\}\pi(\theta)d\theta} \exp\{l_n(\theta)\} m(d\theta) \right\} \\
&= -l_n(\theta) + \log E_{[\theta|Y_1,\ldots,Y_{n-1}]} \exp\{l_n(\theta)\},
\end{aligned}$$

where $E_{[\theta|Y_1,\ldots,Y_{n-1}]}$ is the expectation with respect to the posterior distribution $\Pi(\theta|Y_1,\ldots,Y_{n-1})$. Therefore, we obtain:

$$\begin{aligned} K(\pi_{n-1},\pi_n) =& \int_\Theta \pi_{n-1}(\theta)\log\frac{\pi_{n-1}(\theta)}{\pi_n(\theta)}m(d\theta) \\ =& \log E_{[\theta|Y_1,\ldots,Y_{n-1}]}\exp\{l_n(\theta)\} - E_{[\theta|Y_1,\ldots,Y_{n-1}]}\{l_n(\theta)\}. \end{aligned} \tag{E.7}$$

By the third condition, we have that for any $||\theta-\theta_0|| \leqslant M\epsilon_n$,

$$l_n(\theta) = l_n(\theta_0) + O_{P_{\theta_0}}(\epsilon_n).$$

Combining the above with the second condition, we have

$$\begin{aligned} E_{[\theta|Y_1,\ldots,Y_{n-1}]}\exp\{l_n(\theta)\} =& E_{[\theta|Y_1,\ldots,Y_{n-1}]}\big\{\exp\{l_n(\theta)\}I(||\theta-\theta_0|| \leqslant M\epsilon_n)\big\} + o_{P_{\theta_0}}(1) \\ =& \exp\{l_n(\theta_0)\} + o_{P_{\theta_0}}(1). \end{aligned}$$

Similarly, we have

$$E_{[\theta|Y_1,\ldots,Y_{n-1}]}\{l_n(\theta)\} = l_n(\theta_0) + o_{P_{\theta_0}}(1).$$

Combining the above two with (E.6) and (E.7), we obtain

$$||\pi(\cdot|Y_1,\ldots,Y_n) - \pi(\cdot|Y_1,\ldots,Y_{n-1})|| \to 0, \text{ as } n\to\infty.$$

E.1.12 Proof of Lemma 39

For a kernel $K(x,y)$ on $E\times E$, we define

$$|||K|||_V = \sup_{x\in\mathbb{R}^d}\frac{||K(x,\cdot)||_V}{V(x)} = \sup_{x\in\mathbb{R}^d}\sup_{|f|\leqslant V}\frac{|(Kf)(x)|}{V(x)}.$$

It is easy to verify that $|||\cdot|||_V$ satisfies the triangle inequality. By viewing $\pi(x,y) = \pi(y)$ as a kernel on $E\times E$, we have $|||T-\pi|||_V \leqslant \rho$. Moreover, for any $t\in\mathbb{N}$, we have,

$$\begin{aligned} |||T^t-\pi|||_V =& \sup_{x\in E}\sup_{|f|\leqslant V}\frac{\big|\{(T-\pi)(T^{t-1}-\pi)f\}(x)\big|}{V(x)} \\ =& |||T^{t-1}-\pi|||_V \sup_{x\in E}\sup_{|f|\leqslant V}\frac{|\{(T-\pi)g_f\}(x)|}{V(x)}, \end{aligned}$$

with $g_f(x) = \{(T^{t-1} - \pi)f\}(x)/|||T^{t-1} - \pi|||_V$. By the definition of $|||\cdot|||_V$, we have $|g_f| \leqslant V$ for any f satisfying $|f| \leqslant V$. Combining the above arguments, we obtain

$$\begin{aligned}|||T^t - \pi|||_V \leqslant & |||T^{t-1} - \pi|||_V \cdot |||T - \pi|||_V \\ \leqslant & \rho|||T^{t-1} - \pi|||_V \\ \leqslant & \cdots \leqslant \rho^t.\end{aligned}$$

This implies geometric ergodicity, i.e.

$$||T^t(x, \cdot) - \pi(\cdot)||_V \leqslant V(x)\rho^t, \quad x \in E, \quad t \in \mathbb{N}.$$

For the second part, by the stationarity of π, we have

$$\begin{aligned}||T^t \circ p_0 - \pi||_V = & \sup_{|f| \leqslant V} \int_{\mathbb{X}} \{p_0(x) - \pi(x)\}\{(T^t - \pi)f\}(x)\lambda(dx) \\ \leqslant & \int_{\mathbb{X}} |p_0(x) - \pi(x)|\ V(x) \sup_{|f| \leqslant V} \frac{|\{(T^t - \pi)f\}(x)|}{V(x)}\lambda(dx) \\ \leqslant & \rho^t ||p_0 - \pi||_V.\end{aligned}$$

E.1.13 Proof of Theorem 40

By Lemma 39, for any distribution p_0 on $\mathbb{R}^d$ and any $t \in \mathbb{N}$, we have

$$||T_t^{m_t} \circ p_0 - \pi_t||_V \leqslant \rho_t^{m_t} ||p_0 - \pi_t||_V.$$

Therefore, we have

$$\begin{aligned}||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t||_V \leqslant & \epsilon_t ||Q_{t-1} \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t||_V \\ \leqslant & \epsilon_t ||Q_{t-1} \circ \cdots \circ Q_1 \circ \pi_0 - \pi_{t-1}||_V + \epsilon_t ||\pi_t - \pi_{t-1}||_V.\end{aligned}$$

By Cauchy's inequality,

$$\begin{aligned}||\pi_t - \pi_{t-1}||_V = & \int_{\mathbb{R}^d} |\pi_t(x) - \pi_{t-1}(x)| V(x)\lambda(dx) \\ \leqslant & d_H(\pi_t, \pi_{t-1}) \left[\int_{\mathbb{R}^d} \{\pi_t^{1/2}(x) + \pi_{t-1}^{1/2}(x)\}^2 V^2(x)\lambda(dx) \right]^{1/2} \\ \leqslant & 2\sqrt{C} d_H(\pi_t, \pi_{t-1}) = \alpha_t.\end{aligned}$$

Combining the above two inequalities, we obtain

$$\begin{aligned}||Q_t \circ \cdots \circ Q_1 \circ \pi_0 - \pi_t||_V \leqslant & \epsilon_t ||Q_{t-1} \circ \cdots \circ Q_1 \circ \pi_0 - \pi_{t-1}||_V + \alpha_t \epsilon_t \\ \leqslant & \cdots \leqslant \sum_{s=1}^{t} \left\{ \prod_{u=s}^{t} \epsilon_u \right\} \alpha_s.\end{aligned}$$

Finally, the theorem can be proved by noticing that

$$||\mu|| = \sup_{||f|| \leqslant 1} |\mu(f)| \leqslant \sup_{||f|| \leqslant V} |\mu(f)| = ||\mu||_V$$

for any signed measure μ.

Appendix F

Appendix for Chapter 7

F.1 Proofs of technical results in Chapter 7

F.1.1 Proof of Lemma 52

Use Q to denote a generic quasi-likelihood and Q_0 the quasi-likelihood corresponds to the true parameter (θ_0, η_0). Let P_0 be the true distribution that generates the observations $X_1, \ldots, X_n$. Let $\bar{\Pi}_n$ be any probability measure that supported on the set $\{Q : ||g - g_0||_n \leqslant \rho_n\}$. By the Cauchy inequality,

$$\int \prod_{i=1}^{n} \frac{Q}{Q_0}(X_i) d\bar{\Pi}_n(Q) \cdot \int \prod_{i=1}^{n} \frac{Q_0}{Q}(X_i) d\bar{\Pi}_n(Q) \geqslant 1.$$

Combining the above with Markov inequality and Fubini's theorem, we obtain that for any $C > 0$,

$$\begin{aligned} & P_0\bigg(\int \prod_{i=1}^{n} \frac{Q}{Q_0}(X_i) d\bar{\Pi}_n(Q) \leqslant \exp(-Cn\rho_n^2) \bigg) \\ \leqslant & P_0\bigg(\int \prod_{i=1}^{n} \frac{Q_0}{Q}(X_i) d\bar{\Pi}_n(Q) \geqslant \exp(Cn\rho_n^2) \bigg) \\ \leqslant & \exp(-Cn\rho_n^2) \int \prod_{i=1}^{n} P_0\bigg(\frac{Q_0}{Q}(X_i) \bigg) d\bar{\Pi}_n(Q). \end{aligned} \tag{F.1}$$

By the definition of quasi-likelihood function and assumption (A2), we have

$$\begin{aligned} \log Q_0(X_i) - \log Q(X_i) = & - W_i \int_{F(g(T_i))}^{F(g_0(T_i))} \frac{1}{V(s)} ds + \int_{F(g(T_i))}^{F(g_0(T_i))} \frac{(s - F(g_0(T_i))}{V(s)} ds \\ & \leqslant C_1 |W_i| \cdot |g(T_i) - g_0(T_i)| + C_1 (g(T_i) - g_0(T_i))^2, \end{aligned}$$

where C_1 is some positive constant. Applying assumption Assumption 1(1) to the above inequality, we obtain

$$\prod_{i=1}^{n} P_0\bigg(\frac{Q_0}{Q}(X_i) \bigg) \leqslant \exp\bigg\{ C_2 \sum_{i=1}^{n} (g(T_i) - g_0(T_i))^2 \bigg\} = \exp(C_2 n ||g - g_0||_n^2),$$

for some $C_2 > 0$. Combining the above and (F.1) and choosing $C > C_2 + 1$, we obtain

$$\begin{aligned} & P_0\bigg(\int \prod_{i=1}^{n} \frac{Q}{Q_0}(X_i) d\bar{\Pi}_n(Q) \leqslant \exp(-Cn\rho_n^2) \bigg) \\ \leqslant & \exp(-Cn\rho_n^2) \int \exp(C_2 n ||g - g_0||_n^2) d\bar{\Pi}_n(Q) \leqslant \exp(-n\rho_n^2), \end{aligned} \tag{F.2}$$

where the last step follows by the fact that $\bar{\Pi}_n$ is supported on the set $\{Q : ||g - g_0||_n \leqslant \rho_n\}$.

Using (F.2) to replace the Lemma 8.1 (Ghosal et al., 2000) in the proof of Theorem 2.1 (Ghosal et al., 2000), we can finish the proof.

F.1.2 Proof of Lemma 60

Let $B_n = \{||\theta - \theta_0|| \leqslant M\rho_n,\ \eta \in \mathcal{H}_n\}$. Then by Assumption 2 and the definition of $\mathcal{H}_n$ in (7.25), $\Pi(B_n|X^{(n)}) = 1 - O_P(\delta_n)$. For any measurable $A \subset \mathbb{R}^k$,

$$\begin{aligned}
&\big|\Pi(\theta \in A|X^{(n)}, B_n) - \Pi(\theta \in A|X^{(n)})\big| \\
=&\left|\frac{\Pi(\theta \in A|X^{(n)})\big[1 - \Pi(B_n|X^{(n)})\big] - \Pi(\theta \in A, B_n^c|X_n)}{\Pi(B_n|X^{(n)})}\right| \\
\leqslant& 2\big|1 - \Pi(B_n|X^{(n)})\big|/\Pi(B_n|X^{(n)}) \\
=& O_P(\delta_n).
\end{aligned}$$

Take the supreme over A, we obtain

$$\sup_A \big|\Pi(\theta \in A|X^{(n)}, B_n) - \Pi(\theta \in A|X^{(n)})\big| = O_P(\delta_n).$$

Therefore, to prove (7.33), we only need to prove that

$$\sup_A \big|\Pi(\theta \in A|X_1, \ldots, X_n, B_n) - N_k\big(\widetilde{\Delta}_n, (nI_{\theta_0,\eta_0})^{-1}\big)(A)\big| = O_P[R_n(n^{-1/2}\log n)], \tag{F.3}$$

where,

$$\Pi(\theta \in A|X_1, \ldots, X_n, B_n) = \int_{A\cap\{||\theta-\theta_0||\leqslant M\rho_n\}} \frac{\widetilde{S}_n(\theta)}{\widetilde{S}_n(\theta_0)} d\Pi(\theta) \Big/ \int_{||\theta-\theta_0||\leqslant M\rho_n} \frac{\widetilde{S}_n(\theta)}{\widetilde{S}_n(\theta_0)} d\Pi(\theta). \tag{F.4}$$

Recall the definition of $\widetilde{\Delta}_n$ by (**??**). Since the pdf of a normal random variable with mean $\theta_0 + n^{-1/2}\widetilde{\Delta}_n$ and variance $(nI_{\theta_0,\eta_0})^{-1}$ evaluated at θ is proportional to

$$\exp\left\{(\theta - \theta_0)^T \sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0}(X_i) - \frac{n}{2}(\theta - \theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta - \theta_0) - \frac{1}{2}\widetilde{\Delta}_n^T \widetilde{I}_{\theta_0,\eta_0}\widetilde{\Delta}_n\right\},$$

it suffices to prove

$$\begin{aligned}\Big| \int_A \exp\Big\{(\theta-\theta_0)^T \sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0}(X_i) - \frac{n}{2}(\theta-\theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\Big\} d\theta \\ - \int_{A\cap\{||\theta-\theta_0||\leqslant M\rho_n\}} \frac{\widetilde{S}_n(\theta)}{\widetilde{S}_n(\theta_0)} d\Pi(\theta)\Big| \\ =O_P[R_n(n^{-1/2}\log n)] \int_{\mathbb{R}^k} \exp\Big\{(\theta-\theta_0)^T \sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0}(X_i) \\ - \frac{n}{2}(\theta-\theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\Big\} d\theta,\end{aligned} \tag{F.5}$$

In fact, one can plug in the above equation with $A = A$ and $A = \mathbb{R}^k$ respectively, and simple algebra could lead to (F.3).

By $n\rho_n^2 \gtrsim -\log R_n(n^{-1/2}\log n) \to \infty$ in condition 3 and $\sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0} = O_P(\sqrt{n})$, with M sufficiently large,

$$\begin{aligned}&\Big| \int_{A\cap\{||\theta-\theta_0||>M\rho_n\}} \exp\Big\{(\theta-\theta_0)^T \sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0}(X_i) - \frac{n}{2}(\theta-\theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\Big\} d\theta \\ &=O_P[R_n(n^{-1/2}\log n)] \int_{\mathbb{R}^k} \exp\Big\{(\theta-\theta_0)^T \sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0}(X_i) - \frac{n}{2}(\theta-\theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\Big\} d\theta.\end{aligned} \tag{F.6}$$

By a subsequence argument, the ILAN (7.15) implies that

$$\begin{aligned}\sup_{||\theta-\theta_0||\leqslant M\rho_n} \Big| \log \frac{\widetilde{S}_n(\theta)}{\widetilde{S}_n(\theta_0)} - (\theta-\theta_0)^T \sum_{i=1}^n \widetilde{l}_{\theta_0,\eta_0}(X_i) \\ + \frac{n}{2}(\theta-\theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\Big| / R_n(|\theta-\theta_0|) = O_P(1).\end{aligned} \tag{F.7}$$

For every θ such that $||\theta-\theta_0|| < Mn^{-1/2}\log n$ with M sufficiently large, by the

above, we have

$$\begin{aligned}
&\left|\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)-\frac{n}{2}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\}-\frac{\widetilde{S}_n(\theta)}{\widetilde{S}_n(\theta_0)}\right| \\
\leqslant&\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)-\frac{n}{2}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\} \\
&\qquad\qquad\left|\exp\left\{O_P[R_n(n^{-1/2}\log n)]\right\}-1\right| \\
=&O_P[R_n(n^{-1/2}\log n)]\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)-\frac{n}{2}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\},
\end{aligned} \tag{F.8}$$

where the last step follows by $R_n(n^{-1/2}\log n)=o(1)$.

Therefore, for every θ such that $Mn^{-1/2}\log n\leqslant||\theta-\theta_0||<M\rho_n$ with M sufficiently large, the assumption that $\alpha_n=\sup_{|t|\leqslant\rho_n}R_n(t)/(nt^2)=o(1)$ and (F.7) imply $R_n(|\theta-\theta_0|)=o[n(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)]$. Hence, we have,

$$\begin{aligned}
&\left|\int_{A\cap\{Mn^{-1/2}\log n\leqslant||\theta-\theta_0||<M\rho_n\}}\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)\right.\right. \\
&\qquad\left.\left.-\frac{n}{2}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\}d\theta-\int_{A\cap\{Mn^{-1/2}\log n\leqslant||\theta-\theta_0||<M\rho_n\}}\frac{\widetilde{S}_n(\theta)}{\widetilde{S}_n(\theta_0)}d\Pi(\theta)\right| \\
=&O_P(1)\int_{||\theta-\theta_0||>Mn^{-1/2}\log n}\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)-\frac{n}{4}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\}d\theta \\
=&O_P(e^{-Mc(\log n)^2})\int_{\mathbb{R}^k}\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)-\frac{n}{8}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\}d\theta \\
=&O_P[R_n(n^{-1/2}\log n)]\int_{\mathbb{R}^k}\exp\left\{(\theta-\theta_0)^T\sum_{i=1}^n\widetilde{l}_{\theta_0,\eta_0}(X_i)-\frac{n}{2}(\theta-\theta_0)^T\widetilde{I}_{\theta_0,\eta_0}(\theta-\theta_0)\right\}d\theta,
\end{aligned} \tag{F.9}$$

for M sufficiently large, where $c>0$ is a constant dependent on $\widetilde{I}_{\theta_0,\eta_0}$ and the last step follows by the fact that $\int\exp\{at-bt^2\}dt\backsim b^{-1/2}$ for $b\gg\min(a,1)$.

Finally, (F.6),(F.8) and (F.9) imply (F.5).

F.1.3 Proof of Corollary 49

For each $s = 1, \ldots, k$, taking $A = \mathbb{R} \times \cdots \times A_s \times \cdots \times \mathbb{R}$ in (7.16), where the s-th component is A_s and the rest are $\mathbb{R}$, we obtain

$$\sup_{A_s \subset \mathbb{R}} \left| \Pi(\theta_s \in A_s | X_1, \ldots, X_n) - N\big(\theta_{0,s} + n^{-1/2}\widetilde{\Delta}_{n,s}, n^{-1}\widetilde{I}^{ss}_{\theta_0,\eta_0}\big)(A_s) \right| = O_P(S_n),$$

where $\widetilde{\Delta}_{n,s}$ is the sth component of $\widetilde{\Delta}_n$ and $\widetilde{I}^{ss}_{\theta_0,\eta_0}$ the (s,s)th element of the matrix $\widetilde{I}^{-1}_{\theta_0,\eta_0}$. Let $\widehat{\theta}^B_{n,s}$ to be the median of the marginal posterior distribution of θ_s. Then taking $A_s = (-\infty, \widehat{\theta}^B_{n,s})$ in the above formula yields

$$\left| \Phi\big(n^{1/2}(\widetilde{I}^{ss}_{\theta_0,\eta_0})^{-1/2}(\widehat{\theta}^B_{n,s} - \theta_{0,s} - n^{-1/2}\widetilde{\Delta}_{n,s})\big) - 1/2 \right| = O_P(S_n),$$

where Φ is the cdf of the standard normal distribution. By the continuity of Φ^{-1}, we have

$$n^{1/2}(\widetilde{I}^{ss}_{\theta_0,\eta_0})^{-1/2}(\widehat{\theta}^B_{n,s} - \theta_{0,s} - n^{-1/2}\widetilde{\Delta}_{n,s}) = O_{P_0}(S_n).$$

Concatenating $\widehat{\theta}^B_{n,s}$, $s = 1, \ldots, k$, into a vector provides the desired $\widehat{\theta}^B_n$.

F.1.4 Proof of Corollary 50

We only prove (7.21) here. The proof of (7.20) is similar by noticing the fact that

$$\Pi(\theta_s \in (\widehat{q}_{s,\alpha/2}, \widehat{q}_{s,1-\alpha/2}) | X_1, \ldots, X_n) = 1 - \alpha.$$

By (7.16) and the definition of $A_{n,1-\alpha}$, we have

$$\left| P\big(\theta_0 + n^{-1/2}\widetilde{\Delta}_n + n^{-1/2}\widetilde{I}^{-1/2}_{\theta_0,\eta_0} N \in A_{n,1-\alpha}\big) - (1-\alpha) \right| = O_P(S_n),$$

where N is a random vector that follows $N_k(0, I_k)$, with I_k the k-by-k identity matrix. Therefore, for

$$B_n = n^{1/2}\widetilde{I}^{1/2}_{\theta_0,\eta_0}\big(A_{n,1-\alpha} - \theta_0 - n^{-1/2}\widetilde{\Delta}_n\big),$$

we have

$$P(N \in B_n) = 1 - \alpha + O_P(S_n). \tag{F.10}$$

Note that

$$\begin{aligned} P_0(\theta_0 \in A_{n,1-\alpha}) &= P_0(\theta_0 \in \theta_0 + n^{-1/2}\widetilde{\Delta}_n + n^{-1/2}\widetilde{I}^{-1/2}_{\theta_0,\eta_0} B_n) \\ &= P_0(\widetilde{I}^{1/2}_{\theta_0,\eta_0}\widetilde{\Delta}_n \in -B_n) \\ &= P(N \in -B_n) + O(n^{-1/2}) \\ &= 1 - \alpha + O(S_n), \end{aligned}$$

where the third step follows by the fact that

$$\widetilde{I}^{1/2}_{\theta_0,\eta_0}\widetilde{\Delta}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \widetilde{I}^{-1/2}_{\theta_0,\eta_0}\widetilde{l}_{\theta_0,\eta_0}(X_i) \overset{P_0}{\rightsquigarrow} N(0, I_k)$$

and the Edgeworth expansion, and the last step follows by (F.10), the symmetry of the distribution of $N_k(0, I_k)$, and the fact that $n^{-1/2} = o(S_n)$.

F.1.5 Proof of Lemma 54

Under A3, $\Delta\eta(\theta_n) = O(|\theta_n - \theta_0|) = O(\rho_n)$. If M is sufficiently large, then

$$\begin{aligned} \frac{\int_{\mathcal{H}_n - \Delta\eta(\theta_n)} e^{l_n(\theta_0,\eta)} d\Pi(\eta)}{\int_{\mathcal{H}_n} e^{l_n(\theta_0,\eta)} d\Pi(\eta)} &= \frac{\int_{\mathcal{H}_n - \Delta\eta(\theta_n)} e^{l_n(\theta_0,\eta)} d\Pi(\eta)}{\int_{\mathcal{H}} e^{l_n(\theta_0,\eta)} d\Pi(\eta)} \cdot \frac{\int_{\mathcal{H}} e^{l_n(\theta_0,\eta)} d\Pi(\eta)}{\int_{\mathcal{H}_n} e^{l_n(\theta_0,\eta)} d\Pi(\eta)} \\ &= \frac{\Pi(\mathcal{H}_n - \Delta\eta(\theta_n)|X_1, \ldots, X_n)}{\Pi(\mathcal{H}_n|X_1, \ldots, X_n)} = 1 + O_P(\delta_n), \end{aligned} \tag{F.11}$$

where $\Pi^{\theta_0}(\cdot|X_1, \ldots, X_n)$ is the posterior of η when θ is fixed at θ_0 and the the last step uses Assumption 2. If A4 is true, then by the above observation and (7.27), (A2) holds with $G'_n = G''_n + \delta_n$.

F.1.6 Proof of Lemma 55

Applying a change of variable $\eta \to \widetilde{\eta} + (\theta_n - \theta_0)\widehat{h}$, we obtain

$$\begin{aligned}
&\int_{\mathcal{H}_n} e^{l_n(\theta_0, \eta - \Delta\eta(\theta_n))} d\Pi^{\theta_n}(\eta) \\
=&\int_{\mathcal{H}_n - (\theta_n - \theta_0)\widehat{h}} e^{l_n(\theta_0, \eta - \Delta\eta(\theta_n) + (\theta_n - \theta_0)\widehat{h})} d\Pi^{\theta_n}_{\cdot - (\theta_n - \theta_0)\widehat{h}}(\eta) \\
=&\int_{\mathcal{H}_n - (\theta_n - \theta_0)\widehat{h}} e^{l_n(\theta_0, \eta - \Delta\eta(\theta_n) + (\theta_n - \theta_0)\widehat{h})} d\Pi^{\theta_0}(\eta) \\
=&\big(1 + O_{P_0}[G_n''(\max\{|\theta - \theta_0|, n^{-1/2}\log n\})]\big) \int_{\mathcal{H}_n - (\theta_n - \theta_0)\widehat{h}} e^{l_n(\theta_0, \eta)} d\Pi^{\theta_0}(\eta) \\
=&\big(1 + O_{P_0}[G_n'(\max\{|\theta - \theta_0|, n^{-1/2}\log n\})]\big) \int_{\mathcal{H}_n} e^{l_n(\theta_0, \eta)} d\Pi^{\theta_0}(\eta),
\end{aligned}$$

where the second step follows by the definition of Π, the third step by A5 and the last step by the same argument as (F.11).

F.1.7 Proof of Lemma 53

With the definition of $\widetilde{S}_n$ and the conditions in the lemma, we have

$$\begin{aligned}
\widetilde{S}_n(\theta) =& \int_{\mathcal{H}_n} e^{l_n(\theta, \eta) - l_n(\theta_0, \eta_0)} d\Pi^{\theta}(\eta) \\
=& \exp\Bigg\{ \sqrt{n}(\theta_n - \theta_0)^T \widetilde{g}_n - \frac{1}{2} n (\theta_n - \theta_0)^T \widetilde{I}_{\theta_0, \eta_0} (\theta_n - \theta_0) \\
&+ O_P[G_n(\max\{|\theta - \theta_0|, n^{-1/2}\log n\})] \Bigg\} \int_{\mathcal{H}_n} e^{l_n(\theta_0, \eta - \Delta\eta(\theta)) - l_n(\theta_0, \eta_0)} d\Pi^{\theta}(\eta) \\
=& \exp\Bigg\{ \sqrt{n}(\theta_n - \theta_0)^T \widetilde{g}_n - \frac{1}{2} n (\theta_n - \theta_0)^T \widetilde{I}_{\theta_0, \eta_0} (\theta_n - \theta_0) \\
&+ O_P[R_n(\max\{|\theta - \theta_0|, n^{-1/2}\log n\})]) \Bigg\} \widetilde{S}_n(\theta_0).
\end{aligned}$$

ILAN follows by taking a logarithm of the above.

F.1.8 Proof of Theorem 57

Verification of Assumption 2: We apply Lemma 51 with a modified sieve construction of the nuisance parameter η as van der Vaart and van Zanten (2009), so that the sieve has an upper bound for every ϵ-covering entropy.

Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any d dimensional multi-index $a = (a_1, \ldots, a_d) \in \mathbb{N}_0^d$ define $|a| = a_1 + \cdots + a_d$ and let D^a denote the mixed partial derivative operator $\partial^{|a|}/\partial x_1^{a_1} \cdots \partial x_d^{a_d}$. For any real number b let $\lfloor b \rfloor$ denote the largest integer strictly smaller than b. The Hölder class $C^\gamma([0,1]^d)$ is defined as the set of all d-variate $k = \lfloor \gamma \rfloor$ times differentiable functions f on $[0,1]^d$ such that:

$$||f||_{C^\gamma} = \max_{|\beta| \leqslant k} \sup_{x \in [0,1]^d} |D^\beta f(x)| + \max_{|\beta|=k} \sup_{x \neq y} \frac{|D^\beta(x) - D^\beta(y)|}{|x-y|^{\gamma-k}} < \infty.$$

We use C_1^γ to denote the unit ball in C^γ under the norm $||\cdot||_{C^\gamma}$.

We choose the sieve $\mathcal{F}_n$ as $\mathcal{F}_n^\theta \oplus \mathcal{F}_n^\eta$, with

$$\mathcal{F}_n^\theta = [-c\sqrt{n}, c\sqrt{n}]^k,$$

$$\mathcal{F}_n^\eta = \left(M_n \sqrt{\frac{r_n}{\delta_n}} \mathbb{H}_1^{r_n} + \rho_n C_1^\gamma \right) \cup \Big(\bigcup_{a \leqslant \delta_n} (M_n \mathbb{H}_1^a) + \rho_n C_1^\gamma \Big), \tag{F.12}$$

with c sufficiently large, $\rho_n = n^{-\alpha/(2\alpha+1)} (\log n)^{d+1}$, and

$$D_2 r_n^d \geqslant 2C_0 n \rho_n^2, \quad r_n^{p_0 - d + 1} \leqslant e^{C_0 n \rho_n^2},$$

$$M_n^2 \geqslant 8 C_0 n \rho_n^2, \quad \delta_n = C_1 \rho_n / (2\sqrt{d} M).$$

The only difference between $\mathcal{F}_n^\eta$ and the sieve in van der Vaart and van Zanten (2009) is the remainder term, which is C_1^γ in our case and $B_1 = \{f \in L_2([0,1]^d) : ||f||_\infty\}$ in van der Vaart and van Zanten (2009).

Similar to van der Vaart and van Zanten (2009), we can verify that $\mathcal{F}_n$ satisfies condition a and condition b in Lemma 51 as follows:

By Lemma 4.5 in van der Vaart and van Zanten (2009), for a fixed scaling parameter a and any $\epsilon < 1/2$,

$$\log N(\epsilon, \mathbb{H}_1^a, ||\cdot||_\infty) \leqslant K a^d \left(\log \frac{1}{\epsilon} \right)^{1+d}.$$

For squared exponential kernel, all elements in $\mathbb{H}_1^a$ are infinitely differentiable. With some modifications of their proof, the above can be strength to the following: for any smoothness index $\gamma > 0$,

$$\log N(\epsilon, \mathbb{H}_1^a, ||\cdot||_{C^\gamma}) \leqslant K a^d \left(\log \frac{a^\gamma}{\epsilon} \right) \left(\log \frac{1}{\epsilon} \right)^{d}. \tag{F.13}$$

Therefore, by the relation between the covering entropy of the unit ball of RKHS and small ball probability (Li and Linde, 1999) and similar proof as Lemma 4.6 in van der Vaart and van Zanten (2009) , we have that for any $\gamma > 0$,

$$-\log P(||W^a||_{C^\gamma} \leqslant \epsilon) \leqslant K a^d \left(\log \frac{a}{\epsilon} \right)^{1+d}. \tag{F.14}$$

Denote the right hand side of the above by $\phi_0^a(\epsilon)$. Then by Borell's inequality (van der Vaart and van Zanten, 2008c),

$$P(W^a \notin M\mathbb{H}_1^a + \epsilon C_1^\gamma) \leqslant 1 - \Phi(\Phi^{-1}(e^{-\phi_0^a(\epsilon)}) + M),$$

where Φ is the c.d.f. of standard normal distribution. For $M > 4\sqrt{\phi_0^a(\epsilon)}$ and $\phi_0^a(\epsilon) < 1/4$, the above is bounded by $e^{-M^2/8}$. Combining the above conclusions, our sieve construction and the covering entropy for C_1^γ, we have the following bound for the ϵ-covering entropy for any $\epsilon > 0$,

$$\log N(4\epsilon, \mathcal{F}_n, ||\cdot||_\infty) \lesssim n\rho_n^2 (\log n)^{-(d+1)} \left(\log \left(\frac{n}{\epsilon} \right) \right)^{1+d} + \left(\frac{\rho_n}{\epsilon} \right)^{d/\gamma} + c \log \left(\frac{n}{\epsilon} \right), \tag{F.15}$$

and the following complement probability

$$P(\mathcal{F}_n^c) \lesssim \exp(-C_0 n \rho_n^2). \tag{F.16}$$

Next we verify condition c in Lemma 51:

For the partially linear model, we have,

$$\begin{aligned} K(P^{(n)}_{\theta_0,\eta_0}, P_{\theta,\eta})^{(n)} =& E_0\big\{ \log(dP^{(n)}_{\theta_0,\eta_0}/dP^{(n)}_{\theta,\eta})\big\} \\ =& \frac{1}{2}\sum_{i=1}^{n}[(\theta-\theta_0)U_i + (\eta-\eta_0)(V_i)]^2, \end{aligned}$$

and for any $k \geqslant 2$,

$$\begin{aligned} V_{k,0}(P^{(n)}_{\theta_0,\eta_0}, P^{(n)}_{\theta,\eta}) =& E_0\big\{ |\log(dP^{(n)}_{\theta_0,\eta_0}/dP^{(n)}_{\theta,\eta}) - K(P^{(n)}_{\theta_0,\eta_0}, P^{(n)}_{\theta,\eta})|^k\big\} \\ =& E_0\big\{ \big|\sum_{i=1}^{n} e_i[(\theta-\theta_0)U_i + (\eta-\eta_0)(V_i)]\big|^k \big| U^n, V^n\big\} \\ =& C\big(\sum_{i=1}^{n}[(\theta-\theta_0)U_i + (\eta-\eta_0)(V_i)]^2\big)^{k/2}. \end{aligned}$$

where the expectation is taken with respect to Y^n and the last step follows by the fact that $\sum_{i=1}^{n} e_i[(\theta-\theta_0)U_i + (\eta-\eta_0)(V_i)]$ has a normal distribution with mean zero and variance equal to $\big(\sum_{i=1}^{n}[(\theta-\theta_0)U_i + (\eta-\eta_0)(V_i)]^2\big)^{1/2}$. C is a constant independent of n and ϵ (but depends on k). Therefore, for any $k \geqslant 2$,

$$\begin{aligned} B_n\big(P^{(n)}_0, \epsilon; k\big) =& \big\{(\theta,\eta) : K(P^{(n)}_{\theta_0,\eta_0}, P^{(n)}_{\theta,\eta}) \leqslant n\epsilon^2, V_{k,0}(P^{(n)}_{\theta_0,\eta_0}, P^{(n)}_{\theta,\eta}) \leqslant n^{k/2}\epsilon^k\big\} \\ =& \big\{(\theta,\eta) : K(P^{(n)}_{\theta_0,\eta_0}, P^{(n)}_{\theta,\eta}) \leqslant Cn\epsilon^2\big\}. \end{aligned}$$

By similar proof as Theorem 3.3 in Ghosal and van der Vaart (2007), we have for some constant C_1,

$$P_0(B_n\big(P^{(n)}_0, \rho_n; k)) = P_0(K(P^{(n)}_{\theta_0,\eta_0}, P^{(n)}_{\theta,\eta}) \leqslant Cn\rho_n^2) \geqslant \exp(-C_1 n\rho_n^2).$$

Combining the above conclusions and Lemma 51, we prove Assumption 2 and conclude that for any $r \geqslant 0$,

$$\Pi\bigg\{\bigg(\frac{1}{n}\sum_{i=1}^{n}\big((\theta-\theta_0)^T U_i + (\eta-\eta_0)(V_i)\big)^2\bigg)^{1/2} \leqslant M\rho_n \bigg| X_1, \ldots, X_n\bigg\} = 1 - O(\delta_n), \tag{F.17}$$

where $\delta_n = n^{-r}$. In the sequel, we always fix r at a value such that $\delta_n = o(\sqrt{n}\rho_n^2)$, for example, at $1/2$.

Next, we show that under the condition 2 in the theorem, (F.17) implies $\Pi\big(||\theta - \theta_0|| \leqslant M\rho_n, ||\eta - \eta_0||_n \leqslant M\rho_n \big| X^{(n)}\big) = 1 - O_P(\delta_n)$. For exposition simplicity, we focus on the case that $k = 1$ and for $k > 1$, the proof is essentially the same. In fact, by the central limited theorem conditioning on V_i's, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \big(U_i - E(U_i|V_i)\big) \cdot \big((\eta - \eta_0)(V_i) + (\theta - \theta_0)E(U_i|V_i)\big) \right| = O_P(n^{-1/2})I_n,$$

with $I_n^2 = \sum_{i=1}^n \big((\eta - \eta_0)(V_i) + (\theta - \theta_0)E[U_i|V_i]\big)^2$. Therefore,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \big((\theta - \theta_0)U_i + (\eta - \eta_0)(V_i)\big)^2 \\
=&\frac{1}{n} \sum_{i=1}^{n} \Big((\theta - \theta_0)(U_i - E[U_i|V_i]) + \big((\eta - \eta_0)(V_i) + (\theta - \theta_0)E(U_i|V_i)\big) \Big)^2 \\
=&\big(P(U - E[U|V])^2 + O_{P_0}(n^{-1/2})\big)(\theta - \theta_0)^2 + O_P(n^{-1/2})(\theta - \theta_0)I_n + n^{-1}I_n^2.
\end{aligned}$$

Combining the above with (F.17), we obtain

$$\Pi(||\theta - \theta_0|| \leqslant M\rho_n | X_1, \ldots, X_n) = 1 - O_P(\delta_n).$$

Again applying (F.17) and using the inequality $(a + b)^2 \geqslant b^2/2 - a^2$, we have that for M sufficiently large,

$$\Pi(||\eta - \eta_0||_n \leqslant M\rho_n | X_1, \ldots, X_n) = 1 - O_P(\delta_n).$$

Combining the above two yields

$$\Pi\big(||\theta - \theta_0|| \leqslant M\rho_n, ||\eta - \eta_0||_n \leqslant M\rho_n \big| X^{(n)}\big) = 1 - O_P(\delta_n).$$

Therefore, if we define $H_n = \{\eta \in \mathcal{F}_n^\eta : ||\eta - \eta_0||_n \leqslant M\rho_n\}$, then

$$\Pi\big(||\theta - \theta_0|| \leqslant M\rho_n, \eta \in \mathcal{H}_n \big| X^{(n)}\big) = 1 - O_P(\delta_n). \tag{F.18}$$

Verification of A3: A3 is true with $h^*(v) = -E[U|V = v]$.

Verification of (A1): We verify assumption (A1) with the above choice of $\mathcal{H}_n$. In fact, for the partially linear model, $\Delta\eta(\theta) = -(\theta-\theta_0)^T E[U|V]$. We use the notation $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ to denote the empirical measure and $\mathbb{G}_n = n^{-1/2}\sum_{i=1}^n(\delta_{X_i} - P)$ the empirical process. By the expression of log likelihood (7.8),

$$\begin{aligned}
\log \frac{dP_{\theta,\eta+\Delta\eta(\theta)}}{dP_{\theta_0,\eta}}(X^{(n)}) =& -\frac{1}{2}\sum_{i=1}^n \big[\epsilon_i - (\eta-\eta_0)(V_i) - (\theta-\theta_0)^T(U_i - E[U|V_i])\big]^2 \\
&+ \frac{1}{2}\sum_{i=1}^n \big[\epsilon_i - (\eta-\eta_0)(V_i)\big]^2 \\
=&(\theta-\theta_0)^T\sum_{i=1}^n \tilde{l}_{\theta_0,\eta_0}(X_i) - \frac{n}{2}(\theta-\theta_0)^T I_0(\theta-\theta_0) \\
&+ \frac{1}{2}\sqrt{n}(\theta-\theta_0)^2\mathbb{G}_n\big(U - E[U|V]\big)^2 \\
&- (\theta-\theta_0)^T\sum_{i=1}^n \big(U_i - E[U|V_i]\big)(\eta-\eta_0)(V_i),
\end{aligned}$$

where $g_0(X) = \epsilon\big(U - E[U|V]\big)$ and $I_0 = P\big(U - E[U|V]\big)^2 = E_{\theta_0,\eta_0}g_0^2(X)$.

By central limit theorem, the third term is $O_P\big(\sqrt{n}|\theta-\theta_0|^2\big)$.

A bound of the last term could be achieved by applying the maximal inequality conditioning on V_i's. A key step is the bound (F.15) for the covering entropy of the space $\{\eta-\eta_0 : \eta \in \mathcal{H}_n\}$. Since $||\eta-\eta_0||_n \lesssim \rho_n$ for any $\eta \in \mathcal{H}_n$ and U_i conditioning on V_i are i.i.d. with $E\{U_i - E[U|V_i]|V_i\} = 0$, an application of the maximal inequality (van der Vaart and Wellner, 1996) yields

$$\begin{aligned}
&E\left\{\sup_{\eta\in H_n}\frac{1}{\sqrt{n}}\Big|\sum_{i=1}^n \big(U_i - E[U|V_i]\big)(\eta-\eta_0)(V_i)\Big|\,\Big|V_1,\ldots,V_n\right\} \\
\lesssim& \int_0^{\rho_n}\sqrt{1+\log N(\epsilon, H_n, ||\cdot||_\infty)}d\epsilon \\
\lesssim& \sqrt{n}\rho_n^2 + \rho_n \backsim \sqrt{n}\rho_n^2.
\end{aligned} \tag{F.19}$$

Hence

$$\sup_{\eta\in H_n}(\theta-\theta_0)^T\Big|\sum_{i=1}^{n}\big(U_i-E[U|V_i]\big)(\eta-f_0)(V_i)\Big|$$

$$=O_P\bigg\{n|\theta-\theta_0|\rho_n^2\bigg\}.$$

Therefore, (A1) is true with $G_n(t)=\sqrt{n}t^2+n\rho_n^2 t$.

Verification of A4: Since $\Pi(A_n|X^{(n)})=1-O(\delta_n)$ with $A_n=\{A\leqslant Cn\rho_n^2\}$ for C sufficiently large, where A is the random inverse bandwidth parameter in the GP prior. We can always assume $A\leqslant Cn\rho_n^2$ by conditioning on the event A_n. By Lemma 4.7 in van der Vaart and van Zanten (2009) and the assumption that $E[U|V]\in\mathbb{H}^{t_0}$, for any $a\geqslant t_0$, $||E[U|V=\cdot]||_a\leqslant C_1\sqrt{a}$, where $C_1=||E[U|V=\cdot]||_{t_0}$ is a constant not depending on a. Denote the conditional law of f given $(A=a)$ by Π^a. Do a change of variable $\eta\to\eta-(\theta-\theta_0)E[U|V]$. Since the Radon-Nykodym derivative $d\Pi^a_{\cdot+h^*}/d\Pi^a_{\cdot}(W)=\exp(Uh^*-||h^*||_a^2/2)$ and $\mathrm{Var}U\big(E[U|V]\big)=||E[U|V=\cdot]||_a^2\leqslant C_1a\leqslant C_1n\rho_n^2$ (van der Vaart and van Zanten, 2008c, Lemma 3.1), we have

$$\begin{aligned}\log f(\eta)&=\log d\Pi^a_{\cdot+h^*}/d\Pi^a_{\cdot}(W)\\&=(\theta-\theta_0)U\big(E[U|V]\big)+(\theta-\theta_0)^2||E[U|V=\cdot]||_a^2/2=O_P(G_n''(||\theta-\theta_0||)),\end{aligned}$$

with $G_n''(t)=\sqrt{n}\rho_n t+n\rho_n^2t^2$.

Finally, applying Theorem 56 yields the second order semiparametric BvM theorem with a remainder term

$$G_n(n^{-1/2}\log n)+G_n''(n^{-1/2}\log n)+\delta_n\sim n^{1/2}\rho_n^2\log n.$$

F.1.9 Proof of Theorem 58

Most of the proof is similar to that of Theorem 57. The only difference is that instead of applying the arguments in section 7.4.1, now we apply assumption A5 and Lemma 55.

Verification of assumption A5: By A6 and the form of h, we have

$$\widehat{\Delta}\eta(\theta_n) = (\theta_n - \theta_0)O_P(\kappa_n).$$

Then for $\eta \in \mathcal{H}_n$ and θ such that $|\theta_n - \theta_0| = O_P(\rho_n)$,

$$\begin{aligned}
&l_n\big(\theta_0, \eta + (\theta_n - \theta_0)(\widehat{h} - h^*)\big) - l_n\big(\theta_0, \eta\big) \\
=& -\frac{1}{2}\sum_{i=1}^{n}\big(\epsilon_i + (\eta - \eta_0)(V_i) + (\theta_n - \theta_0)(\widehat{h} - h^*))(V_i)\big)^2 + \frac{1}{2}\sum_{i=1}^{n}\big(\epsilon_i + (\eta - \eta_0)(V_i)\big)^2 \\
=& -(\theta_n - \theta_0)\sum_{i=1}^{n}\big(\epsilon_i + (\eta - \eta_0)(V_i)\big)(\widehat{h} - h^*)(V_i) - \frac{n}{2}(\theta_n - \theta_0)^2||\widehat{h} - h^*||_n.
\end{aligned}$$

By Cauchy's inequality

$$\big|\sum_{i=1}^{n}(\eta - \eta_0)(V_i)(\widehat{h} - h^*)(V_i)\big| \leqslant n||\eta - \eta_0||_n||\widehat{h} - h^*||_n = O_P(n\rho_n\kappa_n).$$

Since

$$E\big|\sum_{i=1}^{n}\epsilon_i(\widehat{h} - h^*)(V_i)\big|^2 = n||\eta - \eta_0||_n^2 = O_P(n\rho_n^2),$$

we have

$$\big|\sum_{i=1}^{n}\epsilon_i(\widehat{h} - h^*)(V_i)\big| = O_P(\sqrt{n}\rho_n).$$

Combining the above three, we have

$$l_n\big(\theta_0, \eta + (\theta_n - \theta_0)(\widehat{h} - h^*)\big) - l_n\big(\theta_0, \eta\big) = O_P(G_n''(|\theta_n - \theta_0|)),$$

with $G_n''(t) = \sqrt{n}\rho_n t + n\kappa_n\rho_n t + n\kappa_n t^2$.

Combining the above with the proof of Theorem 57 yields the second order semi-parametric BvM theorem with a remainder term

$$R_n(t) = n\kappa_n t^2 + \sqrt{n}t^2 + \sqrt{n}\rho_n t + n\rho_n^2 t + n\rho_n\kappa_n t + \delta_n,$$

which implies that $R_n(n^{-1/2}\log n) \backsim n^{1/2}\rho_n^2\log n + n^{1/2}\kappa_n\rho_n\log n$.

F.1.10 Proof of Lemma 47

By Assumption 1(2), for any (θ, η), we have

$$\begin{aligned} & E_0 \log(Q_{\theta,\eta}/Q_{\theta_0,\eta_0}) \\ \leqslant & - C_1^{-1} E_0 \big(m_{\theta,\eta}(T) - m_{\theta_0,\eta_0}(T)\big)^2 \\ \leqslant & - (C_1^2 C_2)^{-1} E_0 |g_{\theta,\eta} - g_0|^2 \\ \leqslant & - 2(C_1^2 C_2)^{-1} \big(|\theta - \theta_0|^2 + E_0|\eta - \eta_0|^2\big), \end{aligned}$$

where the third line follows by the mean value theorem and the fact that $|f(\xi)| = |l(\xi)| \cdot |V(F(\xi))| \in [1/(C_1 C_2), C_1 C_2]$ and the forth line follows by the assumption that $U \in [0,1]^k$. Similarly, we have

$$\begin{aligned} & E_0 \log(Q_{\theta,\eta}/Q_{\theta_0,\eta_0}) \\ \geqslant & - C_1^2 C_2 E_0 \big((\theta - \theta_0)^T U + \eta(V) - \eta_0(V)\big)^2. \end{aligned}$$

Let $\bar{\eta}(\theta)(v) = \eta_0(v) - (\theta - \theta_0) E[U|V = v]$. Then by definition of $\eta^*(\theta)$, we have

$$E_0 \log(Q_{\theta,\eta^*(\theta)}/Q_{\theta_0,\eta_0}) \geqslant E_0 \log(Q_{\theta,\bar{\eta}(\theta)}/Q_{\theta_0,\eta_0}).$$

Combining the above inequalities, we obtain

$$\begin{aligned} & - 2(C_1^2 C_2)^{-1} \big(|\theta - \theta_0|^2 + E_0|\eta^*(\theta) - \eta_0|^2\big) \\ \geqslant & - C_1^2 C_2 E_0 \big((\theta - \theta_0)^T U + \bar{\eta}(\theta)(V) - \eta_0(V)\big)^2 \\ = & - C_1^2 C_2 E_0 (U - E[U|V])^2 |\theta - \theta_0|^2, \end{aligned}$$

which implies

$$\eta^*(\theta) - \eta_0 = O(|\theta - \theta_0|). \tag{F.20}$$

For an arbitrary function $h(V)$ with $||h||$ small, consider

$$\widehat{g}_{\theta,\eta^*(\theta),t} = g_{\theta,\eta^*(\theta)} + th,$$

for t in a neighborhood of 0. Thus,

$$\frac{d}{dt} E_0 \log(Q_{\theta_0,\eta_0}/Q_{\theta,\eta})\big|_{t=0} = 0,$$

which implies

$$\begin{aligned} 0 =& E_0\big[(Y - F(g_{\theta,\eta^*(\theta)}))l(g_{\theta,\eta^*(\theta)})h(V)\big] \\ =& E_0\big\{E_0\big[(F(g_{\theta_0,\eta_0}) - F(g_{\theta,\eta^*(\theta)}))l(g_{\theta,\eta^*(\theta)})\big|V\big]h(V)\big\}. \end{aligned}$$

Since the above equality holds for any h, we have

$$E_0\big[(F(g_{\theta_0,\eta_0}) - F(g_{\theta,\eta^*(\theta)}))l(g_{\theta,\eta^*(\theta)})\big|V = v\big] = 0,\ a.s.$$

The above, (F.20) and Assumption 1(2) imply

$$\begin{aligned} E_0\big[f_0(T)l_0(T)(\theta - \theta_0)^T U\big|V = v\big] + E_0\big[f_0(T)l_0(T)\big|V = v\big](\eta^*(\theta) - \eta_0)(v) \\ = O(|\theta - \theta_0|^2), a.s. \end{aligned}$$

Thus

$$\eta^*(\theta)(v) = \eta_0(v) - (\theta - \theta_0)h^*(v) + O(|\theta - \theta_0|^2),\ \text{as } |\theta - \theta_0| \to 0,$$

with h^* defined by (7.6).

F.1.11 Proof of Theorem 59

To check Assumption 2, we apply the framework of Kleijn and van der Vaart (2006), where the posterior of a misspecified infinite-dimensional Bayesian model is shown to concentrate its mass near the points in the support of the prior that minimize the Kullback-Leibler (KL) divergence with respect to P_0. In the GPLM setting with quasi-likelihood (7.4), the KL divergence minimizer is exactly $(\theta_0, \eta_0$. To study the contraction rate, we can apply the Theorem 2.1 in Kleijn and van der Vaart (2006) (use the fact that $N_t(\epsilon, \mathcal{F}, d) \leqslant N(\epsilon, \mathcal{F}, d) \leqslant N(\epsilon, \mathcal{F}, ||\cdot||_\infty)$ with d the Hellinger distance) with some small modifications similar to the proof of (F.18). The following lemma shows the result.

Lemma 83. *Let $\mathcal{H}_n = \{\eta \in \mathcal{F}_n^\eta : ||\eta - \eta_0||_n \leqslant M\rho_n\}$, with $\mathcal{F}_n^\eta$ the function space defined by* (F.12) *in the proof of Theorem 57 and $\rho_n = n^{-\alpha/(2\alpha+1)}(\log n)^{d+1}$. Then*

the posterior of the Bayesian model specified in the previous subsection satisfies: for any $r > 0$, there exists a sufficiently large M, such that

$$\Pi\big(|\theta - \theta_0| \leqslant M\rho_n, \eta \in \mathcal{H}_n \big| X^{(n)}\big) = 1 - O_P(\exp(-n\rho_n^2)).$$

Next, we prove (A1). By Lemma 83, the posterior of η concentrates its mass in a small neighborhood $\mathcal{H}_n$ of η_0. Write $q_n(\theta, \eta) = \sum_{i=1}^n q_{\theta,\eta}(Y_i)$ and recall that $\Delta\eta(\theta) = \eta^*(\theta) - \eta_0 = (\theta - \theta_0)h(V) + O(|\theta - \theta_0|^2)$.

Lemma 84. *Assume Assumption 1. Then*

$$\begin{aligned} q_n\big(\theta_n, \eta + \Delta\eta(\theta_n)\big) - q_n(\theta_0, \eta) &= (\theta_n - \theta_0)^T \sum_{i=1}^n W_i l_0(T_i)(U_i + h^*(V_i)) \\ &\quad - \frac{1}{2} n(\theta_n - \theta_0)^T \widetilde{I}_{\theta_0,\eta_0}(\theta_n - \theta_0) + O_P[R_n(\max\{|\theta - \theta_0|, n^{-1/2}\log n\})], \end{aligned} \tag{F.21}$$

for every sequence $\{\theta_n\}$ such that $\theta_n = \theta_0 + O_P(\rho_n)$ in P_0 and uniformly for every $\eta \in \mathcal{H}_n$, with

$$\widetilde{I}_{\theta_0,\eta_0} = E_0\big[l_0(T) f_0(T)(U + h^*(V))(U + h^*(V))^T\big],$$

and

$$R_n(t) = nt^3 + \sqrt{n}t^2 + n\rho_n t^2 + n\rho_n^2 t + \sqrt{n}\rho_n^2.$$

Similar to Theorem 58, Theorem 59 can be proved by applying Lemma 47, Lemma 83, Lemma 84, Theorem 60 and Theorem 56.

F.1.12 Proof of Lemma 83

The verification of condition a and condition b in Lemma 51 is the same as the corresponding proof in Theorem 57. This time we apply Lemma 52. For GP priors, we have $\Pi(||\eta - \eta_0||_\infty \leqslant \rho_n) \geqslant \exp(-Cn\rho_n^2)$ (van der Vaart and van Zanten, 2009, Lemma 4.6). Since $||\cdot||_\infty$ is stronger than $||\cdot||_n$, condition e in Lemma 52 is satisfied.

F.1.13 Proof of Lemma 84

Using the definition of q_n and $q_{\theta,\eta}$, we get

$$q_n\big(\theta,\eta+\Delta\eta(\theta)\big)-q_n(\theta_0,\eta)=\sum_{i=1}^{n}W_i\int_{m_{\theta_0,\eta}(T_i)}^{m_{\theta,\eta+\Delta\eta(\theta)}(T_i)}\frac{1}{V(s)}ds$$
$$-\sum_{i=1}^{n}\int_{m_{\theta_0,\eta}(T_i)}^{m_{\theta,\eta+\Delta\eta(\theta)}(T_i)}\frac{(s-m_0(T_i))}{V(s)}ds\triangleq I-II,$$

with $W_i=Y_i-m_0(T_i)$ satisfying $E_0W_i=0$ and Assumption 1(1).

By applying Taylor expansions and Assumption 1(2), we have for any $\xi_0,\xi_1,\xi_2\in\mathbb{R}$,

$$\begin{aligned}\int_{F(\xi_1)}^{F(\xi_2)}\frac{1}{V(s)}ds=&l(\xi_1)(\xi_2-\xi_1)+e_1(\xi_0)(\xi_2-\xi_1)^2+O((\xi_2-\xi_1)^3)\\=&l(\xi_0)(\xi_2-\xi_1)+e_1(\xi_0)(\xi_2-\xi_1)^2+e_2(\xi_0)(\xi_2-\xi_1)(\xi_1-\xi_0)\\&+O\big\{(\xi_2-\xi_1)^3+(\xi_2-\xi_1)^2(\xi_1-\xi_0)+(\xi_2-\xi_1)(\xi_1-\xi_0)^2\big\},\end{aligned}\tag{F.22}$$

$$\begin{aligned}\int_{F(\xi_1)}^{F(\xi_2)}\frac{s-F(\xi_0)}{V(s)}ds=&l(\xi_1)\big(F(\xi_1)-F(\xi_0)\big)(\xi_2-\xi_1)+\frac{1}{2}l(\xi_1)f(\xi_1)(\xi_2-\xi_1)^2\\&+O\big\{(\xi_2-\xi_1)^3+(\xi_2-\xi_1)^2(\xi_1-\xi_0)\big\}\\=&l(\xi_0)f(\xi_0)(\xi_2-\xi_1)(\xi_1-\xi_0)+\frac{1}{2}l(\xi_0)f(\xi_0)(\xi_2-\xi_1)^2\\&+O\big\{(\xi_2-\xi_1)^3+(\xi_2-\xi_1)^2(\xi_1-\xi_0)+(\xi_2-\xi_1)(\xi_1-\xi_0)^2\big\},\end{aligned}\tag{F.23}$$

with $e_1(\xi)$ and $e_2(\xi)$ fixed bounded functions.

By the definition of $g_{\theta,\eta}$ and $\Delta\eta(\theta)$, we have

$$g_{\theta_0,\eta}(T)-g_0(T)=(\eta-\eta_0)(V),$$
$$g_{\theta,\eta+\Delta\eta(\theta)}(T)-g_{\theta_0,\eta}(T)=(\theta-\theta_0)^Th_1(T)+O(|\theta-\theta_0|^2),$$

with $h_1(T) = U + h^*(V)$. Combining the above and the definition of l_0, f_0 and $m_{\theta,\eta}$, and (F.22) with $\xi_0 = g_0$, $\xi_1 = g_{\theta_0,\eta}$ and $\xi_2 = g_{\theta,\eta+\Delta\eta(\theta)}$, we get

$$I = (\theta - \theta_0)^T \sum_{i=1}^n W_i l_0(T_i) h_1(T_i)$$

$$+ (\theta - \theta_0)^T \sum_{i=1}^n W_i e_2(g_0(T_i)) h_1(T_i)(\eta - \eta_0)(V_i) + O_P(\sqrt{n}|\theta - \theta_0|^2),$$

where the last term is obtained by combining the central limit theorem and the fact that $E_0 W_i = 0$ and $E_0 W_i^2 < \infty$.

Since $E_0 W_i e_2(g_0(T_i)) h_1(T_i) = E_0[e_2(g_0(T_i)) h_1(T_i) E_0(W_i|T_i)] = 0$, similar to (F.19), by applying the maximal inequality, we get

$$E_0\left\{ \sup_{\eta \in H_n} \frac{1}{\sqrt{n}} \Big| \sum_{i=1}^n W_i e_2(g_0(T_i)) h_1(T_i)(\eta - \eta_0)(V_i) \Big| \Big| V_1, \ldots, V_n \right\}$$

$$\lesssim \int_0^{\rho_n} \sqrt{1 + \log N(\epsilon, H_n, ||\cdot||_\infty)} d\epsilon$$

$$\lesssim \sqrt{n}\rho_n^2 + \rho_n \backsim \sqrt{n}\rho_n^2.$$

Combining the above two, we get

$$I = (\theta - \theta_0)^T \sum_{i=1}^n W_i l_0(T_i) h_1(T_i) + O_P\big\{\sqrt{n}|\theta - \theta_0|^2 + n|\theta - \theta_0|\rho_n^2\big\}. \tag{F.24}$$

Similarly, using (F.23) and the same choices for ξ_0, ξ_1 and ξ_2, we get

$$II = (\theta - \theta_0)^T \sum_{i=1}^n l_0(T_i) f_0(T_i)(U_i + h^*(V_i))(\eta - \eta_0)(V_i)$$

$$+ \frac{1}{2} \sum_{i=1}^n l_0(T_i) f_0(T_i) \big((\theta - \theta_0)^T h_2(T_i)\big)^2$$

$$+ O_P\big\{n|\theta - \theta_0|^3 + n|\theta - \theta_0|^2 \rho_n + n|\theta - \theta_0|\rho_n^2\big\},$$

where $h_2(t) = u - E[U|V = v]$. By definition of h^*, we have

$$E_0\big[l_0(T_i) f_0(T_i)(U_i + h^*(V_i))(\eta - \eta_0)(V_i)\big]$$

$$= E_0\big[(\eta - \eta_0)(V_i) E_0(l_0(T_i) f_0(T_i)(U_i + h^*(V_i))|V_i)\big] = 0.$$

Therefore, by applying the maximal inequality, we get

$$E_0\Big\{ \sup_{\eta\in H_n} \frac{1}{\sqrt{n}} \Big| \sum_{i=1}^{n} l_0(T_i) f_0(T_i)(U_i + h^*(V_i))(\eta - \eta_0)(V_i) \Big| \Big| V_1, \ldots, V_n \Big\}$$

$$\lesssim \int_0^{\rho_n} \sqrt{1 + \log N(\epsilon, H_n, ||\cdot||_\infty)} d\epsilon$$

$$\lesssim \sqrt{n}\rho_n^2 + \rho_n \backsim \sqrt{n}\rho_n^2.$$

By the central limit theorem, we have

$$\frac{1}{2}\sum_{i=1}^{n} l_0(T_i) f_0(T_i) \big((\theta - \theta_0)^T h_2(T_i)\big)^2$$

$$= \frac{n}{2}(\theta - \theta_0)^T E_0\big[l_0(T) f_0(T) h_1(T)(h_1(T))^T\big](\theta - \theta_0) + O_P(\sqrt{n}|\theta - \theta_0|^2)$$

Combining the above, we have

$$II = \frac{n}{2}(\theta - \theta_0)^T E_0\big[l_0(T) f_0(T) h_1(T)(h_1(T))^T\big](\theta - \theta_0)$$
$$+ O_P\big\{n|\theta - \theta_0|^3 + \sqrt{n}|\theta - \theta_0|^2 + n|\theta - \theta_0|^2\rho_n + n|\theta - \theta_0|\rho_n^2 + \sqrt{n}\rho_n^2\big\}. \quad \text{(F.25)}$$

By (F.24) and (F.25), the lemma is proved.

Bibliography

Aronszajn, N. (1950), "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, 68, 337–404.

Baraniuk, R. G. and Wakin, M. B. (2009), "Random projections of smooth Manifolds," *Found. Comput. Math.*, 9, 51–77.

Barbieri, M. M. and Berger, J. O. (2004), "Optimal Predictive Model Selection," *Ann. Statist.*, 32, 870–897.

Belkin, M. (2003), "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, 15, 1373–1396.

Bhattacharya, A., Pati, D., and Dunson, B. D. (2012), "Adaptive dimension reduction with a Gaussian process prior," *arXiv: 1111.1044*.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2013), "Bayesian shrinkage," *arXiv:1212.6088*.

Bickel, J. P. and Li, B. (2007), "Local polynomial regression on unknown manifolds," *Complex datasets and inverse problem: tomography, networks and beyond, IMS Lecture Notes-Monograph Series,*, 54, 177–186.

Bickel, P. and Kleijn, B. (2012), "The semiparametric Bernstein-von Mises theorem," *Ann. Statist.*, 40, 206–237.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Stat.*, 37, 1705–1732.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer-Verlag, New York.

Birgé, L. (2004), "Model selection for Guassian regression with random design," *Bernoulli*, 10, 1039–1151.

Birgé, L. (2006), "Model selection via testing: An alternative to (penalized) maximum likelihood estimators," *Ann. Inst. H. Poincar Probab. Statist.*, 42, 273–325.

Birman, M. S. and Solomjak, M. J. (1967), "Piece-wise polynomial approximations of functions of the class W_p^α," *Mat. Sbornik*, 3, 295–317.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Boente, G., He, X., and Zhou, J. (2006), "Robust estimates in generalized partially linear models," *Ann. Statist.*, 34, 2856–2878.

Bowman, A. W. and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Clarendon Press, Oxford.

Breiman, L. (2001), "Random Forests," *Mach. Learn.*, 45, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth, Inc, Belmont, CA.

Bühlmann, P. (2006), "Boosting for High-Dimensional Linear Models," *Ann. Statist.*, 34, 559–583.

Bülmann, P. and van de Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer, Heidelberg; New York.

Bunea, F. and Nobel, A. (2008), "Sequential procedures for aggregating arbitrary estimators of a conditional mean," *IEEE Transactions on Information Theory*, 54, 1725–1735.

Bunea, F. and Tsybakov (2007), "Aggregation for Gaussian regression," *Ann. Statist.*, 35, 1674–1697.

Camastra, F. and Vinviarelli, A. (2002), "Estimating the intrinsic dimension of data with a fractal-based method," *IEEE P.A.M.I.*, 24, 1404–1407.

Candes, E. and Tao, T. (2007), "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, 35, 2313–2351.

Carter, K. M., Raich, R., and Hero, A. O. (2010), "On local intrinsic dimension estimation and its applications," *Trans. Sig. Proc.*, 58, 650–663.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The horseshoe estimator for sparse signals," *Biometrika*, 97, 465–480.

Castillo, I. (2012), "Semiparametric BernsteinCvon Mises theorem and bias, illustrated with Gaussian process priors," *The Indian Journal of Statistics*, 74, 194–221.

Castillo, I. and van der Vaart, A. (2012), "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *Submitted to Ann. Statist.*

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D. B., and Carin, L. (2010), "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Trans. Signal Process*, 58, 6140–6155.

Cheng, G. and Kosorok, M. R. (2008a), "General Frequentist Properties of the Posterior Profile Distribution," *Ann. Statist.*, 36, 1819–1853.

Cheng, G. and Kosorok, M. R. (2008b), "Higher Order Semiparametric Frequentist Inference with the Profile Sampler," *Ann. Statist.*, 36, 1786–1818.

Cheng, G. and Kosorok, M. R. (2009), "The penalized profile sample," *Journal of Multivariate Analysis*, 100, 345–362.

Chipman, H. A., George, E. I., and E., M. R. (2010), "BART: Bayesian Additive Regression Trees," *Ann. Appl. Stat.*, 4, 266–298.

Chopin, M. (2002), "A sequential particle filter for static models," *Biometrika*, 89, 539–551.

Cohen, J. E. and Rothblum, U. G. (1993), "Nonnegative Ranks, Decompositions, and Factorizations of Nonnegative Matrices," *Linear Algebra Appl.*, 190, 37.

Comminges, L. and Dalalyan, A. S. (2012), "Tight conditions for consistent variable selection in high dimensionality," *Annals of statistics*, 40, 2667–2696.

Cornelis, B., **Yang, Y.**, Vogelstein, J. T., Dooms, A., Daubechies, I., and Dunson, B. D. (2013), "Bayesian crack detection in ultra high resolution multimodal images of paintings," *accepted by International Conference on Digital Signal Processing*.

Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, John Wilet and Sons, New York.

Datta, G. S. and Mukerjee, R. (2004), *Probability matching priors: Higher order asymptotics*, New York: Springer-Verlag.

de Jonge, R. and van Zanten, H. (2013), "Semiparametric BernsteinCvon Mises for the error standard deviation," *Electronic Journal of Statistics*, 7, 217–243.

De Lathauwer, L., De Moor, B., and Vanderwalle, J. (2000), "A multilinear Singular Value Decomposition," *SIAM J. Matrix Anal. Appl.*, 21, 1253–1278.

Del Moral, P., Doucet, A., and Jasra, A. (2006), "Sequential Monte Carlo samplers," *J. R. Statist. Soc. B*, 68, 411–436.

do Carmo, M. (1992), *Riemannian geometry*, Birkhauser, Boston.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," *J. Amer. Statist. Assoc.*, 104, 1042–1051.

Earlab, D. J. and Deema, M. W. (2005), "Parallel tempering: Theory, applications, and new perspectives," *Phys. Chem. Chem. Phys.*, 7, 3910–3916.

Eldén, L. and Savas, B. (2009), "A Newton-Grassmann method for computing the Best Multilinear Rank-(r_1,r_2,r_3) Approximation of a Tensor," *SIAM J. MATRIX ANAL. APPL.*, 31, 248–271.

Fan, J., Lv, J., and Qi, L. (2011), "Sparse High-Dimensional Models in Economics," *Annu. Rev. Econ.*, 3, 291–317.

Geman, S. and Hwang, C. (1982), "Nonparametric maximum likelihood estimation by the method of sieves," *Ann. Statist.*, 10, 401–414.

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304.

George, E. and McCulloch, R. (1997), "Approaches for Bayesian Variable Selection," *Statist. Sinica*, 7, 339–373.

Geyer, C. J. (1991), "Markov chain Monte Carlo maximum likelihood," *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163.

Ghosal, S. and van der Vaart, A. W. (2007), "Convergence rates of posterior distributions for noniid observations," *Ann. Statist.*, 35, 192–233.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000), "Convergence rates of posterior distributions," *Ann. Statist.*, 28, 500–531.

Gine, E. and Koltchinskii, V. (2005), "Empirical graph Laplacian approximation of Laplace Beltrami operators: Large sample results," *High Dimensional Probability IV*, 51, 238–259.

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Guedj, B. and Alquier, P. (2013), "PAC-Bayesian estimation and prediction in sparse additive models," *Electronic Journal of Statistics*, 7, 264–291.

Hansen, B. (2012), *Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation*, Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Forthcoming.

Harshman, R. (1970), "Foundations of the PARAFAC Procedure: Models and Conditions for an 'exploratory' Multi-modal Factor Analysis," *UCLA working papers in phonetics*, 16, 1–84.

Harshman, R. and Lundy, M. (1994), "Parallel Factor Analysis," *Comput. Statist. Data Anal.*, 18, 39–72.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1, 297–310.

Hastie, T. and Tibshirani, R. (1987), "Nonparametirc logistic and proportional odds regression," *Applied Statistics*, 36, 260–276.

Hiai, F. and Petz, D. (2009), "Riemannian metrics on positive definite matrices related to means," *Linear Algebra and its Applications*, 430, 3105–3130.

Hoerl, A. and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417.

Hoffman, M., Blei, D., and Bach, F. (2010), "Online Learning for Latent Dirichlet Allocation," *Neural Information Processing Systems*.

Ishwaran, H. and Rao, J. S. (2005), "Spike and slab variable selection: Frequentist and Bayesian strategies," *Ann. Statist.*, 33, 730–773.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005), "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling," *Statistical Science*, 20, 50–67.

Jiang, W. (2006), "Bayesian Variable Selection for High Dimensional Generalized Linear Models," *Ann. Statist.*, 35, 1487–1511.

Johnson, R. A. (1970), "Asymptotic Expansions Associated with Posterior Distributions," *Ann. Math. Statist.*, 41, 851–864.

Juditsky, A. and Nemirovski, A. (2000), "Functional aggregation for nonparametric regression," *Ann. Statist.*, 28, 681–712.

Kent, J. T. (1992), "New directions in shape analysis," in *The Art of Statistical Science. A Tribute to G. S. Watson*, pp. 115–127, New York, Wiley.

Kim, Y. D. and Choi, S. (2007), "Nonnegative Tucker decomposition," in *Proceedings of the IEEE CVPR-2007 Workshop on Component Analysis Methods*, Minneapolis, Minnesota, USA.

Kleijn, B. J. K. and van der Vaart, A. W. (2006), "Misspecification in infinite-dimensional Bayesian statistics," *Ann. Statist.*, 34, 837–877.

Koltchinskii, V. and Yuan, M. (2010), "Sparsity in Multiple Kernel Learning," *Ann. Stat.*, 38, 3660–3695.

Koltchiskii, V. and Pancheko, D. (2005), "Complexity of convex combinations and bound the generalization error in classification," *Annals of statistics*, 33, 1455–1496.

Kou, S., Zhou, Q., and Wong, H. W. (2006), "Equi-energy sampler with applications in statistical inference and statistical mechanics," *Ann. Statist.*, 34, 1581–1652.

Kuelbs, J. abd Li, W. V. and Linde, W. (1994), "The Gaussian measure of shifted balls," *Probab. Theory Related Fields*, 98, 143–162.

Kundu, S. and Dunson, D. B. (2011), "Latent factor models for density estimation," *arXiv:1108.2720v2*.

Lafferty, J. and Wasserman, L. (2008), "Rodeo: Sparse, greedy nonparametric regression," *Ann. Stat.*, 36, 28–63.

Lawrence, N. D. (2003), "Gaussian process latent variable models for visualisation of high dimensional data," *Neural Information Processing Systems*.

LeCam, L. (1953), "Locally asymptotically normal families of distributions," *Univ. California Publ. Statist.*, 3, 37–98.

Lehmann, E. L. and Casella, G. (1998), *Theory of point estimation*, New York: Springer-Verlag.

Levina, E. and Bickel, P. (2004), "Maximun likelihood estimation of intrinsic dimension," in *Advances in Neural Information Processing Systems*, vol. 17, Cambridge, MA, USA, The MIT Press.

Li, W. V. and Linde, W. (1999), "Approximation, metric entropy and small ball estimates for Gaussian measures," *The Annals of Probability*, 27, 1556–1578.

Little, A. V., Lee, J., Jung, Y. M., and Maggioni, M. (2009), "Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale SVD," in *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pp. 85–88.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, Springer, New York.

Liu, J. S. and Chen, R. (1998), "Sequential Monte Carlo Methods for Dynamic Systems," *Journal of the American Statistical Association*, 93, 1032–1044.

Liu, S. J., Wong, H. W., and Kong, A. (1995), "Covariance structure and convergence rate of the Gibbs sampler with various scans," *J. R. Statist. Soc. B*, 57, 157–169.

Mammen, E. and van de Geer, S. (1997), "Penalized quasi-likelihood estimation in partial linear models," *Ann. Statist.*, 25, 1014–1035.

Meier, L. van de Geer, S. and Buhlmann, P. (2009), "High-dimensional Additive Modeling," *Ann. Stat.*, 37, 3779–3821.

Meinshausen, N. and Yu, B. (2009), "Lasso-type recovery of sparse representations for high-dimensional data," *Ann. Stat.*, 37, 246–270.

Mengersen, K. L. and Tweedie, R. L. (1996), "Rates of convergence of the Hasting and Metropolis algorithms," *The Annals of Statistics*, 24, 101–121.

Meyn, S. P. and Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer, New York.

Moakher, M. and Zéraï, M. (2011), "The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data," *Journal of Mathematical Imaging and Vision*, 40, 171–187.

Nash, J. (1956), "The imbedding problem for Riemannian manifolds," *Annals of Mathematics*, 63, 20–63.

Neal, R. (2001), "Annealed importance sampling," *Statist. Comput.*, 11, 125–139.

Neal, R. (2003), "Slice sampling," *The Annals of Statistics*, 3, 705–767.

Nemirovski, A. (2000), *Topics in non-parametric statistics.*, Lectures on Probability Theory and Statistics (Saint-Flour, 1998). ecture Notes in Math. 1738, Springer, Berlin.

Nene, S. A., Nayar, S. K., and Murase, H. (1996), "Columbia object image library (COIL-100)," Tech. rep., Columbia University.

Page, G., Bhattacharya, A., and Dunson, D. B. (2013), "Classification via Bayesian nonparametric learning of affine subspaces," *J. Amer. Statist. Assoc.*, 108, 187–201.

Park, M. Y. and Hastie, T. (2007), "L-1 regularization path algorithm for generalized linear models," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69, 659–677.

Polson, N. G. and Scott, J. G. (2010), *Shrink globally, act locally: Sparse Bayesian regularization and prediction*, Bayesian Statistics 9 (J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds.), Oxford University Press, New York.

Raskutti, G., Wainwright, M., and Yu, B. (2011), "Minimax Rates of Estimation for High-Dimensional linear regression over l_q-balls," *IEEE transactions on information theory*, 57, 6976–6994.

Raskutti, G., Wainwright, M. J., and Yu, B. (2012), "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," *Journal of machine learning research*, 13, 389–427.

Rasmussen, C. E. and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, Massachusetts.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2009), "SpAM: Sparse Additive Models," *Journal of the Royal Statistical Society*, 71, 1009–1030.

Reich, B. J., Bondell, H. D., and Li, L. X. (2011), "Sufficient dimension reduction via Bayesian mixture modeling," *Biometrics*, 67, 886–895.

Richardson, S. and Green, P. J. (1997), "On Bayesian analysis of mixtures with an unknow number of components," *J. R. Statist. Soc. B*, 88, 450–456.

Rivoirard, V. and Rousseau, J. (2012), "Bernstein von Mises theorem for linear functionals of the density," *Ann. Statist.*, 40, 1489–1532.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Roberts, O. G. and Rosenthal, S. J. (1997), "Geometric ergodicity and hybrid Markov chains," *Elect. Comm. in Probab.*, 2, 13–25.

Rosenthal, J. S. (1995), "Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558–566.

Rousseau, J. and Mengersen, K. (2011), "Asymptotic behaviour of the posterior distribution in overfitted mixture models," *J. R. Statist. Soc. B*, 73, 689–710.

Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P., and Ferreira, J. (1983), "Coronary risk factor screening in three rural communities," *South African Medical Journal*, 64, 430–436.

Roweis, S. T. and Saul, L. K. (2000), "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 290, 2323–2326.

Savitsky, T., Vannucci, M., and Sha, N. (2011), "Variable selection for nonparametric Gaussian process priors: models and computational strategies," *Statistical Science*, 26, 130–149.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Severini, T. A. and Wong, W. H. (1992), "Profile likelihood and conditionally parametric models," *Ann. Statist.*, 20, 1768–1802.

Shen, X. (2001), "Asymptotic normality of semiparametric and nonparametric posterior distributions," *Journal of the American Statistical Association*, 97, 222–235.

Staicu, A. M. and Reid, N. M. (2008), "On probability matching priors," *The Canadian Journal of Statistics*, 36, 613–622.

Stephan, F. F. (1945), "The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate," *Ann. Math. Statist.*, 16, 50–61.

Stone, C. J. (1982), "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, 10, 1040–1053.

Storvik, G. (2002), "Particle Filters for State-Space Models With the Presence of Unknown Static Parameters," *IEEE Transactions on Signal Processing*, 50, 281–289.

Tenenbaum, J. B., Silva, V., and Langford, J. C. (2000), "A global geometric framework for nonlinear dimensionality reduction," *Science*, 290, 2319–2323.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Association*.

Tokdar, S. (????), "Dimension adaptability of Gaussian process models with variable selection and projection," *arXiv: 1112.0716*.

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010), "Bayesian density regression with logistic Gaussian process and subspace projection," *Bayesian Anal.*, 5, 319–344.

Tsybakov, A. (2003), "Optimal Rates of Aggregation," *Learning Theory and Kernel Machines, Lecture Notes in Computer Science,*, 2777, 303–313.

Tsybakov, B. (2009), *Introction to nonparametric estimation*, Springer, New York.

Tucker, L. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311.

van de Geer, S. (2000), *Empirical processes in M-estimation*, Cambridge University Press, Cambridge, UK.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007), "Super learner," *Genetics and Molecular Biology*, 6.

van der Vaart, A. W. and van Zanten, J. H. (2008a), "Rate of contraction of posterior distributions based on Gaussian process priors," *Ann. Statist.*, 36, 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2008b), "Reproducing kernel Hilbert spaces of Gaussian priors," *IMS Collections*, 3, 200–222.

van der Vaart, A. W. and van Zanten, J. H. (2008c), "Reproducing kernel Hilbert spaces of Gaussian priors," *IMS Collections*, 3, 200–222.

van der Vaart, A. W. and van Zanten, J. H. (2009), "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth," *Ann. Statist.*, 37, 2655–2675.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. A. (2000), *Weak convergence and empirical processes: with applications to statistics, 2nd ed*, Springer.

Vannieuwenhoven, N., Vandebril, R., and Meerbergen, K. (2012), "A New Truncation Strategy for the Higher-Order Singular Value Decomposition," *SIAM J. Sci. Comput.*, 34, 1027–1052.

Verzelen, N. (2012), "Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons," *arXiv:1008.0526*.

Wainwright, M. (2009a), "Information-theoretic limits on sparsity recovery in the high-dimensinal and noisy setting," *IEEE transactions on information theory*, 55, 5728–5741.

Wainwright, M. (2009b), "Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1 constrained quadratic programming (lasso)," *IEEE transactions on information theory*, 55, 2183–2202.

Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gaussian-Newton method," *Biometrika*, 61, 439–447.

Wegkamp, M. (2003), "Model selection in nonparametric regression," *Ann. Statist.*, 31, 252–273.

West, M. (1993), "Approximating posterior distributions by mixtures," *J. R. Statist. Soc. B*, 55, 409–422.

Wu, T. T., Chen, Y. F., and Hastie, T. (2009), "Genome-wide Association Analysis by Lasso Penalized Logistic Regression," *Bioinformatics*, 25, 714–721.

Yang, C., Wan, X., and Yang, Q. (2010), "Identifying Main Effects and Epistatic Interactions from Large-scale SNP Data via Adaptive Group Lasso," *BMC Bioinformatics*, 11.

Yang, Y. (2000), "Combining different procedures for adaptive regression," *J. Multivariate Anal.*, 74, 135–161.

Yang, Y. (2001), "Adaptive regression by mixing," *J. Amer. Statist. Assoc.*, 96, 574–588.

Yang, Y. (2004), "Aggregating regression procedures to improve performance," *Bernoulli*, 10, 25–47.

Yang, Y. and Barron, A. (1999), "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, 27, 1564–1599.

Yang, Y. and Dunson, B. D. (2013), "Bayesian Manifold Regression," *arXiv:1305.0617*.

Ye, G. and Zhou, D. (2008), "Learning and approximation by Gaussians on Riemannian manifolds," *Adv. Comput. Math.*, 29, 291–310.

Yu, B. (1997), "Assouad, Fano, and Le Cam," in *Frestchrift for Lucien Le Cam*, pp. 423–435, New York, Springer.

Zhang, C. H. and Huang, J. (2008), "The sparsity and bias of Lasso selection in high-dimensional linear regression," *Ann. Stat.*, 36, 1567–1594.

Zhang, T. and Golub, G. H. (2001), "Rank-one Approximation to High Order Tensors," *SIAM J. Matrix Anal. Appl.*, 23, 534.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67, 301–320.

Biography

Yun Yang was born in 1988 in Shanghai, China. In July 2011, he received his Bachelors degree in Mathematics from Tsinghua University, Beijing, China. Yun moved to Durham, NC in August 2011 to pursue doctoral studies at the Department of Statistical Science at Duke University.