

ABSTRACT

WANG, KANG. Nonparametric Bayesian Inference Under Shape Restrictions. (Under the direction of Subhashis Ghosal).

Shape restrictions, like monotonicity, imposed on a function of interest, such as a regression or density function, allow for estimation without smoothness assumptions. In the multivariate isotonic regression problem, where the regression function is assumed to be coordinatewise nondecreasing, we propose a Bayesian approach, obtain the posterior contraction rate, construct a universally consistent Bayesian testing procedure, and study the coverage of pointwise credible interval. We set aside the shape restrictions temporarily and endow a prior on block-wise constant regression functions with heights independently normally distributed. The unrestricted block-heights are a posteriori also independently normally distributed by conjugacy. To comply with the shape restrictions, we either project samples from the unrestricted posterior onto the class of multivariate monotone functions or convert the unrestricted posterior samples by a more general “immersion map”. Under an \mathbb{L}_1 -metric, we show that the projection-posterior based on n independent samples contracts around the true monotone regression function at the optimal rate $n^{-1/(2+d)}$. We construct a Bayesian test for multivariate monotonicity based on the posterior probability of a shrinking neighborhood of the class of multivariate monotone functions. We show that the test is universally consistent. We show that the power goes to one against a smooth function class well-separated from the null. To study the coverage of pointwise credible intervals, a block isotonization immersion map is found to be useful. We establish a key weak convergence for the posterior distribution function at a point, leading to an expression for the limiting coverage of the Bayesian credible interval. We find that the limiting coverage is higher than the credibility. Interestingly, the relationship between credibility and limiting coverage does not involve any unknown parameter. Hence by recalibration, we can get a predetermined asymptotic coverage by choosing a credibility level lower than the targeted coverage, thus shortening the credible intervals.

In the multivariate monotone density estimation problem, we consider a nonparametric Bayesian approach. We put a prior on the step heights through binning and a Dirichlet distribution. An arbitrary piece-wise constant probability density is converted to a monotone one by a projection map, taking its \mathbb{L}_1 -projection onto the space of monotone functions, which is subsequently normalized. We construct consistent Bayesian tests for the multivariate monotonicity of a probability density. The test is shown to be consistent. We apply block operations to monotonize the unrestricted posterior samples in studying the coverage. The limiting coverage is explicitly calculated and is higher than the credibility level. By exploring the asymptotic relationship between the coverage and the credibility, we show that a desired asymptomatic coverage can be obtained exactly by starting with an appropriate credibility level.

The concept of k -monotonicity encompasses a family of monotone shape restrictions, including decreasing and convex decreasing as special cases corresponding to $k = 1$ and $k = 2$. We consider Bayesian approaches to estimate a k -monotone density. By utilizing a kernel mixture representation and putting a Dirichlet process or a finite mixture prior on the mixing distribution, we show that the posterior contraction rate in the Hellinger distance is $(n/\log n)^{-k/(2k+1)}$ for a k -monotone density, which is minimax optimal up to a polylogarithmic factor. When the true k -monotone density is a finite J_0 -component mixture of the kernel, the contraction rate improves to the nearly parametric rate $\sqrt{(J_0 \log n)/n}$. By putting a prior on k , we show that the same rates hold even when k is unknown. Applications in modeling the density of p -values in a large-scale multiple testing problem are considered.

© Copyright 2023 by Kang Wang

All Rights Reserved

Nonparametric Bayesian Inference Under Shape Restrictions

by
Kang Wang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2023

APPROVED BY:

Sujit Ghosh

Ryan Martin

Jonathan Williams

Subhashis Ghosal
Chair of Advisory Committee

BIOGRAPHY

The author was born and raised in China. He earned his bachelor's and master's degrees in statistics from Nankai University in 2015 and 2018, respectively. In 2018, he joined the Department of Statistics at North Carolina State University to pursue a Ph.D. degree in statistics.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Dr. Subhashis Ghosal, for his guidance, support, and patience. Without his dedicated assistance, this work would not have been possible. During the three years of working with my advisor, I have grown and learned a lot. I am sure that this experience will have lifelong benefits. I would also like to extend my thanks to the committee members, Drs. Sujit Ghosh, Ryan Martin, and Jonathan Williams. I am grateful for the time and effort they devoted to carefully reading my work and providing me with detailed feedback and insightful comments, all of which have significantly contributed to improving the quality of this work. I would also like to extend my gratitude to my teachers, Professors Zhaojun Wang, Minqian Liu, and Changliang Zou, at Nankai University for their continuous support and encouragement. I would like to sincerely thank my dear friends Ke Zhao, Lin Xiao, Xiaoqian Liu, Yuqi Su, Cheng Wang, Yanzhao Wang, and all my fellow classmates. Their constant assistance, candid advice, inspirational mindset, unwavering support, and integrity have significantly enhanced my journey. Lastly, my heartfelt gratitude goes to my family. Their unconditional love and unwavering support are a constant source of strength.

TABLE OF CONTENTS

List of Tables	vi
List of Figures	vii
Chapter 1 INTRODUCTION	1
1.1 Statistical inference under shape restrictions	1
1.1.1 Examples	3
1.1.2 Least squares estimator and maximum likelihood estimator	7
1.1.3 Bayesian shape-restricted inference	10
1.2 Bayesian nonparametrics	11
1.2.1 Nonparametric modeling methods	12
1.2.2 Frequentist asymptotic properties of posterior measures	14
1.3 Immersion posterior-based Bayesian inference	17
1.4 Notation	20
1.5 Chapter organization	21
 Chapter 2 Multivariate isotonic regression: contraction rates, tests, and frequentist coverage	 23
2.1 Models, priors, and posteriors	23
2.2 Contraction rates of projection posterior	26
2.3 Bayesian tests for multivariate monotonicity	31
2.4 Immersion posterior and frequentist coverage	34
2.4.1 Effect of the immersion map	36
2.4.2 Coverage of Credible Intervals	36
2.5 Simulation	46
2.5.1 Distribution of Z_B	46
2.5.2 Simulation for posterior contraction rate	49
2.5.3 Simulation for Bayesian monotonicity testing	51
2.5.4 Coverage comparison with Deng, Han and Zhang’s method	55
2.5.5 Length comparison with the oracle	56
2.5.6 Effects of the choice of J	58
2.6 Proofs	59
2.6.1 Proofs of results in Section 2.2	59
2.6.2 Proofs of results in Section 2.3	64
2.6.3 Proofs of results in Section 2.4	69
 Chapter 3 Multivariate monotone densities: contraction rates, tests, and frequentist coverage	 89
3.1 Prior and posterior	90
3.2 Contraction rates and testing for multivariate monotonicity	92
3.3 Coverage of pointwise credible intervals	95
3.4 Simulation	99
3.4.1 Global deviation of immersion posteriors	99

3.4.2	Coverage of credible intervals	100
3.5	Proofs	103
Chapter 4	Bayesian Inference for k-Monotone Densities with Applications to Multiple Testing	133
4.1	Definitions and characterization	133
4.2	Posterior Contraction Rates	137
4.3	Adaptation to k	139
4.4	Applications to Multiple Testing	141
4.5	Simulation	143
4.5.1	Estimation accuracy	143
4.5.2	Estimation of the proportion of null hypotheses	145
4.6	Proofs	147
References		162
APPENDICES		169
Appendix A	Auxiliary results for multivariate monotone function estimation	170
A.1	Supporting results for Section 2.2 and 2.3	170
A.2	Supporting results for Section 2.4	175
A.3	Supporting results for Chapter 3	182
Appendix B	Auxiliary results for k -monotone density estimation	185
B.1	Supporting results for Chapter 4	185
B.2	Discrete approximation	188

LIST OF TABLES

Table 1.1	First-year GPA of 2397 freshmen in the University of Iowa in the fall of 1978.	7
Table 2.1	Values of $P(Z_B \leq z)$	48
Table 2.2	Values of $q = \inf\{z : P(Z_B \leq z) \geq p\}$	48
Table 2.3	Values of $P(Z_B \leq z)$ for various z and β , and $Z_B = Z_B^{(1)}, Z_B^{(2)}, Z_B^{(3)}$	49
Table 2.4	Values of $q = \inf\{z : P(Z_B \leq z) \geq p\}$ for various p , and $Z_B = Z_B^{(1)}, Z_B^{(2)}, Z_B^{(3)}$	49
Table 2.5	The Lebesgue \mathbb{L}_1 -distance between the Bayesian projection posterior mean regression function (BP) and the true regression function and between the least squares isotonic regression function (LS) and the true one with standard deviations across all data sets marked in the parentheses.	52
Table 2.6	Percentage of rejections to the null hypothesis of Bayesian projection posterior procedure (BP), linear regression procedure (LR), and piecewise linear fitting (PL) when the true regression functions are coordinatewise increasing.	53
Table 2.7	Percentage of rejections to the null hypothesis of Bayesian projection posterior procedure (BP), linear regression procedure (LR), and piecewise linear fitting (PL) when the true regression functions are not coordinatewise increasing.	54
Table 2.8	Coverage percentage (C) and length (L) comparison,	57
Table 3.1	\mathbb{L}_1 -metric of unrestricted posterior and immersion posterior.	101
Table 3.2	Coverage and length of credible intervals for $g_1(\mathbf{x}_0)$	102
Table 3.3	Coverage and length of credible intervals for $g_2(\mathbf{x}_0)$	103
Table 3.4	Coverage and length of credible intervals for $g_3(\mathbf{x}_0)$	104
Table 3.5	Coverage and length of credible intervals for $g_4(\mathbf{x}_0)$	105
Table 4.1	Average of MSE.	145

LIST OF FIGURES

Figure 1.1	Bivariate isotonic regression function for the first-year GPA data	8
Figure 1.2	Baseball salary data: The black solid squares represent the raw data points, and the red cross signs indicate the fitted salary at every observation.	9
Figure 2.1	Unrestricted and immersion posterior density functions of $f(x_0)$	37
Figure 2.2	Distribution functions of Z_B	47
Figure 2.3	Distribution functions of $Z_B^{(1)}$, $Z_B^{(2)}$ or $Z_B^{(3)}$	50
Figure 2.4	Credible interval length against sample size, grouped by regression functions marked in the subtitles. The blue cross sign marks the oracle confidence interval length under each setting.	58
Figure 2.5	Coverage against J , $f(x_1, x_2) = (x_1 + x_2)^2$	59
Figure 2.6	Coverage against J , $f(x_1, x_2) = e^{x_1 x_2}$	60
Figure 3.1	Function graphs of g_0	132
Figure 4.1	Density plots of the estimated α_0 for p -values from two-sided t-tests ($G = 50$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density. dash line	158
Figure 4.2	Density plots of the estimated α_0 for p -values from two-sided t-tests ($G = 100$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density. .	159
Figure 4.3	Density plots of the estimated α_0 for p -values from one-sided t-tests ($G = 50$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density. .	160
Figure 4.4	Density plots of the estimated α_0 for p -values from one-sided t-tests ($G = 100$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density. .	161

CHAPTER

1

INTRODUCTION

1.1 Statistical inference under shape restrictions

The research on statistical inference under shape constraints has an extensive history that can be traced back to the 1950s; see Ayer et al. (1955); Grenander (1956); Barlow et al. (1972), among others. Scientific problems often involve parameters of interest that are subject to inherent order restrictions, such as monotonicity, unimodality, and convexity. For instance, when evaluating the safety of potentially dangerous substances, the findings from bioassay experiments are frequently used. In these circumstances, it's often physically plausible to make the unimodality assumption on the derivative function of the dose-response curve. In economics, the utility function is to describe the preferences an individual or society has for certain goods or outcomes. The assumption of concavity in the utility function is fundamentally important. A

concave utility function reflects the decreasing marginal utility and influences the shape of demand curves. Researches on multivariate shape restrictions have recently gained significant interest and are increasingly recognized as crucial in a number of scientific studies. The application of an additive structure and shape restrictions on component functions is evidently insufficient. This insufficiency is evident in genetics, where the influence of a single genetic factor on a phenotype can be augmented, or even nullified by other ones. As a consequence, the isotonicity assumption on genetic effects is widely accepted. However, the genetic interaction between genetic factors remains unclear and may differ across various phenotypes, which poses some doubt about employing additive models. Besides the shape-restricted problems arising directly from real-world data as discussed earlier, another important source is the statistical inverse problem. Examples of this include, for example, the deconvolution problem and the current status model with right-censoring. In these situations, shape-restricted inference has shown its appropriateness.

In addition to many scenarios where shape assumptions are both natural and reasonable, shape-restricted inference has significant advantages over parametric or other nonparametric statistical inference methods. In contrast to parametric models such as linear models, inference based solely on shape restrictions reduces the risk caused by model misspecification. Nonparametric function estimation often requires selecting certain tuning parameters, like the bandwidth in kernel smoothing or the degree of polynomials and the positioning of knots in spline smoothing. Nevertheless, optimal choices partly rely on the local regularity of the unknown function being estimated, which can limit the applicability of these methods. However, shape-restricted inference does not necessitate a tuning process. Shape restrictions can effectively regulate the data observed with noise even in a localized manner, leading to a consistent inference method. Moreover, statistical models based on shape restrictions are not limited to dealing with continuous variables; they can also handle categorical or ordinal valued variables, extending beyond the estimation of smooth functions. These characteristics make shape-restricted inference a desirable technique in statisticians' toolbox.

Extensive research has been conducted on non-Bayesian approaches in univariate function estimation problems, primarily focusing on the least squares estimator and the nonparametric maximum likelihood estimator. Recently, there has been a surge of interest in the study of multivariate shape-restricted function estimation problems. The corresponding research in Bayesian nonparametrics has been somewhat limited, and we aim to close this gap in our work. By carrying out a thorough theoretical study of our method, we hope to demonstrate its potential in handling other complex problems by extending the traditional Bayesian paradigm. In the rest of this section, we will provide some concrete examples of shape-restricted problems and the background of the existing approaches.

1.1.1 Examples

In this section, we demonstrate the usefulness and ubiquity of shape-restricted problems through several examples. Some examples are derived from mathematical relations, while others are based on prior knowledge.

Current status data

Current status data arise when each participant in a study is observed only once, and the survival time of interest is known to be either less or greater than the time when observation happens. This specific data type is frequently encountered in cross-sectional studies, reliability theory, and tumorigenicity experiments. By imposing some reasonable assumptions, we can consider the following model. Let X_1, X_2, \dots, X_n be drawn from an unknown distribution function F , which is of primary interest. However, we can not observe X_i directly. For individual i , instead, we can observe a random time T_i and are able to know whether or not X_i is greater than T_i . Moreover, assume that X_i and T_i are independent of each other. Denoting the data by $(T_i, \Delta_i) = (T_i, \mathbb{1}_{X_i \leq T_i})$ and the realized version by (t_i, δ_i) , the log-likelihood function for F is given

by

$$\ell(F) = \sum_{i=1}^n [\delta_i \log F(t_i) + (1 - \delta_i) \log(1 - F(t_i))].$$

The nonparametric maximum likelihood estimator maximizes across the entire class of distribution functions. A crucial characteristic of a distribution function is its monotonically increasing nature. Finding the maximum likelihood estimate naturally leads to a shape-restricted optimization problem, where monotone shape-restricted inference can find its application; see Groeneboom and Wellner (1992); Groeneboom and Jongbloed (2014) for more detail. For simplicity, assume $t_1 < t_2 < \dots < t_n$. By Lemma 2.3 of Groeneboom and Jongbloed (2014), the nonparametric maximum likelihood estimator, $\hat{F}(t_i)$ is given by the left derivative of the convex minorant of the following cumulative sum diagram of points,

$$\left\{ (0, 0), (1, \sum_{j=1}^1 \delta_j), (2, \sum_{j=1}^2 \delta_j), \dots, (n, \sum_{j=1}^n \delta_j) \right\},$$

evaluated at i . That makes a nice connection with the theory of isotonic regression.

Hampel's bird migration problem

Hampel's bird migration problem represents a classic instance of nonparametric density estimation featuring a convexity constraint. This example is detailed in Section 4.3 of Groeneboom and Jongbloed (2014). Scientists are interested in studying the distribution of time during which a population of birds lingers at an oasis while traversing a desert. The durations that the birds spend at the oasis are commonly referred to as sojourn times. However, it is not feasible to directly observe these durations. Instead, multiple captures have been conducted repeatedly, where the captured birds were marked and then released. The capture times are recorded for each captured bird. We propose the following model to address our objective of assessing the distribution of sojourn times. Let X denote the sojourn time of a generic bird in the population, associated with a distribution function F . Additionally, we use a homogeneous

Poisson process to model the capture times denoted by N_t , where $t \geq 0$, with an intensity of λ . In Hampel's model, only the data where a bird is captured exactly twice are used. Given that the bird is captured exactly twice, the conditional distribution function of the sojourn time can be approximated as follows:

$$\tilde{F}(x) = P(X \leq x | N_X = 2) = \frac{\int_0^x P(N_X = 2 | X = y) dF(y)}{\int_0^\infty P(N_X = 2 | X = y) dF(y)} \approx \frac{\int_0^x y^2 dF(y)}{\int_0^\infty y^2 dF(y)}.$$

The last approximation is accurate when λ is small, in view of the fact that $P(N_X = 2 | X = y) = (\lambda y)^2 \exp\{-\lambda y\}/2 \approx (\lambda y)^2/2$. Furthermore, the distribution of the time lapse, denoted by Z , between two catches conditional on the $\{X = x, N_X = 2\}$ is the same as that of the distance of two uniformly and independently distributed random variables on $[0, x]$, which is the Beta(1, 2) distribution scaled by x . We can then write

$$P(Z > z | X = x, N_X = 2) = (1 - z/x) \text{ for } x \in [0, z].$$

Hence,

$$P(Z > z | N_X = 2) = \int P(Z > z | X = x, N_X = 2) d\tilde{F}(x) = \frac{\int_z^\infty (x - z)^2 dF(x)}{\int_0^\infty x^2 dF(x)}.$$

By differentiating about z , we get the density of Z , $g(z) \propto \int_z^\infty (x - z) dF(x)$ and, furthermore, we have $g'(z) \propto -(1 - F(z))$. It is now clear that g should be decreasing and convex. Based on the observed Z given $N_X = 2$, this naturally leads to a shape-restricted problem.

First-year grade point average data

We will now demonstrate an example of multivariate isotonic regression. The data in Table 1.1 is sourced from Section 1.3 of Robertson et al. (1988). The freshmen at the University of Iowa in the fall of 1978 are categorized into different groups based on their composite scores on the ACT Assessment (ACTC) and their high-school percentile rank (HSR). In each group, we

tabulate the average first-year grade point averages (GPA) in Table 1.1. The number of students belonging to each group is indicated below the average GPA in parentheses. In this scenario, students with higher high school rankings and higher ACTC scores tend to perform better in terms of GPA. However, it is difficult to determine whether students with higher high-school rankings but lower ACTC scores have a better GPA than those with lower high-school rankings and higher ACTC scores. Otherwise, for these two predictors, HSR and ACTC, if one predictor dominates the other on average, the admission procedure can simply rely on the dominant predictor. In order to predict the GPA, as a measure of the academic performance of freshmen, we need to seek the appropriate order on the predictors. As discussed earlier, the natural coordinate-wise partial order is a compelling choice, as it ensures that the intrinsic first-year GPA increases. The least squares estimator is obtained by minimizing the Euclidean distance between the observed noisy data and the order-restricted vector, which is obviously automated. The solution is well-defined except for those cells in Table 1.1 that lack observed data. The least squares estimation of the bivariate isotonic regression function is presented in Figure 1.1.

Baseball salary data

Here, we present another example where multivariate isotonic regression is a reasonable model. In this instance, our goal is to study the relationship between baseball salaries and players' performance. The well-known baseball salary dataset has been previously studied in He et al. (1998); Luss et al. (2012) for different purposes. The dataset comprises 1987 salaries of Major League Baseball players along with their 1986 and career performance statistics. The response variable in this study is the player's annual salary on opening day in 1987. We consider two predictors: the number of hits and the number of runs batted in during the 1986 season for each recorded player. The player's performance can be considered as a factor that influences their salary for the following year. Therefore, the use of multivariate isotonic regression is plausible. The raw data and the data fitted by the isotonic regression are plotted in Figure 1.2. The isotonic model efficiently denoises the data without relying on strict structural assumptions, thereby

Table 1.1: First-year GPA of 2397 freshmen in the University of Iowa in the fall of 1978.

ACTC	1-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
$91 \leq \text{HSR} \leq 99$	1.57 (4)	2.11 (5)	2.73 (18)	2.96 (39)	2.97 (126)	3.13 (219)	3.41 (232)	3.45 (47)	3.51 (4)
$81 \leq \text{HSR} \leq 90$	1.80 (6)	1.94 (15)	2.53 (30)	2.68 (65)	2.69 (117)	2.82 (143)	2.75 (70)	2.74 (8)	(0)
$71 \leq \text{HSR} \leq 80$	1.88 (10)	2.32 (13)	2.32 (51)	2.53 (83)	2.58 (115)	2.55 (107)	2.72 (24)	2.76 (4)	(0)
$61 \leq \text{HSR} \leq 70$	2.11 (6)	2.23 (32)	2.29 (59)	2.29 (84)	2.50 (75)	2.42 (44)	2.41 (19)	(0)	(0)
$51 \leq \text{HSR} \leq 60$	1.60 (11)	2.06 (16)	2.12 (49)	2.11 (63)	2.31 (57)	2.10 (40)	1.58 (4)	2.13 (1)	(0)
$41 \leq \text{HSR} \leq 50$	1.75 (6)	1.98 (12)	2.05 (31)	2.16 (42)	2.35 (34)	2.48 (21)	1.36 (4)	(0)	(0)
$31 \leq \text{HSR} \leq 40$	1.92 (7)	1.84 (6)	2.15 (5)	1.95 (27)	2.02 (13)	2.10 (13)	1.49 (2)	(0)	(0)
$21 \leq \text{HSR} \leq 30$	1.62 (1)	2.26 (2)	1.91 (5)	1.86 (14)	1.88 (11)	3.78 (1)	1.40 (2)	(0)	(0)
$\text{HSR} \leq 20$	1.38 (1)	1.57 (2)	2.49 (5)	2.01 (7)	2.07 (7)	(0)	0.75 (1)	(0)	(0)

uncovering a more informative nonlinear and complex relationship between the salary and performance statistics.

1.1.2 Least squares estimator and maximum likelihood estimator

Nonparametric inference often involves a regression function or a density function in modeling. Commonly, a smoothness assumption on a function of interest is imposed, but in some applications, qualitative information, such as monotonicity, unimodality, and convexity, on the shape of the function may be available. This leads to a control on the complexity of the function space analogous to what a smoothness assumption does, allowing convergence without requiring the latter. Monotonicity is the simplest and the most extensively studied shape restriction, especially in the univariate case. In regression analysis, this problem is commonly referred to as isotonic regression when the conditional mean function of the response variable is assumed to be nondecreasing. Starting from the early works on monotone shape-restricted problems,

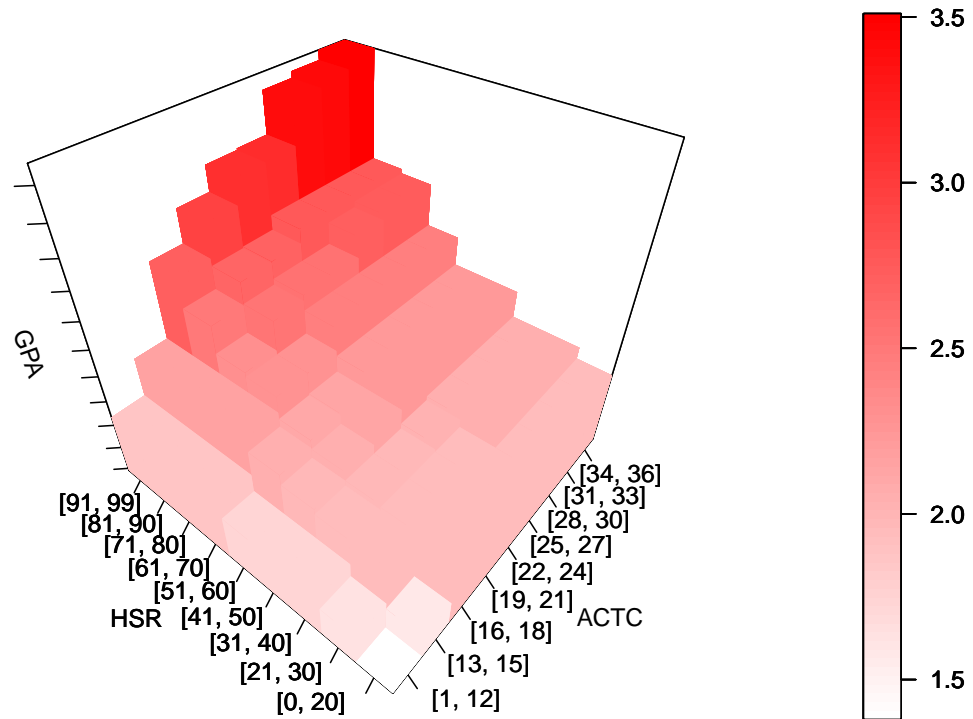


Figure 1.1: Bivariate isotonic regression function for the first-year GPA data

such as Ayer et al. (1955); Brunk (1955), research on non-Bayesian approaches, mainly on the least squares estimator (LSE) and the nonparametric maximum likelihood estimator (MLE), has been fruitful, see Grenander (1956); Barlow et al. (1972); Groeneboom (1985); Robertson et al. (1988). Assuming a non-zero derivative, the pointwise asymptotic distribution of the MLE or the LSE turns out to be the rescaled Chernoff distribution, that is, the minimizer of a quadratically drifted standard two-sided Brownian motion; see Prakasa Rao (1969); Brunk (1970); Wright (1981); Groeneboom (1985). The same limiting distribution can also be found in other problems where monotonicity is implied, such as the monotone hazard rate estimation with randomly right-censored observations in survival analysis; see Huang and Zhang (1994);

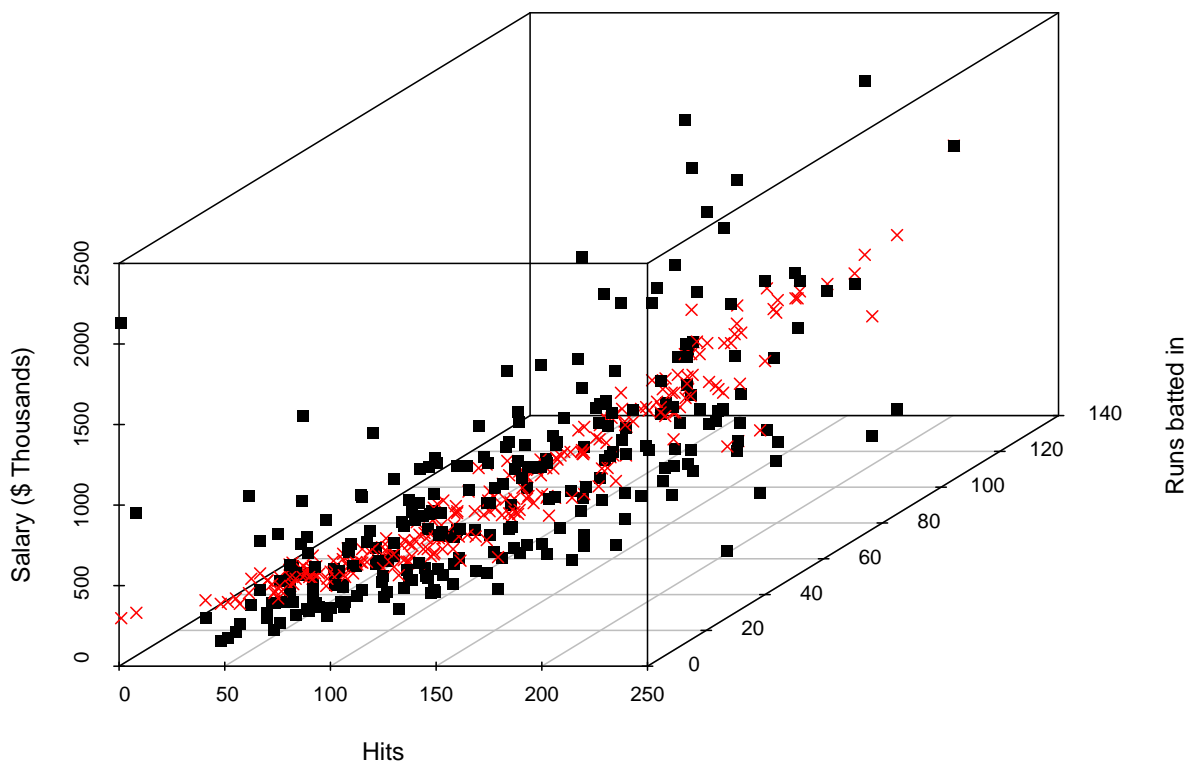


Figure 1.2: Baseball salary data: The black solid squares represent the raw data points, and the red cross signs indicate the fitted salary at every observation.

Huang and Wellner (1995), and many statistical inverse problems, including the current status model and deconvolution problems; see Groeneboom and Jongbloed (2014). Global properties of shape-restricted estimators were also studied extensively; see Groeneboom (1985); Kulikov and Lopuhaä (2005). The convergence rates and limiting distributional behaviors of \mathbb{L}_p - and \mathbb{L}_∞ -distance between monotone shape-restricted estimators and the true function were investigated by Durot (2007); Durot et al. (2012). Nonasymptotic risk bounds for the LSE under a monotone shape restriction were derived by Zhang (2002); Chatterjee et al. (2015); Bellec (2018). Testing for monotonicity was addressed in Hall and Heckman (2000); Gijbels et al. (2000); Ghosal et al. (2000b); Dümbgen and Spokoiny (2001).

Multivariate monotone function estimation was also studied in the literature. Non-Bayesian works focused on the construction of the LSE with respect to various partial orderings on the domain; see Barlow et al. (1972); Robertson et al. (1988). Only the consistency of the isotonic estimator was known until a recent rise in interest in multivariate shape-restricted problems. In a multivariate isotonic regression model, the \mathbb{L}_2 -risk of the LSE, respectively for $d = 2$ and for a general dimension d was studied by Chatterjee et al. (2018); Han et al. (2019). They found that the LSE achieved the optimal minimax rate up to logarithmic factors, and adapted to the parametric rate for a piecewise constant true regression function only when $d \leq 2$. Han (2021) confirmed that the global empirical risk minimizer is indeed rate-optimal in some set structured models even with rapidly diverging entropy integral and thus gave a simpler proof for the optimal convergence rate of the LSE in the multivariate isotonic regression. Deng and Zhang (2020) investigated a block-estimator proposed by Fokianos et al. (2020) and obtained an \mathbb{L}_q -risk bound. This is minimax rate optimal and adapts to the parametric rate up to a logarithmic factor when the true regression function is piecewise constant. Pointwise distributional limits for the block-estimator were obtained by Han and Zhang (2020), which lays the foundation for subsequent inference.

In Han and Zhang (2020), a recently proposed estimator (see Fokianos et al. (2020)) has been thoroughly studied from a local perspective. Although the asymptotic distributional theory is unclear for the LSE in multivariate isotonic regression, the block estimator does stabilize at a limiting distribution. This distribution is characterized by a sup-inf functional of a centered Gaussian process residing in a higher-dimensional Euclidean space, along with a drift term. The reason why the distributional theory is possible is partly due to that

1.1.3 Bayesian shape-restricted inference

The Bayesian approach to shape-restricted problems was also explored, albeit to a lesser extent. Neelon and Dunson (2004) applied piecewise linear structures to the regression function, and put monotone restrictions on the priors for the slope values. Cai and Dunson (2007) proposed

a linear spline model and added an initial Markov random field prior to the coefficients, and then the monotone constraint was incorporated by considering the relation between the nonnegative slopes and coefficients. Wang (2008) adopted the free-knot cubic regression spline model, converted the shape restriction to the coefficients, and then projected the unconstrained coefficients with conventional priors to the target set, inducing the constrained priors. Shively et al. (2009) also used Bayesian splines with constrained normal priors on the coefficients to comply with the monotone shape restriction. Lin and Dunson (2014) addressed this problem by using a Gaussian process prior and projected unconstrained posterior samples to the monotone function class by a min-max formula. Chakraborty and Ghosal (2021c,b, 2022) also used the idea of projection-posterior, making the investigation of frequentist limiting coverage of credible sets possible. Salomond (2014b) used a mixture representation of a nonincreasing density on $[0, \infty)$ and obtained the nearly minimax posterior contraction rate for both a Dirichlet Process and a finite mixture prior on the mixing distribution. Bayesian tests for monotonicity were developed by Salomond (2014a); Chakraborty and Ghosal (2021b, 2022).

The Bayesian approach to multivariate isotonic regression is much less developed. Saarela and Arjas (2011) proposed a Bayesian approach to this problem based on marked point processes and resulted in piecewise constant realizations of the regression function. Lin and Dunson (2014) mentioned that the method of projecting the Gaussian process posterior can also be applied in regression surface case. Nonetheless, the theoretical studies presented in those works are either lacking or inadequate.

1.2 Bayesian nonparametrics

Bayesian nonparametric models have gained significant traction and have been applied to a wide range of modern problems. Their versatility and adaptability have allowed them to address various challenges across different domains. Bayesian nonparametric regression models have found extensive application in regression problems. These models are capable of

capturing complex relationships between input variables and output targets, enabling accurate predictions and improved understanding of underlying patterns in the data. They have been employed in diverse fields such as economics, finance, medicine, genetics, ecology and social sciences to analyze and predict outcomes. Moreover, those important aspects of machine learning, like classification and clustering, have also benefited from Bayesian nonparametrics by leveraging the flexibility of these models, such as latent variable modeling. Bayesian nonparametric models have been employed in areas such as image recognition, natural language processing, fraud detection, document clustering, image segmentation, and recommending systems, among others. Beyond machine learning problems, Bayesian nonparametric models have found application in specialized domains, such as spatial-temporal problems. These models excel at capturing spatial and temporal correlations, making them well-suited for tasks such as weather forecasting, and environmental modeling. We summarize some Bayesian nonparametric methods in the following sections and provide a brief exposition of the frequentist asymptotic properties of posterior measures. It is worth noting that these properties pose greater challenges in nonparametric settings, and certain aspects of them are still open questions to this day.

1.2.1 Nonparametric modeling methods

Nonparametric Bayesian methods are very useful and flexible with broad applications in various fields. These methods provide a powerful framework for modeling complex data structures. In this section, we present an overview of several popular nonparametric Bayesian methods including Dirichlet processes, Gaussian processes, hierarchical models, infinite mixture models, and Bayesian nonparametric regression.

Dirichlet Process, introduced by Ferguson (1973), is a popular nonparametric prior for probability measures in Bayesian nonparametric models for modeling unknown distributions. Dirichlet process mixtures with various kernels have been explored for Bayesian density estimation; see Ferguson (1983); Lo (1984); Brunner and Lo (1989); Escobar and West (1995);

Petrone (1999), among others. Markov chain Monte Carlo computational techniques have been invented to compute posterior characteristics; see Escobar and West (1995); MacEachern and Müller (1998); Walker (2007) and others. The hierarchical Dirichlet process (see Teh et al. (2004)) extends the Dirichlet Process by allowing for the hierarchical modeling of multiple groups or clusters. It provides a way to model both within-group and between-group variations. The Pitman-Yor Process is an extension of the Dirichlet Process that allows for more flexible modeling of power-law distributions. It can capture heavy-tailed behavior and long-range dependencies in the data and has been widely studied and applied in various fields, particularly in natural language processing and computational linguistics; see Teh (2006). Infinite hidden Markov models are nonparametric extensions of traditional Hidden Markov Models; see Beal et al. (2001). They allow for an infinite number of hidden states, which can be useful for modeling sequences with unknown or varying lengths.

Gaussian processes are widely used in Bayesian nonparametric regression and classification. They provide a flexible way to model functions and can capture complex dependence in the data; see, for example, Bernardo et al. (1998). The practical implementation of Gaussian processes involves choosing a suitable covariance kernel function, which determines the smoothness, periodicity, and other properties of the modeled functions. Some commonly used Gaussian process priors include Brownian motion, Riemann-Liouville process, fractional Brownian motion, Matérn process, etc; see Wahba (1978); Leonard (1978); Yang and Tokdar (2015), among others.

The Beta process prior is a commonly used nonparametric prior, as discussed in Hjort (1990). It is often used as the prior distribution for the cumulative hazard rate process in survival analysis and other related fields. Other random structures, such as the Chinese restaurant process, species sampling process, Gibbs process, stick-breaking process, nested Dirichlet process, and Indian buffet process, enhance the flexibility and find applications in various scenarios.

1.2.2 Frequentist asymptotic properties of posterior measures

We will review several concepts and theoretical results relevant to certain sections of the thesis. We start by introducing some definitions concerning the frequentist consistency of posterior measures. For an observation $X^{(n)}$ from a probability distribution $P_\theta^{(n)}$ where θ is the parameter in a topological space Θ .

Definition 1.2.1. *The posterior measures $\Pi(\cdot|X^{(n)})$ are consistent at $\theta_0 \in \Theta$ if $\Pi(U|X^{(n)}) \rightarrow 1$ in $P_{\theta_0}^{(n)}$ -probability for every neighborhood U of θ_0 .*

If the topology space Θ is metrizable, U can be restricted to balls centered at θ_0 . If the convergence can be in $P_{\theta_0}^{(n)}$ -almost surely sense, we say that the posterior measures are strongly consistent at θ_0 . The early posterior consistency result is due to Doob's theorem, which states, in a remarkable manner, that under moderate measurability conditions on the model as well as for separable and complete metric sampling and parameter spaces, and for any Borel prior, the posterior measure is strongly consistent, except for a null set regarding the prior measure. However, from a frequentist perspective, when considering the data from a true model P_{θ_0} and focusing on the convergence to the unknown fixed single point θ_0 , Doob's theorem does not provide a practical way to check. The examples of inconsistency in Bayesian nonparametrics can be found in Cox (1993); Diaconis and Freedman (1998).

Schwartz's theorem and its extensions provide sufficient conditions for posterior consistency with respect to the true model, p_0 .

Theorem 1.2.1 (Schwartz). *$X^{(n)}$ is an i.i.d. sample from a distribution P_0 with probability density function p_0 relative to a dominated class \mathcal{D} . \mathcal{U} is an open neighborhood of p_0 in \mathcal{D} . If there exists a sequence of tests ϕ_n such that*

$$P_0^{(n)} \phi_n \rightarrow 0, \text{ and } \sup_{p \in \mathcal{U}^c} P^{(n)}(1 - \phi_n) \rightarrow 0,$$

and

$$\Pi(p \in \mathcal{D} : \int p_0 \log(p_0/p) < \epsilon) > 0, \text{ for every } \epsilon > 0,$$

then the posterior measures are strongly consistent at p_0 .

A test is defined as a measurable function that maps observations to values between 0 and 1. The existence of uniformly consistent tests, as the first condition in Schwartz's theorem, essentially amounts to the existence of uniformly consistent tests with an exponentially decaying rate with the sample size n . In other words, the error rate of the tests is no slower than $\exp\{-Cn\}$, for some constant $C > 0$. The second condition is commonly referred to as the Kullback-Leibler property. We will provide a brief overview to help understand this theorem and show how these two conditions are applied.

Let \mathcal{V} denote the residual part in the parameter space, concretely defined as the complement of \mathcal{U} , \mathcal{U}^c . Using Bayes formula, the posterior probability of \mathcal{V} is given by

$$\Pi(\mathcal{V}|X^{(n)}) = \frac{\int_{p \in \mathcal{V}} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)}{\int_{p \in \mathcal{D}} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)} \quad (1.1)$$

Let \mathbb{P}_n denote the empirical measure. The integrand in the denominator in the last display can be written as $\exp\{-n\mathbb{P}_n \log p_0/p\}$, which converges to its population counterpart $P_0^{(n)}$ -almost surely. Restricted in the Kullback-Leibler neighborhood of p_0 and in view of the Kullback-Leibler property, the denominator in (1.1) can be lower bounded by a constant multiple of $\exp\{-cn\}$ for some constant $c > 0$ almost surely.

A useful byproduct of this observation, which will be used in a latter argument, is that if the prior probability of the parameter set is bounded by $\exp\{-c'n\}$ for some $c' > c$, then the posterior probability of \mathcal{V} converges to zero almost surely. To see this point, apply Fubini's theorem to the following expectation,

$$P_0^{(n)} \int_{p \in \mathcal{V}} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p) = \Pi(\mathcal{V}) \leq \exp\{-c'n\}. \quad (1.2)$$

By Borel–Cantelli lemma, the posterior probability of \mathcal{V} converges to zero almost surely as $\sum_{n=1}^{\infty} \exp\{-(c' - c)n\} < \infty$.

By taking \mathcal{V} as \mathcal{U}^c and applying the first condition of uniform consistency, with a similar argument as in the last paragraph, Schwartz’s theorem follows.

A more detailed frequentist property than consistency is the convergence rate. Suppose the parameter space (Θ, d) is a metric space. The posterior contraction rate is defined as follows.

Definition 1.2.2. *Let $X^{(n)}$ be an i.i.d. sample from distribution P_{θ_0} , for $\theta_0 \in \Theta$. Let the positive real sequence $\epsilon_n \rightarrow 0$. We say that the posterior measures contracts to θ_0 at rate ϵ_n if*

$$\Pi(\{\theta : d(\theta, \theta_0) \geq M_n \epsilon_n\} | X^{(n)}) \rightarrow 0 \text{ in } P_0^{(n)}\text{-probability,}$$

for any sequences $M_n \rightarrow \infty$.

The main theorem dealing with the posterior contraction rates is due to Ghosal et al. (2000a). For a parameter space (\mathcal{P}, d) , we denote the Kullback-Leibler ball around p_0 by

$$B_2(p_0, \epsilon) = \{p \in \mathcal{P} : \int p_0 \log p_0/p \leq \epsilon^2, \int p_0 (\log p_0/p)^2 \leq \epsilon^2\},$$

where \mathcal{P} is assumed to be a class of probability densities relative to dominated measures for simplicity. We state a version of the theorem concerning posterior contraction rates in the following; see Ghosal et al. (2000a). In the next theorem, we assume d is the Hellinger distance for ease of presentation. Let $N(\epsilon, \mathcal{P}, d)$ denote the ϵ -covering number.

Theorem 1.2.2. *Let $X^{(n)}$ denote an i.i.d. observation from a true distribution P_0 with density p_0 . Let Π denote the prior on \mathcal{P} . \mathcal{P}_n is a sieve of \mathcal{P} . Let ϵ_n and $\bar{\epsilon}_n$ be two real sequences such that $\epsilon_n \leq \bar{\epsilon}_n$ and $n\bar{\epsilon}_n^2 \rightarrow \infty$. There exist positive constants C_1, C_2, C_3 , and C_4 such that*

1. $\Pi(B_2(p_0, \bar{\epsilon}_n)) \geq C_2 \exp\{-C_1 n \bar{\epsilon}_n^2\}$;
2. $\log N(\epsilon_n, \mathcal{P}_n, d) \leq C_3 n \epsilon_n^2$;

$$3. \Pi(\mathcal{P} \setminus \mathcal{P}_n) \leq \exp\{-(C_1 + 4)n\bar{\epsilon}_n^2\}.$$

Then the posterior contraction rates at p_0 is ϵ_n .

In contrast to Schwartz's theorem, Theorem 1.2.2 replaces the uniformly consistent test condition with a relatively easy-to-verify entropy condition. To appreciate this change, we have to mention the minimax Hellinger test theorem between two convex sets that are separated from each other in terms of Hellinger distance; see Le Cam (1986). More specifically, for two subsets \mathcal{P}_1 and \mathcal{P}_2 of \mathcal{P} , there exists a test $\bar{\phi}_n$ such that the type I error over \mathcal{P}_1 and the type II error over \mathcal{P}_2 are both uniformly bounded by $\exp\{-nh(\mathcal{P}_1, \mathcal{P}_2)^2/2\}$. However, this result relies on the convexity of both subsets and in the current context, the composite alternative set, denoted by $\mathcal{U}_n^c = \{p \in \mathcal{P} : d(p, p_0) \geq M\epsilon_n\}$ for some $M > 0$, is not convex. To overcome this issue, we cover the alternative set \mathcal{U}_n^c with Hellinger balls of radius ϵ_n , as described in Le Cam (1973), and then combine the optimal tests against each ball into a single test, which can be easily shown to have the error rate bounded by the ϵ_n -covering number multiplied by the individual error rate. Thus, the existence of the minimax Hellinger test and the use of convex ball covering provide the desired test condition. The residual prior mass allows for more flexibility in applications.

1.3 Immersion posterior-based Bayesian inference

Consider a general statistical model with observation $X \sim P_\theta$, where $\theta \in \Theta_0$. Suppose that the parameter space Θ_0 is a complicated subset of a larger, but simpler to represent, set Θ . This is often the case for shape-restricted inference, where structural constraints, such as monotonicity, convexity, and log-concavity, are imposed on a regression function or a density function. In differential equation models, the parameter space is implicitly described as the set of solutions of a system of ordinary or partial differential equations involving some unknown parameters. In a vector autoregressive process, the set of autoregression coefficients leading to stationary processes may be the parameter space of interest, but it is described by many

complicated constraints. Because of the complicated restrictions on Θ_0 , a prior for θ with support on Θ_0 may be hard to construct, and the corresponding posterior may be difficult to compute. More importantly, the corresponding posterior may be hard to analyze from a frequentist perspective. This may be particularly important for studying delicate properties such as the limiting coverage of a Bayesian credible region.

Often, the distribution P_θ makes sense for any $\theta \in \Theta$, so that Θ_0 can be embedded in Θ , keeping the statistical problem meaningful. For shape-restricted models, this becomes the standard nonparametric regression or the density estimation problem. A differential equation model also embeds in a nonparametric regression model. A prior distribution Π may be specified on Θ , initially disregarding the restriction of θ to Θ_0 . This is typically a standard problem, and often a conjugate prior distribution can be identified. The resulting posterior distribution $\Pi(\cdot|X)$ thus resides in the whole of Θ , and hence is not appropriate to make an inference about θ , which is known to live in Θ_0 . The requirement can be met by considering the random measure induced by a mapping ι from Θ to Θ_0 , in that, we consider the random measure $\Pi^*(B|X) = \Pi(\iota(\theta) \in B|X)$ to make an inference on θ . The map ι immerses θ into the desirable space Θ_0 , and hence will be referred to as the *immersion map*. The induced posterior Π^* will be referred to as the *immersion posterior*. This provides an extension of the Bayesian paradigm since the identity map as the immersion map for the situation $\Theta_0 = \Theta$ reduces the immersion posterior to the classical Bayesian posterior.

The approach has been successfully used in several works including Lin and Dunson (2014); Chakraborty and Ghosal (2021c,b, 2022, 2021a) for shape-restricted problems, and by Bhaumik and Ghosal (2015, 2017b,a); Bhaumik et al. (2022) for differential equation models. These authors used a projection map \mathfrak{p} obtained by minimizing a certain distance from the posterior sample to the restricted space, and the resulting induced random measure is called the projection posterior distribution. The projection map \mathfrak{p} satisfies the appealing property $\mathfrak{p}(\theta) = \theta$ for all $\theta \in \Theta_0$.

While a projection map with respect to an appropriate distance is a natural choice for an

immersion map, the restriction to a projection map is unnecessary for the concept to be used. Depending on the aspect to be studied, there may not be a natural distance associated with it. This happens, for instance, if we are interested in studying the posterior distribution of the function value at a given point. It is also not necessary for the immersion map ι to satisfy $\iota(\theta) = \theta$ for all $\theta \in \Theta_0$. Neither Θ_0 needs to be a subset of Θ , nor the immersion map needs to be defined all over Θ . All that is needed is that an alternative parameter space Θ exists where the model distribution P_θ makes sense, a prior Π can be put on Θ such that the posterior distribution can be computed relatively easily, and the random distribution induced by a map ι from the support of the posterior distribution $\Pi(\cdot|X)$ to Θ_0 can be analyzed theoretically to establish some desirable properties. In most situations, the family of measures $\{P_\theta : \theta \in \Theta\}$ is dominated, so the support of the posterior distribution $\Pi(\cdot|X)$ is contained in the support of the prior distribution Π . The immersion map may be allowed to depend on the sample size like a prior distribution may be allowed to depend on the sample size. Even dependence of ι on the data X may be allowed. Although there is no uniqueness in the choice of the immersion map, the main purpose is to increase flexibility in the posterior measure to achieve a targeted asymptotic frequentist property, such as coverage of a credible region. A choice of an immersion map is therefore guided by a desirable frequentist property. Even if Θ_0 and Θ coincide, the flexibility of the immersion posterior may be helpful to satisfy a desirable convergence property of the immersion posterior that the classical Bayesian posterior may lack.

In many applications, the prior distribution may be actually a sequence of prior distributions specified through a sieve indexed by a discrete variable $J = J_n$ depending on the sample size n . Let Θ_J stand for the sieve (typically a finite-dimensional subset of Θ) and Π_J stand for the prior at that stage concentrated on Θ_J . Then the computation of the posterior in the unrestricted space reduces to a finite-dimensional computation, often also aided by posterior conjugacy. It is then typical that the immersion map ι on Θ_J has the range in $\Theta_0 \cap \Theta_J$ so that the computation of the immersion posterior involves finite-dimensional computations only. Most examples from the existing literature, as well as the method used in this thesis, fall in this

setting.

1.4 Notation

We summarize the notations we shall use in the thesis. The notations \mathbb{R} , \mathbb{N} , and \mathbb{Z} will stand for the real line, the set of natural numbers, and the set of all integers respectively. The positive half-line with and without 0, and the set of nonnegative integers are respectively denoted by $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{> 0}$, and $\mathbb{Z}_{\geq 0}$. Bold Latin or Greek letters will be used to indicate column vectors and the non-bold letter with a subscript will denote a coordinate of the corresponding vector. For example, a_i is the i th coordinate of $\mathbf{a} \in \mathbb{R}^d$. Let $\mathbf{1}$ denote the d -dimensional all-one column vector and $\mathbf{0}$ the all-zero column vector. Let \mathbf{A}^\top denote the transpose of a matrix or a vector \mathbf{A} . For an arbitrary set A , the indicator function will be denoted by $\mathbb{1}_A(\cdot)$, and $\#A$ will denote the cardinality of a finite set A . Let $\lceil a \rceil$ stand for the smallest integer greater than or equal to a real number a . Let $f_+ = \max(f, 0)$ denote the positive part of the function f , and $f(x-)$ (respectively, $f(x+)$) denote the left (respectively, right) limit of f at x when it exists.

For two positive real sequences a_n and b_n , we use $a_n = O(b_n)$ and $a_n = o(b_n)$ when a_n/b_n is bounded and $a_n/b_n \rightarrow 0$ and $O_p(\cdot)$, $o_p(\cdot)$ stand for their stochastic versions. The symbols, \lesssim and \gtrsim , represent less/greater than or equal to up to a universal positive constant. The symbol, \asymp means being equal in order. We use $a_n \ll b_n$ if $a_n/b_n \rightarrow 0$. Say $a_n \sim b_n$ when $a_n/b_n \rightarrow 1$.

For $a, b \in \mathbb{R}$, let $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, let $\mathbf{a} \wedge \mathbf{b} = (a_1 \wedge b_1, \dots, a_d \wedge b_d)^\top$, $\mathbf{a} \vee \mathbf{b} = (a_1 \vee b_1, \dots, a_d \vee b_d)^\top$, and the pointwise product $\mathbf{a} \circ \mathbf{b} = (a_1 b_1, \dots, a_d b_d)^\top$. For a vector $\mathbf{a} \in \mathbb{R}^d$, the Euclidean and the maximum norms are respectively denoted by $\|\mathbf{a}\|$ and $\|\mathbf{a}\|_\infty = \max\{|a_k| : 1 \leq k \leq d\}$. Let $[\mathbf{j}_1 : \mathbf{j}_2] = \{\mathbf{j} \in \mathbb{Z}^d : j_{1,k} \leq j_k \leq j_{2,k}, \text{ for all } 1 \leq k \leq d\}$ stand for the lattice with boundaries $\mathbf{j}_1, \mathbf{j}_2 \in \mathbb{Z}^d$.

For a multivariate function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, let $\partial_k^l f(\mathbf{x}) = \partial^l f(\mathbf{x}) / \partial x_k^l$ for $k \in \{1, \dots, d\}$ and $l \in \mathbb{Z}_{\geq 0}$ at a suitable point $\mathbf{x} \in \mathbb{R}^d$. For a multiple index $\mathbf{l} = (l_1, \dots, l_d)^\top \in \mathbb{Z}_{\geq 0}^d$, we use $\partial^{\mathbf{l}} = \partial_1^{l_1} \dots \partial_d^{l_d}$, $\mathbf{l}! = l_1! \dots l_d!$, and $\mathbf{x}^{\mathbf{l}} = x_1^{l_1} \dots x_d^{l_d}$.

For $p > 0$, let $\mathbb{L}_p(\mu)$ denote the set of real-valued functions defined on $[0, 1]^d$ with respect to a measure μ whose p th power is integrable. For $p \geq 1$ and $f \in \mathbb{L}_p(\mu)$, the \mathbb{L}_p -norm of f is denoted by $\|f\|_{p,\mu}$. For a distance ρ on functions, a function f and a set of functions \mathcal{F} , let $\rho(f, \mathcal{F}) = \inf\{\rho(f, g) : g \in \mathcal{F}\}$. For ease of notation, let $\mathbb{L}_p(A)$, $p \geq 1$, and $\mathbb{L}_\infty(A)$ be the spaces of Lebesgue p -integrable and bounded functions on A respectively. Let $\mathbb{L}_p[\mathbf{a}, \mathbf{b}]$, $1 \leq p \leq \infty$, stand for the Lebesgue \mathbb{L}_p -space on a multivariate interval $[\mathbf{a}, \mathbf{b}]$. For $1 \leq p \leq \infty$, let $\|\cdot\|_p$ be the p -norm and define $d_p(f, \mathcal{S}) = \inf\{\|f - s\|_p : s \in \mathcal{S}\}$, for $f \in \mathbb{L}_p(A)$ and $\mathcal{S} \subseteq \mathbb{L}_p(A)$. The Hellinger distance between two densities p_1 and p_2 is defined by $d_H(p_1, p_2) = \|\sqrt{p_1} - \sqrt{p_2}\|_2$. The Kullback-Leibler divergence and Kullback-Leibler variation are respectively given by $K(p_1, p_2) = \int p_1 \log(p_1/p_2)$ and $V(p_1, p_2) = \int p_1 [\log(p_1/p_2)]^2$. For a semimetric space (\mathcal{T}, d) , the metric entropy refers to the logarithm of the covering number $\mathcal{N}(\epsilon, \mathcal{T}, d)$, while the bracketing entropy refers to the logarithm of the bracketing number $\mathcal{N}_{[]}(\epsilon, \mathcal{T}, d)$; see Section 2.1 of van der Vaart and Wellner (1996) for details.

Let $N(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2 . The binomial distribution with parameters n and p is denoted by $\text{Bin}(n, p)$. Let Δ_J be the unit J -simplex, $J = 1, 2, \dots$. The k -dimensional Dirichlet distribution with parameters $(\alpha_1, \dots, \alpha_k)$ is denoted by $\text{Dir}(k; \alpha_1, \dots, \alpha_k)$. We use $\text{Ga}(a, b)$ denoting Gamma distribution with shape parameter a and scale parameter b . Let $\mathcal{L}(\cdot)$ denote the law of a random element. Equality in distribution is denoted by $=_d$. Convergence in probability under a measure P is denoted by \rightarrow_p . Distributional equality will be denoted by $=_d$ and weak convergence by \rightsquigarrow .

1.5 Chapter organization

The rest of the thesis is organized as follows. In Chapter 2, we study the nonparametric multivariate isotonic regression problem by using an immersion posterior scheme. We obtain the posterior contraction rates, construct a universally constant Bayesian testing procedure for the multivariate monotonicity in the regression problem, and study the frequentist coverage

of the credible intervals of the regression function value at an interior point theoretically and numerically. In Chapter 3, our focus is on the estimation of multivariate decreasing density. We achieve this by employing Dirichlet priors on a finite partition of the support, followed by a monotonization operation. We demonstrate that the posterior converges to the true density at an optimal rate by carefully selecting the partition. Additionally, we have developed a consistent test procedure for multivariate monotonicity. We demonstrate its uniform consistency against a smooth density class, with an explicit separation rate from the null hypothesis. The asymptotic frequentist coverage of the pointwise credible interval has also been studied. It has been confirmed that the limiting coverage is slightly larger than the credibility based on the posterior quantile-based credible interval. In Chapter 4, we focus on the k -monotone density estimation problem. We introduce the notion of a k -monotone density, characterize it through a mixture representation, and present an important approximation result using finite mixtures. By the integral representation, we employ the Dirichlet process prior and finite mixture prior on the mixing distribution. We present results on the posterior contraction rates and consider Bayesian estimation with an unknown parameter, k . We also discuss the important application to multiple testing and demonstrate, through simulations under various settings, the usefulness of k -monotone density modeling.

CHAPTER

2

MULTIVARIATE ISOTONIC REGRESSION: CONTRACTION RATES, TESTS, AND FREQUENTIST COVERAGE

2.1 Models, priors, and posteriors

We adopt the coordinatewise partial ordering on \mathbb{R}^d , that is, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{x} \preceq \boldsymbol{y}$ if and only if $x_k \leq y_k$ for all $1 \leq k \leq d$. We also use $\boldsymbol{x}_2 \succeq \boldsymbol{x}_1$ if $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$. We say that a function f on \mathbb{R}^d is multivariate monotone if $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ for all $\boldsymbol{x} \preceq \boldsymbol{y}$. The class of all multivariate monotone functions on $[0, 1]^d$ will be denoted by \mathcal{M} .

We consider the nonparametric multivariate regression model

$$Y = f(\mathbf{X}) + \varepsilon, \quad (2.1)$$

where \mathbf{X} is the d -dimensional predictor and ε is an error term with zero mean and finite variance, independent of \mathbf{X} when \mathbf{X} is assumed to be random. We shall assume, essentially without loss of generality, that the domain of \mathbf{X} is $[0, 1]^d$. Instead of a traditional smoothness assumption on the regression function f , we only assume that f is multivariate monotonic. However, when we study the frequentist coverage of pointwise credible intervals, we would like to make assumptions on the local regularity of the regression function. To construct the likelihood function, we use the working Gaussian model by assuming that ε is normally distributed, but the actual data-generating process need not be so.

We observe the data \mathbb{D}_n consisting of n samples $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ independently from the model. The predictor variable \mathbf{X} may be deterministic or obtained independently from a fixed distribution G , independent of the random error variable ε . To make an inference on f , we adopt a Bayesian approach by putting an appropriate prior distribution on f and other parameters of the model.

We put a prior distribution on f through a sieve of piecewise constant functions with gradually refining intervals of constancy, forming a partition of $[0, 1]^d$. For $\mathbf{J} = (J_1, \dots, J_d)^\top \in \mathbb{N}^d$, let $I_j = \prod_{k=1}^d ((j_k - 1)/J_k, j_k/J_k]$ be a hyperrectangle in $[0, 1]^d$, indexed by a d -dimensional vector j , for $j \in [1 : \mathbf{J}] \setminus \{\mathbf{1}\}$ and $I_1 = \prod_{k=1}^d [0, 1/J_k]$. Then $\{I_j\}_{j \in [1 : \mathbf{J}]}$ forms a partition of $[0, 1]^d$. We define the piecewise constant functions

$$f = \sum_{j \in [1 : \mathbf{J}]} \theta_j \mathbb{1}_{I_j},$$

where $\theta_j \in \mathbb{R}$ for all j . As we follow the immersion posterior approach, we do not initially impose the order restriction. A prior is imposed on $f = \sum_{j \in [1 : \mathbf{J}]} \theta_j \mathbb{1}_{I_j}$ by giving independent

Gaussian priors to θ_j , namely,

$$\theta_j \sim \text{N}(\zeta_j, \sigma^2 \lambda_j^2), \quad \text{independently for all } j \in [1 : J], \quad (2.2)$$

where ζ_j and λ_j are hyperparameters. The values of the prior parameters, ζ_j and λ_j , will not affect our asymptotic results. However, in practice, when very little prior information is available, it is sensible to choose $\zeta_j = 0$ and λ_j large for all j .

To facilitate Bayesian inference, we construct a likelihood based on the working model assumption that $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \text{N}(0, \sigma^2)$, although the actual distribution may be non-normal. For a given f , let

$$p_{f,\sigma}(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y - f(\mathbf{x}))^2}{2\sigma^2}\right]$$

stand for the conditional density of Y given $\mathbf{X} = \mathbf{x}$. By the representation $f = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j}$, equivalently, we have

$$Y_i \sim \text{N}\left(\sum_{j \in [1:J]} \theta_j \mathbb{1}\{\mathbf{X}_i \in I_j\}, \sigma^2\right), \quad (2.3)$$

which leads to, in the unrestricted parameter space, a Gaussian joint likelihood for $(\theta_j : j \in [1 : J])$ without any cross-product terms in the exponent. This gives independent Gaussian posterior distribution for each θ_j , given σ , such that by conjugacy,

$$\theta_j | \mathbb{D}_n, \sigma \sim \text{N}((N_j \bar{Y}|_{I_j} + \zeta_j \lambda_j^{-2}) / (N_j + \lambda_j^{-2}), \sigma^2 / (N_j + \lambda_j^{-2})), \quad (2.4)$$

where $N_j = \#\{i : \mathbf{X}_i \in I_j\}$ and $\bar{Y}|_{I_j} = \sum_{i=1}^n Y_i \mathbb{1}\{\mathbf{X}_i \in I_j\} / N_j$.

The parameter σ^2 can be estimated by maximizing the marginal likelihood function given by

$$(2\pi\sigma^2)^{-n/2} \prod_{j \in [1:J]} (1 + \lambda_j^2 N_j)^{-1/2} \exp\left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (Y_i - \sum_{j: \mathbf{X}_i \in I_j} \zeta_j)^2 - \sum_{j \in [1:J]} \frac{N_j^2 (\bar{Y}|_{I_j} - \zeta_j)^2}{N_j + \lambda_j^{-2}} \right\}\right],$$

and the resulting estimator

$$\hat{\sigma}_n^2 = \frac{1}{n} \left[\sum_{i=1}^n \left(Y_i - \sum_{j: X_i \in I_j} \zeta_j \right)^2 - \sum_{j \in [1:J]} \frac{N_j^2 (\bar{Y}|_{I_j} - \zeta_j)^2}{N_j + \lambda_j^{-2}} \right], \quad (2.5)$$

may be plugged into the expression (2.4). Alternatively, in a fully Bayesian framework, we can give σ^2 an Inverse-Gamma prior $\text{IG}(b_1, b_2)$ with parameters $b_1 > 0$, $b_2 > 0$, and obtain that the posterior distribution of σ^2 is given by

$$\sigma^2 | \mathbb{D}_n \sim \text{IG}(b_1 + n/2, b_2 + n\hat{\sigma}_n^2/2).$$

It will be shown in Lemma A.1.3 that the marginal maximum likelihood estimator of σ^2 as well as the posterior for σ^2 concentrate in a shrinking neighborhood of its true value σ_0^2 . Then it easily follows that the asymptotic behavior of the posterior distribution of f is identical with that when σ is known to be σ_0 . Hence, it suffices to study the asymptotic behavior of the posterior distribution given σ .

We introduce some additional notations. Let $G_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ denote the empirical distribution of \mathbf{X} . For a deterministic predictor variable \mathbf{X} , this is a sequence of deterministic distributions, while for a random \mathbf{X} , this sequence is random. Let f_0 stand for the true value of the regression function f , σ_0 stand for the true value of σ , and let P_0 denote the true distribution of (\mathbf{X}, Y) . The expectation with respect to P_0 will be denoted by E_0 .

2.2 Contraction rates of projection posterior

The usual approach to Bayesian inference for model (2.1) with $f \in \mathcal{M}$ would be to put a prior on f supported within \mathcal{M} , and obtain the posterior distribution to make an inference. However, the shape restriction in \mathcal{M} forbids certain natural priors, such as the one on step functions with the step heights independently normally distributed, which allows fast calculations through conjugacy. A compliant prior will have to maintain the order restriction on the step heights,

which makes the posterior computation more challenging. More importantly, this will make frequentist analyses such as posterior contraction rates and limiting coverage of credible regions extremely hard. The projection-posterior approach provides a simple tool to “correct” a non-compliant posterior distribution by projecting posterior samples on the relevant parameter space and uses the resulting induced distribution to make inference, as in Lin and Dunson (2014) and Chakraborty and Ghosal (2021b, 2022). A generalization of this approach uses a broader “immersion map”, as described in the introduction. To obtain posterior contraction rate in terms of a global metric like an \mathbb{L}_1 -distance, we follow the projection-posterior approach in studying the posterior contraction rates as an important triangle inequality implied by the projection plays a central role in the derivation.

To study the asymptotic properties of the posterior distribution of f in the setting of a deterministic predictor, we consider the $\mathbb{L}_1(G_n)$ -distance, while for a random predictor arising from a distribution G , we also use the $\mathbb{L}_1(G)$ -distance. It will be seen that the projection posterior inherits the convergence properties of the original posterior if the same metric is used to obtain the projection, and hence it will be sufficient to study the unrestricted posterior, which can be done using traditional tools like moment bounding or by applying the general theory of posterior contraction (cf., Ghosal and van der Vaart (2017)). For random predictors, another alternative is to use the Lebesgue \mathbb{L}_1 -distance. If G admits a density bounded above and below, then the $\mathbb{L}_1(G)$ -distance and the Lebesgue \mathbb{L}_1 -distance are equivalent, and hence they lead to the same rate. It is also sensible to consider \mathbb{L}_p -distances for p different from 1, but the weaker \mathbb{L}_p -approximation property (see Lemma A.1.2) will lead to a suboptimal contraction rate $n^{-1/(pd+2)}$ for $1 < p \leq 2$. For the univariate case $d = 1$, Chakraborty and Ghosal (2021b) improved the \mathbb{L}_p -rate to the optimal rate $n^{-1/3}$ up to a logarithmic factor by using variable knots and by putting a prior on the knots, but the corresponding improved approximation result does not seem to be obtainable in the multivariate case.

We make the following assumptions.

Assumption 1 (Design). The predictor variables $\mathbf{X}_1, \dots, \mathbf{X}_n$ are deterministic and that $\max\{N_j :$

$j \in [1 : J] \lesssim n/J^d$, or are sampled i.i.d. from a distribution G with bounded density g .

Assumption 2 (Data). The true regression function $f_0 \in \mathcal{M}$ and the true distribution of the regression error ε has mean zero and true variance σ_0^2 .

Assumption 3 (Prior). The parameters ζ_j and λ_j in the prior on the coefficients θ_j satisfy $\max_j |\zeta_j| < \infty$ and $0 < \min_j \lambda_j^2 \leq \max_j \lambda_j^2 < \infty$.

If the number J of steps in each direction is not chosen deterministically, then it is given a prior supported on \mathbb{N} satisfying the tail condition

$$\exp\{-b_2 J^d \log J\} \leq \pi(J) \leq \exp\{-b_1 J^d \log J\}, \quad (2.6)$$

where b_1 and b_2 are positive hyperparameters.

In this section, we only use $\mathbf{J} = (J, \dots, J)^\top$. Define

$$\mathcal{F}_J = \{f : f = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j}, \text{ for } \theta_j \in \mathbb{R}, \forall j\}.$$

Let $\mathcal{M}_J = \mathcal{F}_J \cap \mathcal{M}$. To comply with the shape constraints, we project the posterior of f onto the monotone function space \mathcal{M}_J through the map

$$f \mapsto f^* \in \arg \min \{\rho(f, h) : h \in \mathcal{M}_J\}, \quad (2.7)$$

provided the minimizer exists, where ρ is the metric of interest. Note that for $f = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j} \in \mathcal{F}_J$, the condition of monotonicity is equivalent to that the array of the coefficients lies in the convex cone

$$\mathcal{C} = \{\boldsymbol{\theta} = (\theta_j : j \in [1 : J]) : \theta_{j_1} \leq \theta_{j_2}, \text{ if } j_1 \preceq j_2\}. \quad (2.8)$$

In this paper, ρ will be taken as the $\mathbb{L}_p(G^*)$ -distance for a distribution G^* on $[0, 1]^d$, possibly depending on n (such as G_n), and some $p \geq 1$, usually 1. By minimizing the $\mathbb{L}_p(G^*)$ -distance over

\mathcal{M}_J , we will get the projection posterior samples, and the corresponding induced distribution as the projection-posterior distribution to make an inference. Let the Lebesgue measure on $[0, 1]^d$ be denoted by λ . The following result shows that the projection posterior given by the $\mathbb{L}_p(\lambda)$ -projection onto \mathcal{M} charges only \mathcal{M}_J .

Proposition 2.2.1. *For any f in \mathcal{F}_J and $p \geq 1$, its $\mathbb{L}_p(\lambda)$ -projection onto \mathcal{M} , f^* , exists, and f^* is also the solution of the minimization problem $\min\{\|f - h\|_{p,\lambda} : h \in \mathcal{M}_J\}$.*

However, for a general distribution G^* , the $\mathbb{L}_p(G^*)$ -projection of $f \in \mathcal{F}_J$ onto \mathcal{M}_J is not necessarily the $\mathbb{L}_p(G^*)$ -projection onto \mathcal{M} . That means, given $f \in \mathcal{F}_J$, the minimization problem $\min\{\|f - h\|_{p,G^*} : h \in \mathcal{M}\}$ may possess no solution in \mathcal{F}_J , as the minimization problem also depends on the weighting distribution G^* . This is different from the univariate case, where the same minimizing problem always has solutions in \mathcal{M}_J .

We focus on the $\mathbb{L}_p(G^*)$ -projection onto \mathcal{M}_J . For $f \in \mathcal{F}_J$, the minimizing problem then becomes,

$$\min_{\theta^* \in \mathcal{C}} \sum_{j \in [1:J]} |\theta_j - \theta_j^*|^p G^*(I_j). \quad (2.9)$$

The solution of isotonic optimization problem in (2.9) is available in some R packages like ‘isotone’, see de Leeuw et al. (2009). It is a convex optimization problem with a set of linear constraints in (2.8), so a general convex optimization algorithm, such as an active-set method or an interior-point method, can be applied. However, algorithms specially designed for isotonic regression may obtain the solution faster. By the algorithms given in Stout (2013), problem (2.9) can be solved in $O(J^d \log J)$ steps when $d = 2$, and in $O(J^{2d} \log J)$ steps when $d \geq 3$. It is clear that the solution is unique if $p > 1$ and $G^*(I_j) > 0$ for all j , by the strict convexity of the $\mathbb{L}_p(G^*)$ -norm. For the $\mathbb{L}_1(G^*)$ -norm, the solution may not be unique, but any solution may be chosen to define the projection-posterior. The convergence properties are not affected by the choice. For the choice $G^* = G_n$ primarily used in this paper, the minimization in (2.9) reduces

to

$$\min_{\theta^* \in \mathcal{C}} \sum_{j \in [1:J]} N_j |\theta_j^* - \theta_j|^p, \quad (2.10)$$

while the use of the Lebesgue measure leads to the unweighted isotonization problem of the minimization of $\sum_{j \in [1:J]} |\theta_j^* - \theta_j|^p$ subject to the restriction that $\theta^* \in \mathcal{C}$.

Let a sample from the projection-posterior defined by the minimization of an \mathbb{L}_1 -distance, be denoted by $f^* = \sum_{j \in [1:J]} \theta_j^* \mathbb{1}_j$. The first part of the following theorem under abstract conditions gives the projection-posterior contraction rates with respect to a variety of \mathbb{L}_1 -metrics. In the second part of the theorem, the conclusion is specialized to the empirical \mathbb{L}_1 -metric or the \mathbb{L}_1 -metric with respect to the distribution of the predictor under easily verifiable conditions.

Theorem 2.2.1. *Let J be deterministic, Assumptions 2–3 hold and let G^* be a distribution on $[0, 1]^d$ possibly depending on n and $\mathbf{X}_1, \dots, \mathbf{X}_n$ satisfying the conditions that*

$$\mathbb{E} \left[\max_{j \in [1:J]} G^*(I_j) \right] \lesssim J^{-d}, \quad \mathbb{E} \left[\sum_{j \in [1:J]} G^*(I_j) (N_j + 1)^{-1} \right] \lesssim J^d / n. \quad (2.11)$$

Let f^ be the $\mathbb{L}_1(G^*)$ -projection of f sampled from the unrestricted posterior on \mathcal{F}_J . Moreover, assume that either σ is known, or a consistent estimator is plugged-in, or that the posterior distribution of σ is consistent. Then for $\epsilon_n = \max\{\sqrt{J^d/n}, J^{-1}\}$, we have that*

$$\mathbb{E}_0 \Pi(\|f^* - f_0\|_{1, G^*} > M_n \epsilon_n | \mathbb{D}_n) \rightarrow 0 \text{ for any } M_n \rightarrow \infty. \quad (2.12)$$

The optimal $\mathbb{L}_1(G^)$ -rate $n^{-1/(2+d)}$ is obtained above by choosing $J \asymp n^{1/(2d+1)}$.*

Further, let Assumption 1 hold, and if the predictor is random, assume that $J^d(\log n)/n \rightarrow 0$. Then the assertion (2.12) holds for G^ the empirical distribution G_n for both deterministic and random predictor, and also for $G^* = G$ if the predictor is random with distribution G .*

The optimal rate above reduces to the \mathbb{L}_1 -optimal rate $n^{-1/3}$ in the univariate case obtained by Chakraborty and Ghosal (2021b). We may also like to study the posterior contraction rate

with respect to the \mathbb{L}_p -metric. However, for $p > 1$, the \mathbb{L}_p -approximation rate by the step function f_j is weaker, only $J^{-1/p}$, at monotone functions with jumps; see Remark 7. Hence the \mathbb{L}_p -contraction rate of the corresponding procedure will be suboptimal.

The distribution of a random predictor \mathbf{X} is often unknown, but we can compute the $\mathbb{L}_1(G_n)$ -projection. The following corollary asserts that for random predictors with density bounded and bounded away from 0, the $\mathbb{L}_1(G_n)$ -projection posterior achieves the same posterior contraction rate with respect to the $\mathbb{L}_1(\lambda)$ -metric (and hence also under the $\mathbb{L}_1(G)$ -metric, which is equivalent under the assumed condition).

Corollary 2.2.2. *Let X_1, \dots, X_n be i.i.d. with distribution G admitting a density function g bounded and bounded away from 0. Let J be deterministic, $J \rightarrow \infty$ and $J^d(\log n)/n \rightarrow 0$. Then under Assumptions 2 and 3, for $\epsilon_n = \max\{\sqrt{J^d/n}, J^{-1}\}$ and any $M_n \rightarrow \infty$, $E_0\Pi(\|f^* - f_0\|_{1,\lambda} > M_n \epsilon_n | \mathbb{D}_n) \rightarrow 0$ where f^* is the $\mathbb{L}_1(G_n)$ -projection of f sampled from the unrestricted posterior.*

2.3 Bayesian tests for multivariate monotonicity

Next, we shall construct a Bayesian test for the multivariate coordinatewise monotonicity. A natural Bayesian test is based on the posterior probability of the region under the null hypothesis, that is, reject the hypothesis if $\Pi(f \in \mathcal{M} | \mathbb{D}_n)$ is less than 0.5. However, such a test cannot be consistent since, non-monotone functions will also lie in any neighborhood of a monotone function, so posterior consistency does not imply that the test will be consistent. In numerical experiments, we observe that the Lebesgue \mathbb{L}_1 -distance between a sample from the unrestricted posterior and the set \mathcal{M} is often positive for sample size up to 1000. To avoid a false rejection of the null hypothesis, we enlarge the class of monotone functions to include functions separated by a distance at most δ_n , where δ_n decreases with n appropriately. Then we consider the posterior probability of the enlarged set, $\Pi(\rho(f, \mathcal{M}) \leq \delta_n | \mathbb{D}_n)$, where ρ is a suitable metric, usually an \mathbb{L}_1 -distance. This idea was also pursued in Salomond (2014a) and Chakraborty and Ghosal (2021b) for Bayesian tests for monotonicity in the univariate case, respectively using

the \mathbb{L}_∞ - and an \mathbb{L}_1 -distance. Below, we consider random predictors obtained from a fixed distribution G independently. The following result shows that the resulting test is consistent at the null and at all fixed alternatives, and the power goes to one at an alternative belonging to a Hölder smooth class $\mathcal{H}(\alpha, L)$ (see Definition A.1.1) even if the alternative approaches the null, provided that happens sufficiently slowly.

Theorem 2.3.1. *Let Assumptions 1–3 hold for a random predictor with distribution G , and let ρ stand for the $\mathbb{L}_1(G)$ -distance. Let $\gamma \in (0, 1)$ and $M_n \rightarrow \infty$ be predetermined and $J \asymp n^{1/(2+d)}$. Then for the test $\phi_n = \mathbb{1}\{\Pi(\rho(f, \mathcal{M}_J) \leq M_n n^{-1/(d+2)} | \mathbb{D}_n) < \gamma\}$, we have*

- (i) $E_0 \phi_n \rightarrow 0$ for any fixed $f_0 \in \mathcal{M}$;
- (ii) $E_0(1 - \phi_n) \rightarrow 0$ for any fixed integrable $f_0 \notin \overline{\mathcal{M}}$, where $\overline{\mathcal{M}}$ is the $\mathbb{L}_1(G)$ -closure of \mathcal{M} ;
- (iii) $\sup\{E_0(1 - \phi_n) : f_0 \in \mathcal{H}(\alpha, L), \rho(f_0, \mathcal{M}) > \tau_n(\alpha)\} \rightarrow 0$, where

$$\tau_n(\alpha) = \begin{cases} C n^{-\alpha/(2+d)}, & \text{for some } C > 0 \text{ if } \alpha < 1, \\ C M_n n^{-1/(2+d)}, & \text{for any } C > 1 \text{ if } \alpha = 1. \end{cases}$$

The separation rate $n^{-\alpha/(2+d)}$ appearing above for consistency at smooth alternatives is weaker than the corresponding rate $n^{-\alpha/(2\alpha+d)}$ for estimation. This is because the value of $J \asymp n^{1/(2+d)}$ is optimal for estimating monotone functions, but is suboptimal for estimating α -smooth functions. The problem can be avoided simultaneously for all $\alpha \leq 1$ by putting a prior on J and using a larger enlargement in terms of the weaker Hellinger distance on the density

$$p_{f,\sigma}(x, y) = (\sigma \sqrt{2\pi})^{-1} \exp[-(y - f(x))^2 / (2\sigma^2)] \quad (2.13)$$

of (\mathbf{X}, Y) (with respect to the product of G and the Lebesgue measure) with size dependent on the random J drawn from its posterior distribution. In this case, the posterior sampling

is more involved as the posterior probabilities of each value of J also need to be obtained, which involves computations of a large matrix and its determinant, and a stronger separation is needed in terms of the weaker Hellinger metric.

Theorem 2.3.2. *Let σ be known, Assumptions 1–3 hold for a random predictor with distribution G , and Lebesgue density g bounded away from zero. Assume ε is sub-Gaussian. Let ρ stand for the Hellinger metric on the density of (X, Y) induced on the regression function, that is,*

$$\rho^2(f_1, f_2) = 2 \left\{ 1 - (2\pi\sigma^2)^{-1/2} \int \exp[-(f_1(x) - f_2(x))^2 / (8\sigma^2)] dG(x) \right\}. \quad (2.14)$$

Let J be given a prior satisfying (2.6). Consider the test

$$\phi_n = \mathbb{1}\{\Pi(\rho(f, \mathcal{M}_J) \leq M_0 \sqrt{(J^d \log n)/n} | \mathbb{D}_n) < \gamma\},$$

for a predetermined $\gamma \in (0, 1)$ and a sufficiently large $M_0 > 0$. Assume that f_0 is bounded. Then

- (i) for any fixed $f_0 \in \mathcal{M}$, $E_0 \phi_n \rightarrow 0$;
- (ii) for any fixed f_0 integrable on $[0, 1]^d$, and $f_0 \notin \overline{\mathcal{M}}$, $E_0(1 - \phi_n) \rightarrow 0$, where $\overline{\mathcal{M}}$ is the $\mathbb{L}_1(G)$ -closure of \mathcal{M} ;
- (iii) for alternatives in the Hölder function class, we have for a sufficiently large constant $C > 0$,

$$\sup\{E_0(1 - \phi_n) : f_0 \in \mathcal{H}(\alpha, L), \rho(f_0, \mathcal{M}) > C(n/\log n)^{-\alpha/(1+2\alpha)}\} \rightarrow 0.$$

Remark 1. In both results on testing, we can allow deterministic predictors with ρ replaced by the $\mathbb{L}_1(G_n)$ -distance to derive properties (i) and (iii). This follows from a similar proof by obtaining posterior contraction with respect to the $\mathbb{L}_1(G_n)$ -metric using Theorem 8.26 of Ghosal and van der Vaart (2017) for deterministic predictors.

Remark 2. As the distribution G of the random predictor is typically unknown, the tests used in Theorems 2.3.1 and 2.3.2 are not generally computable. If G admits a density also bounded

away from 0, then the $\mathbb{L}_1(G)$ -metric and the Hellinger metric given by (2.14) may be respectively replaced by the Lebesgue \mathbb{L}_1 -metric and by ρ defined by

$$\rho^2(f_1, f_2) = 2\{1 - (2\pi\sigma^2)^{-1/2} \int \exp[-(f_1(x) - f_2(x))^2 / (8\sigma^2)] dx\}. \quad (2.15)$$

Then the conclusions of the theorems hold. For Theorem 2.3.1, this follows by following the same arguments by using Part (iii) of Theorem 2.2.1 instead of Part (ii). For Theorem 2.3.2, we use the equivalence of the metrics (2.14) and (2.15) under the assumed condition and the equivalence of the projections. Moreover, the conclusion in Part (iii) of both theorems can be strengthened by replacing the Hölder class with the corresponding Sobolev class $\mathcal{W}(\alpha, L)$; see Definition C.6 of Ghosal and van der Vaart (2017). This is because the approximation rate $J^{-\alpha}$ for α -smooth function by step function with J intervals in each direction holds also for the more general Sobolev class, as the \mathbb{L}_2 -norm is stronger than the \mathbb{L}_1 -norm.

2.4 Immersion posterior and frequentist coverage

The unrestricted posterior distribution of f given σ is induced from (2.4) by the representation $f = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j}$. To obtain the immersion posterior distribution to make an inference, we consider three possible immersion maps.

Recall that

$$\mathcal{M}_J = \left\{ f = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j} : \theta_j \in \mathbb{R} \text{ and } \theta_{j_1} \leq \theta_{j_2} \text{ if } j_1 \preceq j_2 \right\}, \quad (2.16)$$

consisting of the coordinatewise nondecreasing functions taking constant value on every I_j .

Based on the isotonization procedure introduced in Fokianos et al. (2020), consider transformations $\underline{\iota}$ and $\bar{\iota}$ acting on $f = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j} \in \mathcal{K}_J$ mapping to an element of \mathcal{M}_J defined

by

$$\underline{\iota}(f)(\mathbf{x}) = \max_{j_1 \leq j_0(\mathbf{x})} \min_{\substack{j_0(\mathbf{x}) \leq j_2 \\ N_{[j_1:j_2]} > 0}} \frac{\sum_{j \in [j_1:j_2]} N_j \theta_j}{N_{[j_1:j_2]}}, \quad (2.17)$$

$$\bar{\iota}(f)(\mathbf{x}) = \min_{j_0(\mathbf{x}) \leq j_2} \max_{\substack{j_1 \leq j_0(\mathbf{x}) \\ N_{[j_1:j_2]} > 0}} \frac{\sum_{j \in [j_1:j_2]} N_j \theta_j}{N_{[j_1:j_2]}}, \quad (2.18)$$

where $j_0(\mathbf{x}) = \lfloor \mathbf{x} \circ \mathbf{J} \rfloor$, $N_{[j_1:j_2]} = \sum_{j \in [j_1:j_2]} N_j$, and $\mathbf{x} \in [0, 1]^d$, for j_1, j_2 in \mathbb{Z}^d . The immersion posterior can be derived through the immersion map, ι , which is chosen to be either $\underline{\iota}$ or $\bar{\iota}$. This is determined by examining the resulting induced distribution of

$$f_* = \underline{\iota}(f), \quad (2.19)$$

$$f^* = \bar{\iota}(f). \quad (2.20)$$

It is obvious that $\iota(f) \in \mathcal{M}_J$ and $\iota(f) = f$ if $f \in \mathcal{M}_J$ and $N_j > 0$ for all $j \in [\mathbf{1} : \mathbf{J}]$. Generally, $\underline{\iota}(f)(\mathbf{x}) \leq \bar{\iota}(f)(\mathbf{x})$ for any $\mathbf{x} \in [0, 1]^d$, but this may fail to hold if $N_j = 0$ for some j , see Deng et al. (2021). To neutralize the effect stemming from the order of minimization and maximization, we propose using the average of $\underline{\iota}$ and $\bar{\iota}$, leading to another immersion map $\iota = (\underline{\iota} + \bar{\iota})/2$, by which f is mapped to

$$\tilde{f} = (f_* + f^*)/2. \quad (2.21)$$

The projection map for the univariate case is typically computed by the pool adjacent violator algorithm (see Section 2.3 of (Barlow et al. 1972)), which requires $O(J)$ computations for a function with J steps. The computation of f_* or f^* requires no more than $(\prod_{k=1}^d J_k)^3$ operations by the brute-force search utilizing the block max-min or min-max formulas.

2.4.1 Effect of the immersion map

To see the effect of the immersion map on the posterior distribution of the function value at a point $\mathbf{x}_0 = (0.5, 0.5) \in [0, 1]^2$, we conduct a small simulation study and compare the unrestricted and immersion posterior density for a randomly generated sample of three different sizes $n = 100, 200$, and 500 , and several different regression functions:

1. $f_{0,1}(x_1, x_2) = x_1 + x_2$;
2. $f_{0,7}(x_1, x_2) = \sqrt{x_1 + x_2}$;
3. $f_{0,9}(x_1, x_2) = \mathbb{1}\{x_1 < 1/3\} + 2\mathbb{1}\{1/3 \leq x_1 < 2/3\} + 3\mathbb{1}\{x_1 \geq 2/3\}$.

Let X_1 and X_2 be distributed independently and uniformly on $[0, 1]$ and error $\varepsilon \sim N(0, \sigma^2)$ with true value of σ to be 0.1 . We choose the number of grid points $J_1 = J_2 = J = \lceil n^{1/4} \log_{10} n \rceil$. The random heights, $\{\theta_{(j_1, j_2)} : j_1, j_2 \leq J\}$, are endowed with the independent Gaussian prior $N(0, 1000\sigma^2)$. The variance σ^2 is estimated using the maximum marginal likelihood method. We plot both the unrestricted posterior density and the estimated immersion posterior density in the same figure. The latter is based on $2,000$ posterior samples transformed by the immersion map $(\bar{t} + \underline{t})/2$.

As evident from Figure 2.1, the immersion posterior density functions exhibit lower variance across all instances, albeit to varying degrees depending on the true regression functions and sample sizes. Furthermore, the modes of the immersion posterior are nearer to the true value. The impacts of the other immersion maps, \bar{t} and \underline{t} , on the posterior were found to be similar.

2.4.2 Coverage of Credible Intervals

Let $\mathbf{x}_0 \in (0, 1)^d$ be fixed. Suppose that we want to make an inference on $f(\mathbf{x}_0)$. For a given $0 < \gamma < 1$, consider a $(1 - \gamma)$ -credible interval with endpoints the $\gamma/2$ and $(1 - \gamma/2)$ quantiles of $f_*(\mathbf{x}_0)$, $f^*(\mathbf{x}_0)$, or $\tilde{f}(\mathbf{x}_0)$ defined in (2.19) – (2.21). To obtain the limiting frequentist coverage of

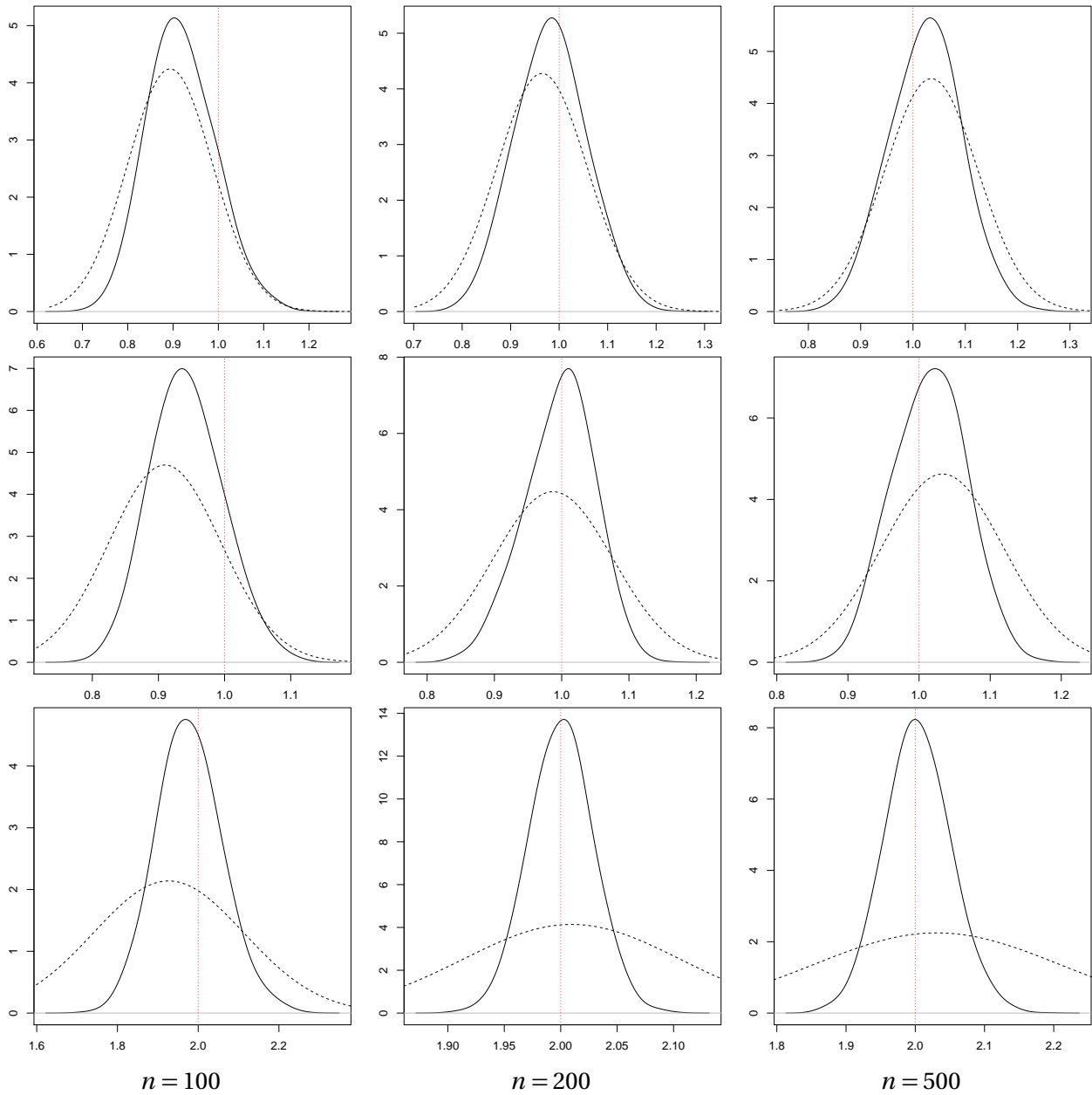


Figure 2.1: Unrestricted and immersion posterior density functions of $f(x_0)$. The solid black line stands for the immersion posterior density, while the black dashed line represents the unrestricted posterior density. The true function value is indicated by the red dotted vertical line. The rows correspond to functions (i), (ii), and (iii), respectively, while the columns represent sample sizes of $n = 100, 200, \text{ and } 500$.

these credible intervals, we obtain the weak limit of the immersion posterior distributions of f for all three immersion maps \underline{l} , \bar{l} and $(\underline{l} + \bar{l})/2$ at $\mathbf{x} = \mathbf{x}_0$.

The first assumption is about the local regularity of the true regression function f_0 near a point of interest \mathbf{x}_0 . This assumption, as in Han and Zhang (2020), is an essential ingredient to establish the limiting distribution.

Assumption 4. Let $f_0 \in \mathcal{M}$. For $\mathbf{x}_0 \in (0, 1)^d$ and $1 \leq k \leq d$, let β_k be the order of the first non-zero derivative of f at \mathbf{x}_0 along the k th coordinate, that is, $\beta_k = \min_{l \geq 1} \{l : \partial_k^l f_0(\mathbf{x}_0) \neq 0\}$ and $\beta_k = \infty$ if $\partial_k^l f_0(\mathbf{x}_0) = 0$ for all $l \geq 1$. Without loss of generality, we may assume that f_0 depends on its first s arguments locally at \mathbf{x}_0 , that is, $1 \leq \beta_1, \dots, \beta_s < \infty$, and that $\beta_{s+1} = \dots = \beta_d = \infty$ for some $0 \leq s \leq d$. Define an index set $L = \{\mathbf{l} : 0 < \sum_{k=1}^s l_k / \beta_k \leq 1 \text{ and } l_k = 0, \text{ for } k = s+1, \dots, d\}$. For a positive sequence $\omega_n \downarrow 0$, set $\mathbf{r}_n = (\omega_n^{1/\beta_1}, \dots, \omega_n^{1/\beta_s}, 1, \dots, 1)^\top$. For any $t > 0$,

$$\lim_{\omega_n \downarrow 0} \omega_n^{-1} \sup_{\substack{\mathbf{x} \in [0, 1]^d, \\ |x_k - x_{0,k}| \leq t r_{n,k}, \\ 1 \leq k \leq d}} \left| f_0(\mathbf{x}) - f_0(\mathbf{x}_0) - \sum_{\mathbf{l} \in L} \frac{\partial^{\mathbf{l}} f_0(\mathbf{x}_0)}{\mathbf{l}!} (\mathbf{x} - \mathbf{x}_0)^{\mathbf{l}} \right| = 0. \quad (2.22)$$

Assumption 4 takes into account varying convergence rates across different coordinates, according to their respective smoothness levels. Each term in the expansion contributes towards approximation rates larger than or equal to ω_n . Let

$$L_0 = \{\mathbf{l} : 0 < \sum_{k=1}^s l_k / \beta_k < 1 \text{ and } l_k = 0 \text{ for } k = s+1, \dots, d\}, \quad (2.23)$$

$$L^* = \{\mathbf{l} : \sum_{k=1}^s l_k / \beta_k = 1 \text{ and } l_k = 0 \text{ for } k = s+1, \dots, d\}. \quad (2.24)$$

Under Assumption 4, a unique feature for functions in \mathcal{M} is that the derivatives of order $\mathbf{l} \in L_0$ are zero (see Lemma 1 of Han and Zhang (2020)). Only those derivatives corresponding to the index set L^* can be nonzero. Thus, the nonzero terms in the expansion of (2.22) contribute the same approximation rate ω_n . However, Assumption 4 cannot exclude the nonzero mixed derivatives. Additional assumptions will be needed when we want to eliminate the mixed

derivative terms.

Next, we make the following assumption on the distributions of the covariate \mathbf{X} and the error ε from the data generating process (2.1).

Assumption 5. The covariate \mathbf{X} has a density g such that $a_1 \leq g(\mathbf{x}) \leq a_2$ for all $\mathbf{x} \in [0, 1]^d$ and $0 < a_1 \leq a_2 < \infty$. Suppose g is continuous in a neighborhood of the set $\{(x_{0,1}, \dots, x_{0,s}, x_{s+1}, \dots, x_d) : x_k \in [0, 1], \text{ for } s+1 \leq k \leq d\}$. The random error ε , with mean 0 and variance σ_0^2 , has a finite $2(\sum_{k=1}^s \beta_k^{-1} + 1)$ -th moment.

Let H_1 and H_2 be two independent centered Gaussian processes indexed by $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$ with the covariance kernel

$$\prod_{k=1}^s (u_k \wedge u'_k + v_k \wedge v'_k) D_s(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}'), \quad (2.25)$$

where $D_d(\mathbf{u}, \mathbf{v}) = g(\mathbf{x}_0)$, where g is the probability density function of \mathbf{X} , and for $s = 0, \dots, d-1$, and $D_s(\mathbf{u}, \mathbf{v})$ is given by

$$\int_{\substack{x_k \in [(x_0 - \mathbf{u})_k, (x_0 + \mathbf{v})_k] \cap [0, 1] \\ s+1 \leq k \leq d}} g(x_{0,1}, \dots, x_{0,s}, x_{s+1}, \dots, x_d) dx_{s+1} \cdots dx_d. \quad (2.26)$$

Additionally, we define a Gaussian process

$$\begin{aligned} U(\mathbf{u}, \mathbf{v}) = & \frac{\sigma_0 H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} + \frac{\sigma_0 H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \\ & + \sum_{l \in L^*} \frac{\partial^l f_0(x_0)}{(l+1)!} \prod_{k=1}^s \frac{v_k^{l_k+1} - (-u_k)^{l_k+1}}{u_k + v_k} \end{aligned} \quad (2.27)$$

indexed by $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$, and its functionals

$$Z_* = \sup_{\substack{\mathbf{u} \geq \mathbf{0} \\ u_k \leq x_{0,k} \\ s+1 \leq k \leq d}} \inf_{\substack{\mathbf{v} \geq \mathbf{0} \\ v_k \leq 1 - x_{0,k} \\ s+1 \leq k \leq d}} U(\mathbf{u}, \mathbf{v}), \quad Z^* = \inf_{\substack{\mathbf{v} \geq \mathbf{0} \\ v_k \leq 1 - x_{0,k} \\ s+1 \leq k \leq d}} \sup_{\substack{\mathbf{u} \geq \mathbf{0} \\ u_k \leq x_{0,k} \\ s+1 \leq k \leq d}} U(\mathbf{u}, \mathbf{v}). \quad (2.28)$$

The following result describes the asymptotic behavior of the normalized immersion poste-

rior distributions of $f(\mathbf{x}_0)$. Recall that \mathbb{D}_n represents the data and $r_{n,k}$ in Assumption 4 is the convergence rate along the k th direction through adjusting the overall rate ω_n according to the local smoothness levels. The weak limit of the normalized immersion posterior distribution function plays a central role in the study of the limiting coverage of the credible intervals based on the immersion posterior quantiles.

Theorem 2.4.1. *Let $\omega_n = n^{-1/(2+\sum_{k=1}^s \beta_k^{-1})}$ and let $\mathbf{r}_n = (\omega_n^{1/\beta_1}, \dots, \omega_n^{1/\beta_s}, 1, \dots, 1)^\top$. Suppose that \mathbf{J} satisfies $J_k \gg r_{n,k}^{-1}$, for each $k = 1, \dots, d$, and $\prod_{k=1}^d J_k \ll n\omega_n$. Under Assumptions 4 and 5, for any $z \in \mathbb{R}$, we have*

$$\Pi(\omega_n^{-1}(f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \leq z | \mathbb{D}_n) \rightsquigarrow \mathbb{P}(Z_* \leq z | H_1); \quad (2.29)$$

$$\Pi(\omega_n^{-1}(f^*(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \leq z | \mathbb{D}_n) \rightsquigarrow \mathbb{P}(Z^* \leq z | H_1); \quad (2.30)$$

$$\Pi(\omega_n^{-1}(\tilde{f}(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \leq z | \mathbb{D}_n) \rightsquigarrow \mathbb{P}((Z_* + Z^*)/2 \leq z | H_1). \quad (2.31)$$

Furthermore, for any $(z_1, z_2) \in \mathbb{R}^2$,

$$\begin{aligned} & \Pi(\omega_n^{-1}(f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \leq z_1, \omega_n^{-1}(f^*(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \leq z_2 | \mathbb{D}_n) \\ & \rightsquigarrow \mathbb{P}(Z_* \leq z_1, Z^* \leq z_2 | H_1). \end{aligned} \quad (2.32)$$

Remark 3. We make some remarks on Theorem 2.4.1:

1. The weak limit is understood in the usual sense for random variables since we consider the limiting behavior of the random probability measure of a fixed set $(-\infty, z]$. We refer to the proof technique of Han and Zhang (2020), which provides the distributional theory for the block-estimator in general multivariate isotonic regression, especially the small and large deviation arguments therein.
2. For the choice of J_k , the lower bound $r_{n,k}^{-1}$ is essential for Theorem 2.4.1. That eliminates the effect of the roughness of piecewise constant function approximation in view of the local contraction rate. But the upper bound, $n\omega_n$ in Theorem 2.4.1, is not fundamentally

necessary for the validity of the weak limit. Instead, we can set the hyperparameters λ_j large enough, specifically, $\min \lambda_j^2 \gg \omega_n^{-1} \sqrt{n}$, to obtain the limiting theory. The rest proof of Theorem 2.4.1 will not be affected much without such an upper bound except for the treatment of σ^2 . The estimation of σ^2 is not a hard problem and any consistent procedure will work. For any $\beta_k \geq 1$ and any $0 \leq s \leq d$, we observe that $r_k \leq n^{-1/3}$ for all $1 \leq k \leq d$. Without the local smoothness information, we can choose $J_k \gg n^{1/3}$. On the other side, if we assume that $\beta_k = 1, 1 \leq k \leq d$, is the leading case for the multivariate regression function, we can then choose $J_k \gg n^{1/(2+d)}$. One may raise concerns that selecting J in this manner seems not optimal when the true regression function is less smooth. However, this concern typically does not pose a big issue in practice. When lacking prior information about the regression function, an appropriately large J_k can be chosen and a non-informative prior should be applied to θ_j , such as a normal prior with a large variance. Our empirical study indicates that opting for a larger J can indeed enhance the performance of our method. For practical applications, we recommend selecting J_k no less than 15, particularly when dealing with a smaller sample size. Notably, \mathbf{J} is not a tuning parameter in this context; the immersion posterior is governed by shape restrictions rather than a tuning process like bandwidth selection in kernel smoothing. The choice of \mathbf{J} does not influence the contraction rate or the distributional theory, distinguishing it from typical tuning parameters.

3. It is also important to note that we employ a working normal model to derive the posterior distribution. The validity of this method remains intact even when the model is misspecified. The finite moment condition for the random error ε can be relaxed to the second order, as in Han and Zhang (2020), by selecting a sufficiently large λ_j^2 as in the last point.

The covariance kernels of the processes H_1 and H_2 depend on g , and the limiting Gaussian process also involves the derivative values of f_0 at \mathbf{x}_0 . A considerable simplification happens in some special cases where the parameters appear through a scale parameter in the kernel. It will

be seen shortly that this fact has a far-reaching implication in that the limiting coverage of a credible interval constructed from the immersion posterior is free of the unknown parameters of the model. If L^* defined by (2.24) only contains $\beta_k e_k$ for $k = 1, \dots, s$, where e_k denotes the standard unit vector in \mathbb{R}^d with one in the k th component and zero elsewhere, then the limiting processes in Theorem 2.4.1 can be further simplified by self-similarity. A factor depending on f_0 comes out as a multiplicative constant, and the remaining factor is only a known functional of H_1 and H_2 . The case $s = d$ stands for the regular case that all directional derivatives of f_0 at \mathbf{x}_0 are positive at a certain order. Then the covariance kernel further simplifies as a completely known function and a factor involving derivatives of the regression function and predictor density g . The result is precisely formulated in the result below.

Proposition 2.4.1. *If $L^* = \{\beta_k e_k : 1 \leq k \leq s\}$, then*

$$\begin{aligned} & \sup_{\mathbf{u} \geq 0} \inf_{\mathbf{v} \geq 0} \left\{ \frac{\sigma_0 H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} + \frac{\sigma_0 H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \\ & \quad \left. + \sum_{k=1}^s \left[\frac{\partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \cdot \frac{v_k^{\beta_k + 1} - (-u_k)^{\beta_k + 1}}{u_k + v_k} \right] \right\} \\ & =_d A_\beta \cdot \sup_{\mathbf{u} \geq 0} \inf_{\mathbf{v} \geq 0} \left\{ \frac{H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} + \frac{H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \\ & \quad \left. + \sum_{k=1}^s \frac{v_k^{\beta_k + 1} - (-u_k)^{\beta_k + 1}}{u_k + v_k} \right\}, \end{aligned}$$

where $A_\beta = (\sigma_0^2 \prod_{k=1}^s \left(\frac{\partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \right)^{1/\beta_k})^{1/(2 + \sum_{k=1}^s \beta_k^{-1})}$.

Furthermore, if $s = d$, then the above expression further simplifies to

$$\tilde{A}_\beta \sup_{\mathbf{u} \geq 0} \inf_{\mathbf{v} \geq 0} \left\{ \frac{\tilde{H}_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\tilde{H}_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \sum_{k=1}^d \frac{v_k^{\beta_k + 1} - (-u_k)^{\beta_k + 1}}{u_k + v_k} \right\},$$

where $\tilde{A}_\beta = \left(\frac{\sigma_0^2}{g(\mathbf{x}_0)} \prod_{k=1}^d \left(\frac{\partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \right)^{1/\beta_k} \right)^{1/(2 + \sum_{k=1}^d \beta_k^{-1})}$, and \tilde{H}_1 and \tilde{H}_2 are two independent centered Gaussian processes with covariance kernel given by $\prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k)$, $(\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}') \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$.

The same conclusion also applies to the inf sup-functional obtained by switching the positions of the supremum and the infimum.

Remark 4 (Univariate case). We specialize to the univariate case $s = d = 1$, with a general β , expanding from the case $\beta = 1$ studied by Chakraborty and Ghosal (2021c). Then

$$\tilde{H}_i(u, v) =_d W_i(v) + W_i(-u) =_d W_i(v) - W_i(-u), \quad (u, v) \in \mathbb{R}_{\geq 0}^2, \quad (2.33)$$

where W_1, W_2 are two independent standard two-sided Brownian motions starting from 0. Observe that the sup-inf functional

$$\begin{aligned} & \sup_{u>0} \inf_{v>0} \left\{ \frac{\tilde{H}_1(u, v)}{u+v} + \frac{\tilde{H}_2(u, v)}{u+v} + \frac{v^{\beta+1} - (-u)^{\beta+1}}{u+v} \right\} \\ & =_d \sup_{u>0} \inf_{v>0} \left\{ \frac{(W_1(v) + W_2(v) + v^{\beta+1}) - (W_1(-u) + W_2(-u) + u^{\beta+1})}{v - (-u)} \right\}, \end{aligned}$$

coincides with the slope of the greatest convex minorant of the process $W_1(t) + W_2(t) + t^{\beta+1}$. By the switching relation (cf. Groeneboom and Jongbloed (2014), page 56), for any $z \in \mathbb{R}$,

$$\begin{aligned} & \mathbb{P} \left(\tilde{A}_\beta \sup_{u>0} \inf_{v>0} \left\{ \frac{\tilde{H}_1(u, v)}{u+v} + \frac{\tilde{H}_2(u, v)}{u+v} + \frac{v^{\beta+1} - (-u)^{\beta+1}}{u+v} \right\} \leq z \right) \\ & = \mathbb{P} \left(\operatorname{argmin} \{ W_1(t) + W_2(t) + t^{\beta+1} - \tilde{A}_\beta^{-1} z t : t \in \mathbb{R} \} \geq 0 \right). \end{aligned}$$

If $\beta = 1$, the last display can be further simplified by applying the change of variable, $t = s + z/(2\tilde{A}_1)$, and is equal to

$$\mathbb{P}(2\tilde{A}_1 \operatorname{argmin} \{ W_1(s) + W_2(s) + s^2 : s \in \mathbb{R} \} \leq z),$$

with $\tilde{A}_1 = (\sigma_0^2 f'(x_0)/(2g(x_0)))^{1/3}$. This reproduces the main result of Chakraborty and Ghosal (2021c).

Now we are ready for the evaluation of the limiting coverage of an immersion posterior

credible interval for $f(\mathbf{x}_0)$. Let

$$Q_{n,\gamma}^{(1)} = \inf\{z : \Pi(f_*(\mathbf{x}_0) \leq z | \mathbb{D}_n) \geq 1 - \gamma\} \quad (2.34)$$

stand for the $(1 - \gamma)$ -quantile of $f_*(\mathbf{x}_0)$. Similarly, let $Q_{n,\gamma}^{(2)}$ and $Q_{n,\gamma}^{(3)}$ stand for that of $f^*(\mathbf{x}_0)$ and $\tilde{f}(\mathbf{x}_0)$ respectively. Let $\tilde{U}(\mathbf{u}, \mathbf{v})$ stand for the Gaussian process

$$\frac{\tilde{H}_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\tilde{H}_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \sum_{k=1}^d \frac{v_k^{\beta_k+1} - (-u_k)^{\beta_k+1}}{u_k + v_k} \quad (2.35)$$

indexed by $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$.

The following result gives the ultimate conclusion of the paper about asymptotic coverage of credible intervals for the regression function value at an interior point.

Theorem 2.4.2. *Under the assumed setup, Assumptions 4 and 5, and the condition that $L^* = \{\beta_k e_k : 1 \leq k \leq s\}$, the asymptotic coverage of the quantile-based one-sided credible interval $(-\infty, Q_{n,\gamma}^{(1)}]$ is given by*

$$\begin{aligned} & \mathbb{P}\left(\mathbb{P}\left(\sup_{\mathbf{u} \geq 0} \inf_{\mathbf{v} \geq 0} \left\{ \frac{H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} + \frac{H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \right. \right. \\ & \left. \left. \left. + \sum_{k=1}^s \frac{v_k^{\beta_k+1} - (-u_k)^{\beta_k+1}}{u_k + v_k} \right\} \leq 0 \mid H_1\right) \leq 1 - \gamma\right). \end{aligned}$$

If $Q_{n,\gamma}^{(1)}$ is replaced by $Q_{n,\gamma}^{(2)}$, the above limit is changed by swapping the order of the supremum and infimum operations. If $Q_{n,\gamma}^{(1)}$ is replaced by $Q_{n,\gamma}^{(3)}$, the above limit is changed by replacing the expression on the right with the average of the sup inf and inf sup operations.

Moreover, if $s = d$,

- (i) $\mathbb{P}_0(f_0(\mathbf{x}_0) \leq Q_{n,\gamma}^{(1)}) \rightarrow \mathbb{P}(Z_B^{(1)} \leq 1 - \gamma)$;
- (ii) $\mathbb{P}_0(f_0(\mathbf{x}_0) \leq Q_{n,\gamma}^{(2)}) \rightarrow \mathbb{P}(Z_B^{(2)} \leq 1 - \gamma)$;
- (iii) $\mathbb{P}_0(f_0(\mathbf{x}_0) \leq Q_{n,\gamma}^{(3)}) \rightarrow \mathbb{P}(Z_B^{(3)} \leq 1 - \gamma)$,

where

$$Z_B^{(1)} = \mathbb{P}(\sup_{u \geq 0} \inf_{v \geq 0} \tilde{U}(u, v) \leq 0 | \tilde{H}_1),$$

$$Z_B^{(2)} = \mathbb{P}(\inf_{v \geq 0} \sup_{u \geq 0} \tilde{U}(u, v) \leq 0 | \tilde{H}_1),$$

and

$$Z_B^{(3)} = \mathbb{P}\left(\frac{1}{2} \left\{ \sup_{u \geq 0} \inf_{v \geq 0} \tilde{U}(u, v) + \inf_{v \geq 0} \sup_{u \geq 0} \tilde{U}(u, v) \right\} \leq 0 | \tilde{H}_1\right).$$

Proof. We observe that $f_0(\mathbf{x}_0) \leq Q_{n,\gamma}^{(1)}$ if and only if

$$\Pi(f_*(\mathbf{x}_0) \leq f_0(\mathbf{x}_0) | \mathbb{D}_n) = \Pi(\omega_n^{-1}(f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \leq 0 | \mathbb{D}_n) \leq 1 - \gamma.$$

Hence by Theorem 2.4.1 and Proposition 2.4.1, as the multiplicative positive constant in the limiting process can be dropped because the interval $(-\infty, 0]$ remains invariant under a scale-change, the first conclusion follows immediately. The special cases follow from the second part of Proposition 2.4.1. \square

Remark 5. For $d = 1$, $Z_B^{(1)}$, $Z_B^{(2)}$ and $Z_B^{(3)}$ all coincide, and may be simply denoted by Z_B as in Chakraborty and Ghosal (2021c).

The distributions of $Z_B^{(1)}$ and $Z_B^{(2)}$ are related, as shown next.

Proposition 2.4.2. *For any $z \in [0, 1]$, we have $\mathbb{P}(Z_B^{(1)} \leq z) = \mathbb{P}(Z_B^{(2)} \geq 1 - z)$, and $Z_B^{(3)}$ is symmetrically distributed about $1/2$.*

From Theorem 2.4.2 and Proposition 2.4.2, it follows that the limiting coverage of a one-sided Bayesian credible interval for $f(\mathbf{x}_0)$ using one of the three proposed immersion posteriors can be evaluated, is free of the true regression function (and also is free of the density g of the predictor if $s = d$, and hence depends only on the credibility level), but in general, need not be equal to the credibility. Nevertheless, a targeted limiting coverage can be obtained by starting with a certain credibility level that can be explicitly computed by back-calculation. As in the univariate monotone problems studied by Chakraborty and Ghosal (2021c,b), numerical

calculations show that the required credibility to obtain a specific limiting coverage is less than the targeted coverage, the opposite of the phenomenon Cox (1993) observed for smoothing problems. However, unlike in the univariate case where the limiting Bayes-Chernoff distribution determining the asymptotic coverage of the credible interval is symmetric, the corresponding random variables $Z_B^{(1)}$ and $Z_B^{(2)}$ for the posterior based on the immersion maps $\underline{\iota}$ and $\bar{\iota}$ appearing in the multivariate case are not symmetric. This has implications for the limiting coverage of a two-sided credible interval, which is more commonly used in practice. For instance, for $0 < \gamma < 1/2$, a two-sided $(1 - \gamma)$ -credible interval $[Q_{n,1-\gamma/2}, Q_{n,\gamma/2}]$ based on the immersion posterior using the map $\underline{\iota}$, the limiting coverage is given by $P(Z_B^{(1)} \leq 1 - \gamma/2) - P(Z_B^{(1)} \leq \gamma/2)$. The corresponding limit for the immersion posterior using the map $\bar{\iota}$ is $P(Z_B^{(2)} \leq 1 - \gamma/2) - P(Z_B^{(2)} \leq \gamma/2)$. Interestingly, a separate table for the distribution function of $Z_B^{(2)}$ is not needed, as it can be obtained from that of $Z_B^{(1)}$ in view of Proposition 2.4.2. The symmetry of $Z_B^{(3)}$, however, implies that the credibility level $1 - \gamma$ needed to make the asymptotic coverage of an equal-tailed $(1 - \gamma)$ -credible interval $1 - \alpha$ is obtained by choosing $1 - \gamma = 1 - 2F_{Z_B^{(3)}}^{-1}(\alpha/2)$, which is readily obtained once the cumulative distribution function $F_{Z_B^{(3)}}$ of $Z_B^{(3)}$ is tabulated.

2.5 Simulation

2.5.1 Distribution of Z_B

In this section, we present tables detailing the distribution and quantiles of Z_B for the case $d = 1$ when $\beta = 1, 3, 5$, as well as those for $Z_B^{(1)}, Z_B^{(2)}, Z_B^{(3)}$ for the case $d = 2$ when $\beta = (1, 1), (1, 3), (3, 3)$. The distributions of these variables are simulated using the Monte Carlo method, with the Gaussian processes concerned being generated by discrete approximation. The quantile table can function as a recalibration reference, to achieve the exact frequentist asymptotic coverage.

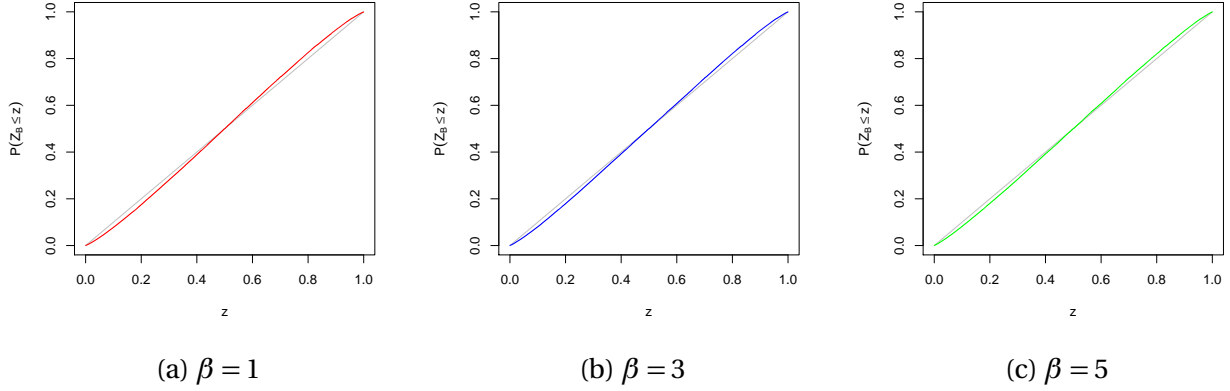


Figure 2.2: Distribution functions of Z_B .

Case $d = 1$

First, we generate approximations to the Gaussian processes \tilde{H}_1 and \tilde{H}_2 . Let \tilde{H} denote either \tilde{H}_1 or \tilde{H}_2 . To approximate \tilde{H} , we generate $14m$ independent standard Gaussian random variables, specifically $\{\zeta_j : j = 1, \dots, 7m\}$ and $\{\zeta'_j : j = 1, \dots, 7m\}$, where $m = 50$. Then \tilde{H} can be approximated as follows:

$$\tilde{H}(u, v) \approx \frac{1}{\sqrt{m}} \left[\sum_{j=1}^{[mu]} \zeta_j + \sum_{j=1}^{[mv]} \zeta'_j \right], \quad (2.36)$$

for $u, v \in [0, 7]$. Given each instance of \tilde{H}_1 , we generate 500 realizations of \tilde{H}_2 . For each realization, we compute the sup-inf functional. The proportion of non-positive outcomes then serves as a sample value of Z_B . We repeat the generation process 50,000 times to obtain the approximate distribution function of Z_B .

In Figure 2.2, we draw the simulated distribution functions of Z_B with $\beta = 1, 3$, and 5. We give the values of $P(Z_B \leq z)$ with different smoothness levels for selected z values in Table 2.1 and the values of the quantiles of Z_B 's distribution in Table 2.2.

Table 2.1: Values of $P(Z_B \leq z)$

z	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.990	0.995
$\beta = 1$	0.719	0.772	0.826	0.875	0.923	0.965	0.985	0.994	0.997
$\beta = 3$	0.715	0.768	0.821	0.870	0.921	0.963	0.983	0.994	0.997
$\beta = 5$	0.716	0.768	0.820	0.869	0.919	0.962	0.983	0.993	0.997

Table 2.2: Values of $q = \inf\{z : P(Z_B \leq z) \geq p\}$

p	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.990	0.995
$\beta = 1$	0.683	0.730	0.777	0.825	0.878	0.932	0.964	0.994	0.997
$\beta = 3$	0.687	0.734	0.781	0.829	0.882	0.935	0.966	0.986	0.992
$\beta = 5$	0.686	0.734	0.782	0.831	0.882	0.936	0.966	0.986	0.994

Case $d = 2$

To approximate $\tilde{H}(\mathbf{u}, \mathbf{v})$, for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, we generate 4 random matrices $\zeta^{(1)}$, $\zeta^{(2)}$, $\zeta^{(3)}$ and $\zeta^{(4)}$ with independent standard Gaussian random variables. The dimensions of these 4 matrices are $[m t_1] \times [m t_2]$, $[m s_1] \times [m t_2]$, $[m s_1] \times [m s_2]$, and $[m t_1] \times [m s_2]$, for $m = 5$ and $t_1 = t_2 = s_1 = s_2 = 5$. $\tilde{H}(\mathbf{u}, \mathbf{v})$ is then approximated by

$$\frac{1}{m} \left(\sum_{i=1}^{\lfloor m v_1 \rfloor} \sum_{j=1}^{\lfloor m v_2 \rfloor} \zeta_{ij}^{(1)} + \sum_{i=1}^{\lfloor m u_1 \rfloor} \sum_{j=1}^{\lfloor m v_2 \rfloor} \zeta_{ij}^{(2)} + \sum_{i=1}^{\lfloor m u_1 \rfloor} \sum_{j=1}^{\lfloor m u_2 \rfloor} \zeta_{ij}^{(3)} + \sum_{i=1}^{\lfloor m v_1 \rfloor} \sum_{j=1}^{\lfloor m u_2 \rfloor} \zeta_{ij}^{(4)} \right),$$

for $u_1, u_2, v_1, v_2 \in [0, 5]$.

To get a sample of any one of $Z_B^{(1)}$, $Z_B^{(2)}$ or $Z_B^{(3)}$, we first generate a sample of \tilde{H}_1 . Given this sample, we generate 500 realizations of \tilde{H}_2 . We then compute the three functionals that define $Z_B^{(1)}$, $Z_B^{(2)}$ and $Z_B^{(3)}$. The conditional probabilities are approximated by the frequency of non-positive functional values. This process is repeated 50,000 times for $\beta = (1, 1)$, $(3, 1)$, and $(3, 3)$, to estimate the distribution of $Z_B^{(1)}$, $Z_B^{(2)}$ or $Z_B^{(3)}$.

Since in higher-dimensional cases, $Z_B^{(1)}$ and $Z_B^{(2)}$ are not equal in distribution and their distribution functions are not symmetric about 0.5, we give both the values of $P(Z_B^{(1)} \leq z)$ and $P(Z_B^{(2)} \leq z)$ for some selected z values in Table 2.3. The corresponding distribution functions are plotted in Figure 2.3. We present the quantiles of $Z_B^{(3)}$ with different smoothness levels in

Table 2.4.

Table 2.3: Values of $P(Z_B \leq z)$ for various z and β , and $Z_B = Z_B^{(1)}, Z_B^{(2)}, Z_B^{(3)}$.

z	$\beta = (1, 1)$			$\beta = (3, 1)$			$\beta = (3, 3)$		
	$Z_B^{(1)}$	$Z_B^{(2)}$	$Z_B^{(3)}$	$Z_B^{(1)}$	$Z_B^{(2)}$	$Z_B^{(3)}$	$Z_B^{(1)}$	$Z_B^{(2)}$	$Z_B^{(3)}$
0.700	0.705	0.752	0.725	0.704	0.741	0.721	0.708	0.735	0.718
0.750	0.762	0.803	0.778	0.760	0.791	0.773	0.762	0.787	0.771
0.800	0.817	0.851	0.832	0.814	0.842	0.827	0.817	0.838	0.825
0.850	0.871	0.898	0.880	0.868	0.889	0.877	0.868	0.885	0.874
0.900	0.921	0.939	0.927	0.917	0.932	0.924	0.918	0.930	0.922
0.950	0.966	0.975	0.968	0.964	0.971	0.966	0.964	0.970	0.965
0.975	0.985	0.989	0.987	0.983	0.987	0.986	0.984	0.986	0.985
0.990	0.995	0.997	0.995	0.995	0.997	0.995	0.995	0.996	0.994
0.995	0.997	0.998	0.998	0.998	0.998	0.998	0.997	0.998	0.997

Table 2.4: Values of $q = \inf\{z : P(Z_B \leq z) \geq p\}$ for various p , and $Z_B = Z_B^{(1)}, Z_B^{(2)}, Z_B^{(3)}$.

p	$\beta = (1, 1)$			$\beta = (3, 1)$			$\beta = (3, 3)$		
	$Z_B^{(1)}$	$Z_B^{(2)}$	$Z_B^{(3)}$	$Z_B^{(1)}$	$Z_B^{(2)}$	$Z_B^{(3)}$	$Z_B^{(1)}$	$Z_B^{(2)}$	$Z_B^{(3)}$
0.700	0.697	0.653	0.677	0.699	0.665	0.681	0.695	0.669	0.684
0.750	0.741	0.699	0.724	0.743	0.711	0.728	0.741	0.715	0.732
0.800	0.787	0.749	0.771	0.789	0.759	0.776	0.787	0.763	0.778
0.850	0.833	0.801	0.819	0.835	0.811	0.823	0.833	0.815	0.825
0.900	0.881	0.855	0.872	0.883	0.865	0.876	0.883	0.869	0.878
0.950	0.933	0.917	0.928	0.937	0.925	0.931	0.935	0.927	0.933
0.975	0.963	0.951	0.959	0.965	0.957	0.962	0.965	0.959	0.964
0.990	0.983	0.977	0.982	0.985	0.981	0.984	0.985	0.983	0.984
0.995	0.991	0.987	0.990	0.991	0.989	0.992	0.993	0.991	0.992

2.5.2 Simulation for posterior contraction rate

We conduct a numerical study to assess the finite sample performance of the projection posterior methods for the estimation of isotonic regression functions. We use the projection posterior sample mean as our Bayesian estimator and compare the empirical \mathbb{L}_1 -distance

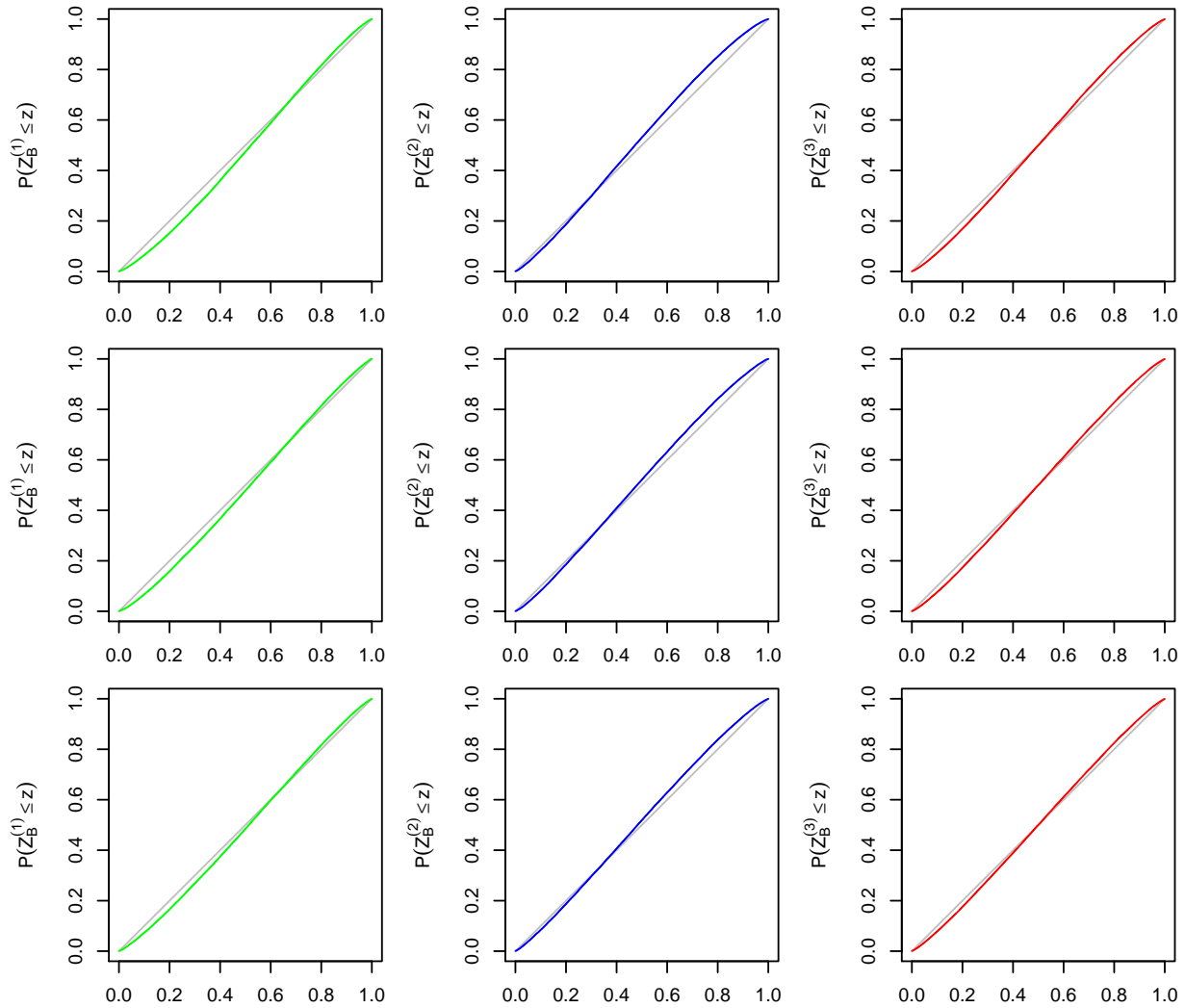


Figure 2.3: Distribution functions of $Z_B^{(1)}$, $Z_B^{(2)}$ or $Z_B^{(3)}$. The three plots in the first row are for $\beta = (1, 1)$; the second row is for $\beta = (3, 1)$; the last row is for $\beta = (3, 3)$.

between our estimator and the true regression function with that of the least square estimator on data sets of different sizes. We consider monotone regression functions:

- $f_1(x_1, x_2) = x_1 + x_2$,
- $f_2(x_1, x_2) = \exp\{x_1 x_2\}$,
- $f_3(x_1, x_2) = (x_1 + x_2)^2$,
- $f_4(x_1, x_2) = \sqrt{x_1 + x_2}$,

- $f_5(x_1, x_2) = (1 + \exp\{-6(x_1 + x_2 - 1)\})^{-1}$,
- $f_6(x_1, x_2) = 0$.

For each of sample size $n = 100, 200$, and 500 , and each regression function, we generate 20 data sets from the true regression model $Y = f_0(\mathbf{X}) + \varepsilon$ with \mathbf{X} uniformly distributed over $[0, 1]^2$ and independent errors $\varepsilon \sim N(0, 0.1^2)$. Set $J = \lceil n^{1/4} \log_{10} n \rceil$, which is chosen slightly larger than the optimal one to get a better approximation in lower sample sizes. For each data set, we generate $M = 1000$ unrestricted posterior sample functions. Then we compute the \mathbb{L}_1 -projection posterior, by the "activeSet" function in the R package "isotone". With the projection posterior samples, we then compute the empirical \mathbb{L}_1 -distance of the projection posterior mean function and the data-generating regression function. For the least square estimator, we use the same piecewise constant representation of the regression functions to obtain a function estimator on the whole range of \mathbf{X} and to make a fair comparison with our method. The least squares isotonic estimator is obtained by using the R package "isotonic.pen". We summarize the results in Table 2.5.

We can see from the table the Bayesian projection posterior estimator has a smaller \mathbb{L}_1 -error than the least squares estimator except for the last case of a constant function.

2.5.3 Simulation for Bayesian monotonicity testing

To test for $H_0 : f_0 \in \mathcal{M}$, we choose $J = \lceil n^{1/4} \rceil$, $\gamma = 0.5$ and $M_n = a(\log n)^b$, where a and b are two parameters to be determined. We run the procedure on several datasets of different sizes with both coordinatewise increasing and nonincreasing regression functions. Then we obtain the posterior samples of $\rho(f, \mathcal{M}_J)$, denoted by d . We fit a linear model of $\log(d n^{1/4})$ over $\log \log n$ to find the estimates of $\log a$ and b , which leads to $a = 0.237$ and $b = 0.234$. In the following simulation, we will choose $M_n = 0.237(\log n)^{0.234}$.

Since a test, frequentist or Bayesian, for multivariate monotonicity does not seem to exist in the literature before, we consider the following hypothesis testing procedure as the baseline

Table 2.5: The Lebesgue \mathbb{L}_1 -distance between the Bayesian projection posterior mean regression function (BP) and the true regression function and between the least squares isotonic regression function (LS) and the true one with standard deviations across all data sets marked in the parentheses.

	$n = 100$		$n = 200$		$n = 500$	
	BP	LS	BP	LS	BP	LS
f_1	0.054 (0.003)	0.059 (0.005)	0.045 (0.003)	0.050 (0.003)	0.034 (0.002)	0.041 (0.002)
f_2	0.049 (0.004)	0.051 (0.006)	0.040 (0.004)	0.043 (0.004)	0.030 (0.002)	0.034 (0.002)
f_3	0.085 (0.006)	0.089 (0.011)	0.072 (0.004)	0.074 (0.004)	0.055 (0.002)	0.058 (0.002)
f_4	0.040 (0.003)	0.045 (0.004)	0.032 (0.003)	0.038 (0.004)	0.024 (0.002)	0.030 (0.002)
f_5	0.051 (0.005)	0.052 (0.006)	0.041 (0.003)	0.044 (0.002)	0.032 (0.002)	0.044 (0.002)
f_6	0.032 (0.006)	0.021 (0.009)	0.026 (0.004)	0.018 (0.007)	0.021 (0.004)	0.012 (0.003)

method. We confine to the normal linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$, $i = 1, \dots, n$.

The hypothesis testing of multivariate monotonicity for affine functions becomes

$$H_0 : \beta_1 \geq 0 \text{ and } \beta_2 \geq 0, \text{ against } H_1 : \beta_1 < 0 \text{ or } \beta_2 < 0.$$

Given the significance level $\eta = 0.05$, we use the Bonferroni adjustment since we have only two parameters to be tested. We reject the null hypothesis when any one of the t-values of β_1 and β_2 is smaller than $t_{n-3, 1-\eta/2}$. To study the level of these two procedures, we consider functions, f_1, \dots, f_6 used in the last section. For the comparison of the power performance, we consider the following nonincreasing functions on $[0, 1]^2$:

- $f_7(x_1, x_2) = (x_1 + x_2 - 1)^2$.
- $f_8(x_1, x_2) = 2(x_1 + x_2 - 1)^3 - (x_1 + x_2 - 1)$.
- $f_9(x_1, x_2) = (x_1 + x_2 - 1)^3 - 0.5(x_1 + x_2 - 1)$.

- $f_{10}(x_1, x_2) = \sin((x_1 + x_2)\pi)$.
- $f_{11}(x_1, x_2) = x_1 - x_2$.
- $f_{12}(x_1, x_2) = \exp\{-10(x_1 + x_2 - 1)^2\} + x_1 + x_2$.

Even though the linear model is misspecified, it can summarize the overall trend of the regression function through the sign of the coefficients, and hence is appropriate. We also consider fitting a nonparametric regression using piecewise linear functions and test for the linear hypothesis that the slope coefficients on each piece in each direction are all nonnegative. Specifically, we take $J = 3$ and take the partition I_j for $j = (1, 1), \dots, (3, 3)$. On each I_j , we fit a linear model and test whether any t-value of the slope coefficients $\beta_{1,j}$ and $\beta_{2,j}$ is smaller than $t_{N_j-3, 1-\eta/18}$ by the Bonferroni adjustment.

We generate 200 datasets for each sample size $n = 100, 200$, and 500 . The predictors \mathbf{X} and ε are generated in the same way as in the last subsection. For the Bayesian procedure, we generate 200 posterior samples for each dataset and project each posterior sample to the monotone function class \mathcal{M} , denoting the projection posterior sample as f^* . Then $\rho_n(f, \mathcal{M})$ is obtained by computing $\rho_n(f, f^*)$, where ρ_n is the empirical \mathbb{L}_1 -distance. The results are summarized in Tables 2.6 and 2.7.

Table 2.6: Percentage of rejections to the null hypothesis of Bayesian projection posterior procedure (BP), linear regression procedure (LR), and piecewise linear fitting (PL) when the true regression functions are coordinatewise increasing.

	$n = 100$			$n = 200$			$n = 500$		
	BP	LR	PL	BP	LR	PL	BP	LR	PL
f_1	0	0	0	0	0	0	0	0	0
f_2	0.5	0	0.5	0	0	0	0	0	0
f_3	0	0	0	0	0	0	0	0	0
f_4	0.5	0	0	0	0	0	0	0	0
f_5	0	0	0.5	0	0	0	0	0	0
f_6	0	3	3	0	5	3.5	0	7	3

Table 2.7: Percentage of rejections to the null hypothesis of Bayesian projection posterior procedure (BP), linear regression procedure (LR), and piecewise linear fitting (PL) when the true regression functions are not coordinatewise increasing.

	$n = 100$			$n = 200$			$n = 500$		
	BP	LR	PL	BP	LR	PL	BP	LR	PL
f_7	64.5	8.5	73	93.5	10	99.5	100	10.5	100
f_8	100	84.5	83	100	98.5	100	100	100	100
f_9	35.5	7.6	28.5	96	94.5	66.5	100	100	100
f_{10}	100	100	100	100	100	100	100	100	100
f_{11}	100	100	98.5	100	100	100	100	100	100
f_{12}	35	0	55	89.5	0	94.5	100	0	100

We can see from Tables 2.6 and 2.7 that all three methods can control the Type I error rate of the test to a low level, even though the linear regression model is misspecified in case f_2 to f_5 . That is because the coefficients should be nonnegative when we project any coordinatewise nondecreasing function onto the linear function space. Noting that the null hypothesis is composite in the linear regression and the piecewise linear regression methods and the coefficients of the projected linear functions of f_2 to f_5 are all strictly greater than zero, it is thus reasonable that the results in table 2.6 looks conservative. However, in the case, f_6 , where the slope coefficients are zero and on the boundary of the null hypothesis, the Bonferroni adjustment seems not that conservative, giving an error rate very close to the nominal level even in the piecewise linear fitting where there are 18 slope coefficients to be tested. The nonparametric Bayesian test we proposed controls the Type I error at a very low level, especially when the sample size is moderately large. We can further adjust the value of M_n to make the type I error close to the nominal level 0.05 and thus a higher power would be expected. The nonparametric Bayesian method and the piecewise linear fitting method both have high power, as they can detect all kinds of violations to coordinatewise monotonicity, global or local, as the sample size increases. However, for some regression functions such as f_{12} , where there is a small bump in the middle of the function graph, the linear regression totally breaks down as it focuses on the global nature. The same conclusion also applies in the case f_7 .

The proposed methods enjoy power enhancement when the signal-to-noise ratio increases. We can see this by comparing cases f_8 and f_9 . From these two cases, we also notice that the proposed Bayesian method has a better capability of capturing the local violation than others.

2.5.4 Coverage comparison with Deng, Han and Zhang's method

For pointwise inference in multivariate isotonic regression, Deng et al. (2021) constructed the confidence interval by the asymptotic distribution of pivotal statistics. Their method will be referred to as DHZ in the following. Let $\hat{u}(x_0)$ and $\hat{v}(x_0)$ be such that

$$\begin{aligned}\hat{f}^-(x_0) &= \max_{u \leq x_0} \min_{\substack{v \geq x_0 \\ \#\{i: \mathbf{X}_i \in [u:v]\} > 0}} \bar{Y}|_{[u:v]} = \min_{\substack{v \geq x_0 \\ \#\{i: \mathbf{X}_i \in [\hat{u}(x_0):v]\} > 0}} \bar{Y}|_{[\hat{u}(x_0):v]}, \\ \hat{f}^+(x_0) &= \min_{v \geq x_0} \max_{\substack{u \leq x_0 \\ \#\{i: \mathbf{X}_i \in [u:v]\} > 0}} \bar{Y}|_{[u:v]} = \max_{\substack{u \leq x_0 \\ \#\{i: \mathbf{X}_i \in [u:\hat{v}(x_0)]\} > 0}} \bar{Y}|_{[u:\hat{v}(x_0)]},\end{aligned}$$

and $\hat{f}(x_0) = (\hat{f}^-(x_0) + \hat{f}^+(x_0))/2$. Under the same data generating conditions as in Theorem 2.4.2 and additionally assuming \mathbf{X} is uniform distributed, Deng et al. (2021) showed that

$$\frac{\sqrt{\#\{i: \mathbf{X}_i \in [\hat{u}(x_0): \hat{v}(x_0)]\}}}{\sigma} (\hat{f}(x_0) - f(x_0)) \rightsquigarrow K_\beta,$$

where K_β is a universal distribution that depends solely on the local regularity β . Let $1 - \gamma \in (0.5, 1)$ be the confidence level. They proposed the following confidence interval for $f_0(x_0)$:

$$\left[\hat{f}(x_0) - \frac{c_\gamma \hat{\sigma}}{\sqrt{\#\{i: \mathbf{X}_i \in [\hat{u}(x_0): \hat{v}(x_0)]\}}}, \hat{f}(x_0) + \frac{c_\gamma \hat{\sigma}}{\sqrt{\#\{i: \mathbf{X}_i \in [\hat{u}(x_0): \hat{v}(x_0)]\}}}, \right] \quad (2.37)$$

where c_γ is the critical value obtained by simulating the limiting distribution of K_β and $\hat{\sigma}$ is a consistent estimator of σ .

We consider five regression functions: (1) $f_1(x_1, x_2) = (x_1 + x_2)^2$; (2) $f_2(x_1, x_2) = \sqrt{x_1 + x_2}$; (3) $f_3(x_1, x_2) = x_1 x_2$; (4) $f_4(x_1, x_2) = e^{x_1 + x_2}$; (5) $f_5(x_1, x_2) = e^{x_1 x_2}$. Set $\varepsilon_i \sim N(0, 1)$ and $X_1, X_2 \sim \text{Unif}(0, 1)$, mutually independent, for $i = 1, \dots, n$. We consider sample sizes $n = 200, 500, 1000$,

and 2000. To construct credible intervals, we choose $J = \lceil n^{1/3} \log(\log n) \rceil$ according to the discussion in Remark 3. We compare the coverage and length of our immersion credible interval (IB), the recalibrated credible interval (IB(adj)), and the DHZ's confidence interval under two credible/confidence levels 0.95 and 0.90. The coverage percentage and the average length are calculated over 2000 replications. The result is summarized in Table 2.8.

The unadjusted credible intervals generally overcover the true function value for larger sample sizes, whereas the recalibrated credible intervals provide more accurate coverage to different extents for different functions. DHZ's method yields more precise coverage at the given confidence level when the sample sizes are relatively smaller. However, our credible intervals are generally shorter and exhibit less variation compared to DHZ's confidence intervals. The variation observed in our method across different regression functions may be attributed to the roughness of the partition used. In practical applications, a slightly larger J can be set, provided that the credible intervals can be computed within a reasonable time.

2.5.5 Length comparison with the oracle

To compare with the oracle confidence interval length; see Deng et al. (2021), we use the critical value $c = 1.85$ for confidence level 0.95 and calculate the oracle interval length by

$$2c(n/\sigma)^{-1/(2+d)} \left[\prod_{k=1}^d (\partial_k f(\mathbf{x}_0)/2) \right]^{1/(2+d)}.$$

We use the simulation results in Section 2.5.4 for our methods, IB and IB(adj), The results are presented in Figure 2.4, overlapped with the oracle confidence interval length. In most settings, the median of the credible interval length, adjusted or unadjusted, is smaller than the oracle confidence interval length. Moreover, our methods produce credible intervals significantly shorter than DHZ's confidence intervals using the data generated under the same condition from the simulation study in Section 2.5.4.

Table 2.8: Coverage percentage (C) and length (L) comparison,

f	level	n	IB		IB(adj)		DHZ	
			C	L	C	L	C	L
f_1	0.05	200	93.6	0.903(0.145)	90.0	0.805(0.132)	92.4	1.138(0.600)
		500	98.8	0.777(0.111)	97.7	0.692(0.101)	95.0	0.959(0.490)
		1000	97.0	0.630(0.086)	94.4	0.562(0.078)	94.8	0.827(0.435)
		2000	97.4	0.535(0.072)	95.4	0.476(0.066)	94.9	0.686(0.324)
	0.10	200	88.1	0.761(0.126)	81.5	0.668(0.112)	86.1	0.898(0.473)
		500	96.9	0.656(0.097)	94.4	0.576(0.087)	90.0	0.757(0.387)
		1000	92.8	0.532(0.075)	88.8	0.467(0.068)	88.8	0.652(0.343)
		2000	94.3	0.451(0.063)	90.4	0.397(0.057)	89.7	0.541(0.256)
f_2	0.05	200	91.0	0.503(0.089)	87.0	0.447(0.081)	95.2	0.722(0.339)
		500	96.6	0.380(0.061)	93.8	0.338(0.055)	95.4	0.546(0.303)
		1000	94.9	0.308(0.047)	91.7	0.274(0.043)	94.8	0.439(0.252)
		2000	95.9	0.253(0.039)	93.3	0.225(0.035)	95.3	0.357(0.175)
	0.10	200	85.0	0.423(0.077)	79.0	0.371(0.068)	89.8	0.570(0.268)
		500	92.7	0.320(0.052)	88.0	0.280(0.047)	90.3	0.431(0.239)
		1000	90.0	0.259(0.040)	84.7	0.227(0.036)	88.9	0.346(0.199)
		2000	91.8	0.213(0.033)	87.0	0.186 (0.030)	90.6	0.281(0.138)
f_3	0.05	200	91.8	0.476(0.084)	87.2	0.423(0.076)	94.7	0.740(0.410)
		500	96.0	0.371(0.061)	93.4	0.329(0.055)	95.0	0.532(0.246)
		1000	95.0	0.293(0.046)	91.6	0.260(0.042)	95.4	0.433(0.207)
		2000	95.6	0.242(0.037)	93.0	0.215(0.034)	94.8	0.353(0.165)
	0.10	200	84.9	0.400(0.072)	79.6	0.350(0.064)	89.4	0.584(0.323)
		500	92.2	0.311(0.053)	87.8	0.273(0.047)	89.7	0.419(0.194)
		1000	90.0	0.246(0.040)	84.7	0.216(0.036)	90.3	0.341(0.163)
		2000	91.6	0.204(0.033)	86.9	0.178(0.029)	89.8	0.279(0.131)
f_4	0.05	200	97.0	1.260(0.188)	94.4	1.122(0.170)	89.2	1.234(0.621)
		500	99.8	1.086(0.133)	99.5	0.968(0.120)	92.8	1.077(0.538)
		1000	99.0	0.869(0.100)	97.9	0.774(0.092)	94.4	0.927(0.437)
		2000	99.7	0.728(0.083)	98.2	0.649(0.075)	94.8	0.800(0.405)
	0.10	200	92.8	1.063(0.162)	87.9	0.932 (0.144)	83.4	0.974(0.490)
		500	99.3	0.917(0.115)	98.4	0.805(0.103)	87.0	0.850(0.424)
		1000	97.0	0.733(0.088)	93.8	0.644(0.079)	89.1	0.731(0.345)
		2000	97.2	0.615(0.072)	94.9	0.540(0.064)	89.3	0.631(0.320)
f_5	0.05	200	94.0	0.540(0.093)	90.2	0.480(0.083)	95.4	0.798(0.398)
		500	97.4	0.432(0.068)	95.6	0.384(0.062)	94.2	0.597(0.316)
		1000	96.5	0.338(0.051)	93.4	0.301(0.047)	95.1	0.491(0.265)
		2000	96.9	0.280(0.042)	94.4	0.249(0.038)	95.5	0.401(0.195)
	0.10	200	88.2	0.454(0.079)	82.6	0.398(0.070)	90.3	0.629(0.314)
		500	94.4	0.364(0.059)	90.6	0.319(0.053)	89.1	0.471(0.249)
		1000	92.2	0.285(0.045)	87.8	0.249(0.040)	91.1	0.387(0.209)
		2000	93.4	0.236(0.036)	89.2	0.207(0.033)	90.1	0.317(0.154)

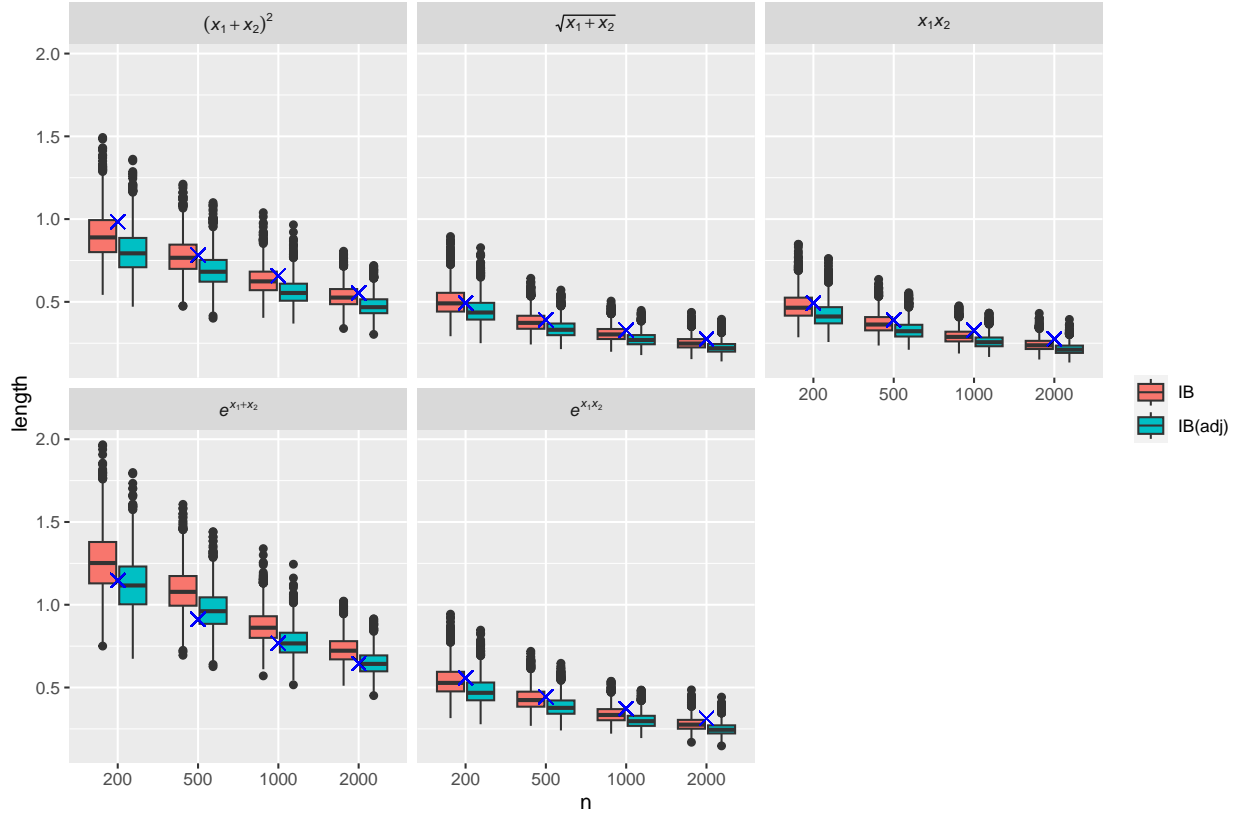


Figure 2.4: Credible interval length against sample size, grouped by regression functions marked in the subtitles. The blue cross sign marks the oracle confidence interval length under each setting.

2.5.6 Effects of the choice of J

We have conducted simulations to see the effect of J . In this simulation study, we choose two regression functions $f(x_1, x_2) = (x_1 + x_2)^2$ and $f(x_1, x_2) = e^{-x_1 x_2}$, and three different sample sizes, $n \in \{200, 500, 1000\}$. We construct credible intervals for $f(0.5, 0.5)$. Let σ^2 be known and set as 1. The prescribed credibility is set at 95%. Let J vary from 5 to 30. We generate 1000 posterior samples for each data set. For each combination of (f, n, J) , we replicate the immersion Bayes (IB) procedure and the adjusted immersion Bayes (IB(adj)) for 1000 times, respectively. The rest of the setup keeps the same as in the previous sections. We calculate the frequency of credible intervals including the true regression function value (Coverage). The results are plotted in Figure 2.5 and 2.6. To understand the simulation results, it is worthwhile to note that we choose

$x_0 = (0.5, 0.5)$, which lies in the middle of a certain piece when J is odd, while on the edge of a certain piece when J is even. This explains the wiggly pattern of the line plot. Despite being a very coarse approximation, the coverage remains highly precise even for small values of odd J . However, for even J , we have to increase its value to get precise coverage. It is worth noting that the choice of J is highly flexible and the results remain robust even when choosing a very large value of J in comparison to the relatively low sample size. Consider, for instance, when $n = 200$. Our methods exhibit desirable behavior when we select $J \geq 25$. This choice gives a partition of the set $[0, 1]^2$ consisting of $J^2 \geq 25^2$ pieces, a number significantly greater than the sample size. By noting the nice results for large J , the adaptation for less smooth regression functions will not be an issue in practice.

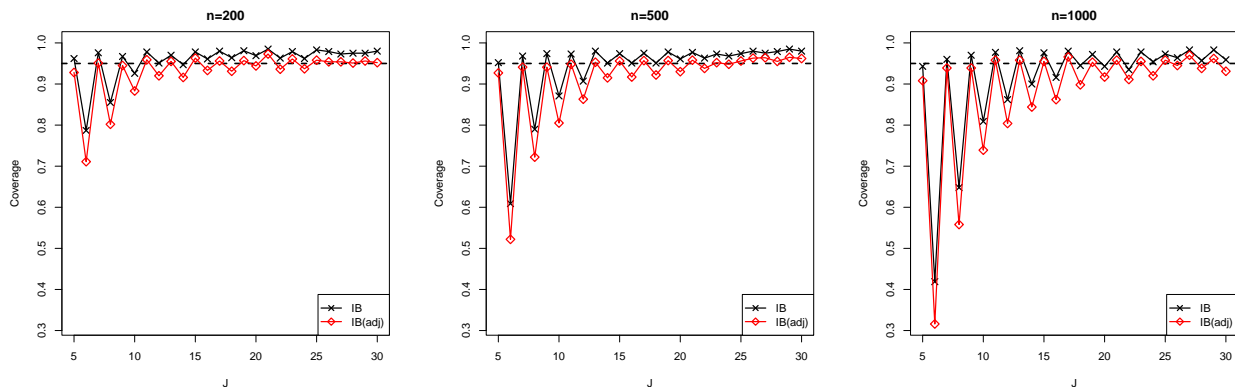


Figure 2.5: Coverage against J , $f(x_1, x_2) = (x_1 + x_2)^2$.

2.6 Proofs

2.6.1 Proofs of results in Section 2.2

Proof of Proposition 2.2.1. For a given h , let

$$\bar{h} = \sum_{j \in [1:J]} \lambda(I_j)^{-1} \int_{I_j} h d\lambda \cdot \mathbb{1}_{I_j}.$$

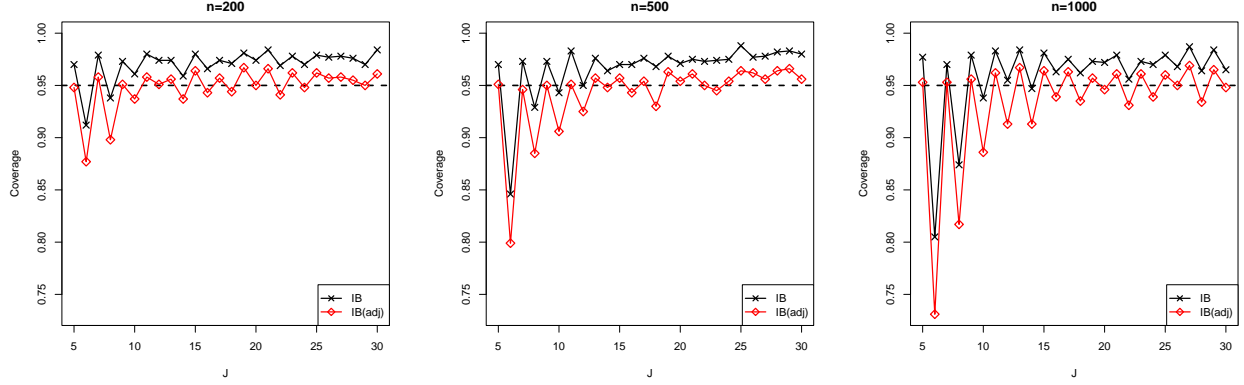


Figure 2.6: Coverage against J , $f(x_1, x_2) = e^{x_1 x_2}$.

Clearly, $\bar{h} \in \mathcal{M}$ if $h \in \mathcal{M}$. Since f is constant on I_j , for every $x \in I_j$,

$$\left| \frac{\int_{I_j} h d\lambda}{\lambda(I_j)} - f(x) \right|^p = \frac{\left| \int_{I_j} (h - f) d\lambda \right|^p}{\lambda(I_j)^p} \leq \frac{\int_{I_j} |h - f|^p d\lambda}{\lambda(I_j)}, \quad (2.38)$$

by Jensen's inequality. Taking integrals on both sides of (2.38) over I_j , it follows that

$$\int_{I_j} |\bar{h} - f|^p d\lambda \leq \int_{I_j} |h - f|^p d\lambda.$$

Hence the monotone projection of $f \in \mathcal{F}_J$ onto \mathcal{M} also belongs to \mathcal{F}_J . The existence of f^* is ensured by the convexity and the closedness of \mathcal{C} and the convexity of L_p -losses. □

Proof of Theorem 2.2.1. Since the posterior for σ is consistent, it is sufficient to condition on the value of σ lying in a small neighborhood of σ_0 , unless σ is known. Let

$$f_{0,J} = \sum_{j \in [1:J]} f_0(j/J) \mathbb{1}_{I_j}.$$

Then $f_{0,J} \in \mathcal{M}_J$. As f^* is the $\mathbb{L}_1(G^*)$ -projection of f onto \mathcal{M}_J and $f_0 \in \mathcal{M}$,

$$\begin{aligned}
& \|f^* - f_0\|_{1,G^*} \\
& \leq \|f^* - f\|_{1,G^*} + \|f - f_{0,J}\|_{1,G^*} + \|f_{0,J} - f_0\|_{1,G^*} \\
& \leq 2\|f - f_{0,J}\|_{1,G^*} + \|f_{0,J} - f_0\|_{1,G^*}.
\end{aligned} \tag{2.39}$$

By Lemma A.1.2, $\|f_{0,J} - f_0\|_{1,G^*} \lesssim J^{-1}$ as $G^*(I_j) \lesssim J^{-1}$ is assumed. Hence it suffices to bound $\|f - f_{0,J}\|_{1,G^*}$.

Without loss of generality, we assume that $N_j > 0$ for all j . Let $\bar{f}_{0,J} = \sum_{j \in [1:J]} \theta_{0,j} \mathbb{1}_{I_j}$, where $\theta_{0,j} = N_j^{-1} \sum_{i: \mathbf{X}_i \in I_j} f_0(\mathbf{X}_i)$. Then Lemma A.1.2 applied twice and the triangle inequality give $\|f_{0,J} - \bar{f}_{0,J}\|_{1,G^*} \lesssim J^{-1}$. Therefore it suffices to show that

$$\mathbb{E}_0 \Pi(\|f - \bar{f}_{0,J}\|_{1,G^*} > M_n \sqrt{J^d/n} | \mathbb{D}_n) \rightarrow 0.$$

Applying the Cauchy-Schwarz inequality first and then Markov's inequality,

$$\begin{aligned}
& \Pi(\|f - \bar{f}_{0,J}\|_{1,G^*} > M_n \sqrt{J^d/n} | \mathbb{D}_n, \sigma) \\
& = \Pi\left(\sum_{j \in [1:J]} G^*(I_j) |\theta_j - \theta_{0,j}| > M_n \sqrt{J^d/n} | \mathbb{D}_n, \sigma\right) \\
& \leq \Pi\left(\sum_{j \in [1:J]} G^*(I_j) |\theta_j - \theta_{0,j}|^2 > M_n^2 J^d/n | \mathbb{D}_n, \sigma\right) \\
& \leq M_n^{-2} J^{-d} \sum_{j \in [1:J]} n G^*(I_j) \mathbb{E}[(\theta_j - \theta_{0,j})^2 | \mathbb{D}_n, \sigma].
\end{aligned} \tag{2.40}$$

We decompose

$$\mathbb{E}[(\theta_j - \theta_{0,j})^2 | \mathbb{D}_n, \sigma] = \text{Var}(\theta_j | \mathbb{D}_n, \sigma) + (\mathbb{E}(\theta_j | \mathbb{D}_n, \sigma) - \theta_{0,j})^2. \tag{2.41}$$

We observe that

$$\sum_{j \in [1:J]} nG^*(I_j) \text{Var}(\theta_j | \mathbb{D}_n, \sigma) \leq \frac{\sigma^2}{\min\{1, \min_j \{\lambda_j^{-2}\}\}} \sum_{j \in [1:J]} \frac{nG^*(I_j)}{N_j + 1}. \quad (2.42)$$

From (2.4), we know

$$\begin{aligned} & \sum_{j \in [1:J]} nG^*(I_j) (\mathbb{E}(\theta_j | \mathbb{D}_n, \sigma) - \theta_{0,j})^2 \\ &= \sum_{j \in [1:J]} nG^*(I_j) \left(\frac{N_j \bar{\epsilon}|_{I_j} + \lambda_j^{-2} \zeta_j - \theta_{0,j} \lambda_j^{-2}}{N_j + \lambda_j^{-2}} \right)^2 \\ &\lesssim \sum_{j \in [1:J]} \frac{nG^*(I_j) N_j^2 (\bar{\epsilon}|_{I_j})^2}{(N_j + 1)^2} + \sum_{j \in [1:J]} \frac{nG^*(I_j)}{(N_j + 1)^2} \\ &\lesssim \sum_{j \in [1:J]} \frac{nG^*(I_j)}{N_j + 1} \end{aligned} \quad (2.43)$$

by noting that $\mathbb{E}[(\bar{\epsilon}|_{I_j})^2 | \mathbf{X}, \sigma] = \sigma^2 / N_j$. Hence, the expectations of the expressions in (2.42) and (2.43) are bounded by a constant multiple of J^d in view of (2.11). Combining these with (2.40) and (2.41), it follows that

$$\mathbb{P}(\|f - \bar{f}_{0,J}\|_{2,G^*} > M_n \sqrt{J^d/n} | \mathbb{D}_n, \sigma) \lesssim M_n^{-2}, \quad (2.44)$$

and hence the first part of the theorem is established.

If $\max\{N_j : j \in [1 : J]\} \lesssim n/J^d$, then Lemma A.1.3 ensures that the estimator and the posterior for σ are consistent. For $G^* = G_n$, the condition (2.11) holds because $nG_n(I_j)(N_j + 1)^{-1} \leq 1$. If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. with a bounded density g , then

$$\max\{N_j : j \in [1 : J]\} \lesssim n/J^d$$

by Lemma A.1.1, provided that $J^d(\log n)/n \rightarrow 0$. If $G^* = G_n$ for either random or deterministic predictors \mathbf{X}_i , (2.42) is bounded by J^d up to some positive constant. If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. G ,

then owing to $N_j \sim \text{Bin}(n, G(I_j))$, we have that

$$\mathbb{E}_0[(N_j + 1)^{-1}] = \begin{cases} \frac{1 - (1 - G(I_j))^{n+1}}{(n+1)G(I_j)}, & \text{if } G(I_j) > 0; \\ 1, & \text{if } G(I_j) = 0, \end{cases} \quad (2.45)$$

so that

$$nG(I_j)\mathbb{E}[(N_j + 1)^{-1}] \leq 1,$$

implying that (2.11) holds for $G^* = G$. This completes the proof of the second part of the theorem. \square

Proof of Corollary 2.2.2. For f^* the $\mathbb{L}_1(G_n)$ -projection, by the triangle inequality,

$$\|f^* - f_0\|_{1,G} \leq \|f^* - \bar{f}_{0,J}\|_{1,G} + \|\bar{f}_{0,J} - f_0\|_{1,G},$$

where, as in the proof of the last theorem, $\bar{f}_{0,J} = \sum_{j \in [1:J]} \theta_{0,j} \mathbb{1}_{I_j}$, with $\theta_{0,j} = N_j^{-1} \sum_{i: X_i \in I_j} f_0(\mathbf{X}_i)$.

From Lemma A.1.2, we know that $\|\bar{f}_{0,J} - f_0\|_{1,G} \lesssim J^{-1}$ under the assumption of bounded density.

As f^* is the $\mathbb{L}_1(G_n)$ -projection of f onto \mathcal{M}_J , from Theorem 2.2.1,

$$\mathbb{E}_0 \Pi(\|f^* - \bar{f}_{0,J}\|_{1,G_n} > M_n \epsilon_n | \mathbb{D}_n) \rightarrow 0, \quad (2.46)$$

since we also have $\|\bar{f}_{0,J} - f_0\|_{1,G_n} \lesssim J^{-1}$ by Lemmas A.1.1 and A.1.2. Thus it suffices to show that

$$\mathbb{E}_0 \Pi(|\|f^* - \bar{f}_{0,J}\|_{1,G} - \|f^* - \bar{f}_{0,J}\|_{1,G_n}| > M_n \epsilon_n | \mathbb{D}_n) \rightarrow 0. \quad (2.47)$$

Clearly, we have

$$\begin{aligned} & \left| \|f^* - \bar{f}_{0,J}\|_{1,G} - \|f^* - \bar{f}_{0,J}\|_{1,G_n} \right| \\ & \leq \sum_{j \in [1:J]} G_n(I_j) |\theta_j^* - \theta_{0,j}| \cdot \max_j |G(I_j)/G_n(I_j) - 1|. \end{aligned} \quad (2.48)$$

Under the additional condition on the lower bound for g , Lemma A.1.1 implies that the last factor is $O_{P_0}(1)$. Thus (2.48) is bounded by a constant multiple of $\|f^* - \tilde{f}_{0,J}\|_{1,G_n}$ on an event with P_0 -probability tending to 1. Then this claim follows from Theorem 2.2.1. As g is bounded and bounded away from 0, $\|f^* - f_0\|_{1,G} \asymp \|f^* - f_0\|_{1,\lambda}$, then the corollary follows. \square

2.6.2 Proofs of results in Section 2.3

Proof of Theorem 2.3.1. (i) Since $\rho(f, \mathcal{M}_J) \leq \|f - f_{0,J}\|_{1,G}$, the conclusion follows from Theorem 2.2.1.

(ii) By the definition of projection and the triangle inequality,

$$\rho(f, \mathcal{M}_J) \geq \|f_0 - f^*\|_{1,G} - \|f - f_0\|_{1,G} \geq \rho(f_0, \mathcal{M}_J) - \|f - f_0\|_{1,G}. \quad (2.49)$$

Thus by the triangle inequality,

$$\begin{aligned} \Pi(\rho(f, \mathcal{M}) \leq M_n n^{-1/(d+2)} | \mathbb{D}_n) \\ \leq \Pi(\|f - f_0\|_{1,G} \geq \rho(f_0, \mathcal{M}_J) - M_n n^{-1/(d+2)} | \mathbb{D}_n). \end{aligned} \quad (2.50)$$

Since $\rho(f_0, \mathcal{M}_J) \geq \rho(f_0, \mathcal{M})$ and the latter is a fixed positive constant, to conclude the proof, it suffices to show that the posterior for f is consistent at f_0 in the $\mathbb{L}_1(G)$ -metric. Let

$$\theta_{0,j} = \int_{I_j} f_0 dG / G(I_j)$$

and then $f_{0,J} = \sum_j \theta_{0,j} \mathbb{1}_{I_j}$. By the martingale convergence theorem, $\|f_0 - f_{0,J}\|_{1,G} \rightarrow 0$. Proceeding as in the proof of Theorem 2.2.1, we conclude that

$$E_0 \Pi(\|f - f_{0,J}\|_{1,G} > M_n \sqrt{J^d/n} | \mathbb{D}_n) \rightarrow 0, \quad (2.51)$$

so posterior consistency holds in terms of the $\mathbb{L}_1(G)$ -distance.

(iii) For $f_0 \in \mathcal{H}(\alpha, L)$, we have $\|f_0 - f_{0,J}\|_{1,G} \lesssim J^{-\alpha}$. Together with (2.51), which is valid even when f_0 is not fixed, it follows that the $\mathbb{L}_1(G)$ -posterior contraction rate at f_0 is

$$\max\{\sqrt{J^d/n}, J^{-\alpha}\} \asymp n^{-\alpha/(2+d)}$$

for the choice $J \asymp n^{1/(2+d)}$. For $\alpha < 1$, the expression on the right hand side of (2.50) is, for large n , bounded by

$$\Pi(\|f - f_0\|_{1,G} \geq C n^{-\alpha/(2+d)}/2) \rightarrow_{P_0} 0,$$

since $n^{-\alpha/(2+d)} \gg n^{-1/(2+d)}$. If $\alpha = 1$, the corresponding bound for the event of interest reduces to

$$\Pi(\|f - f_0\|_{1,G} \geq (C-1)M_n n^{-1/(d+2)}/2 | \mathbb{D}_n) \rightarrow_{P_0} 0.$$

□

Proof of Theorem 2.3.2. With $p_{f,\sigma}$ defined by (2.13), the Hellinger distance between $p_{f_1,\sigma}$ and $p_{f_2,\sigma}$ is $\rho(f_1, f_2)$ and the Kullback-Leibler divergences are given by

$$K(p_{f_0,\sigma}; p_{f,\sigma}) = \frac{1}{2\sigma^2} \|f - f_0\|_{2,G}^2, \quad V(p_{f_0,\sigma}; p_{f,\sigma}) = \frac{1}{\sigma^2} \|f - f_0\|_{2,G}^2.$$

Thus the Kullback-Leibler ball

$$\{f : K(p_{f_0,\sigma}; p_{f,\sigma}) \leq \epsilon^2, V(p_{f_0,\sigma}; p_{f,\sigma}) \leq \epsilon^2\}$$

contains the $\mathbb{L}_2(G)$ -ball $\{f : \|f - f_0\|_{2,G} \leq C\epsilon\}$ for some $C > 0$, and hence to study posterior contraction at a true f_0 , it suffices to lower bound the prior probability of the latter. Since

$$\|f_0 - f_{0,J}\|_{2,G}^2 \leq (f_0(\mathbf{1}) - f_0(\mathbf{0})) \|f_0 - f_{0,J}\|_{1,G},$$

to keep $\|f_0 - f_{0,J}\|_{2,G}$ within a targeted ϵ (which may or may not depend on n), J should be sufficiently large to make $\|f_0 - f_{0,J}\|_{1,G} \leq c\epsilon^2$ for some sufficiently small $c > 0$. If a value \bar{J} , possibly depending on n , achieves this, then using (2.6), we can lower bound the required $\mathbb{L}_2(G)$ -prior concentration by

$$\begin{aligned} & \Pi(\bar{J})\Pi(\|f - f_{0,\bar{J}}\|_{2,G} \leq C\epsilon | J = \bar{J}) \\ & \geq \Pi(\bar{J})\Pi(\cap_{j=1}^{\bar{J}} \{|\theta_j - \theta_{0,j}| \leq C_1\epsilon^2\}) \\ & \gtrsim \exp\{-b_2\bar{J}^d \log \bar{J} - C_2\bar{J}^d \log(1/\epsilon)\} \end{aligned}$$

for some constant $C_1, C_2 > 0$. Let J_n stand for a sufficiently large multiple of $(n\epsilon^2)^{1/d}$. There are two situations to be considered. If $\epsilon > 0$ is fixed at an arbitrarily small number, then \bar{J} may be chosen as a sufficiently large constant. Then the lower bound for prior concentration in ϵ -neighborhood is a fixed positive number. Hence it follows that

$$\Pi(J \geq J_n)/\Pi(\|f - f_0\|_{2,G} \leq C\epsilon) = o(e^{-2n\epsilon^2}), \quad (2.52)$$

and hence by Theorem 8.20 of Ghosal and van der Vaart (2017), $\Pi(J > J_n | \mathbb{D}_n) \rightarrow_{P_0} 0$. If $\epsilon = \epsilon_n \rightarrow 0$ is chosen so that $n\epsilon_n^2 \rightarrow \infty$ and the corresponding $\bar{J} = \bar{J}_n$ satisfies $\log \bar{J}_n \lesssim \log n$, and it holds that $\log(1/\epsilon_n) \lesssim \log n$ and $\bar{J}_n^d \log n \lesssim n\epsilon_n^2$, then for the choice $J_n = L(n\epsilon_n^2/\log n)^{1/d}$ for some sufficiently large constant $L > 0$, it again follows that

$$\Pi(J \geq J_n)/\Pi(\|f - f_0\|_{2,G} \leq C\epsilon_n) = o(e^{-2n\epsilon_n^2}).$$

Hence by Theorem 8.20 of Ghosal and van der Vaart (2017) again, $\Pi(J > J_n | \mathbb{D}_n) \rightarrow_{P_0} 0$.

First, we establish an auxiliary estimate essential to prove assertions (i), (ii), and (iii). We claim that for any bounded measurable f_0 (not necessarily monotone or smooth) and a given

$\delta > 0$, if $\log J_n \lesssim \log n$, there exists a sufficiently large constant $M_0 > 0$ such that

$$\mathbb{E}_0 \Pi(\|f - f_{0,J}\|_{2,G} \geq M_0 \sqrt{J^d(\log n)/n}, J \leq J_n | \mathbb{D}_n) < \delta, \quad (2.53)$$

when n is large enough. The posterior probability in the expectation of the last display can be written as

$$\sum_{J=1}^{J_n} \Pi(J | \mathbb{D}_n) \Pi\left(\sum_{j \in [1:J]} (\theta_j - \theta_{0,j})^2 G(I_j) \geq M_0^2 J^d (\log n)/n \mid \mathbb{D}_n\right). \quad (2.54)$$

By Markov's inequality and Assumption 3,

$$\begin{aligned} \max_{J \leq J_n} \Pi\left(\sum_{j \in [1:J_n]} (\theta_j - \theta_{0,j})^2 G(I_j) \geq M_0^2 J^d (\log n)/n \mid \mathbb{D}_n\right) \\ \leq \max_{J \leq J_n} \frac{n}{M_0^2 J^d \log n} \sum_{j \in [1:J_n]} G(I_j) [\text{Var}(\theta_j | \mathbb{D}_n) + (\mathbb{E}(\theta_j | \mathbb{D}_n) - \theta_{0,j})^2] \end{aligned}$$

which is bounded in probability by a constant multiple of

$$\max_{J \leq J_n} \frac{n}{M_0^2 J^d \log n} \sum_{j \in [1:J]} G(I_j) [(N_j + \lambda_j^{-2})^{-1} + (\bar{Y}|_{I_j} - \theta_{0,j})^2] \quad (2.55)$$

It is clear that $G(I_j) \asymp J^{-d}$. By Lemma A.1.1,

$$\mathbb{P}_0\left(\bigcap_{J=1}^{J_n} \{C_1 n/J^d \leq \min_j N_j \leq \max_j N_j \leq C_2 n/J^d\}\right) \rightarrow 1,$$

provided $n/J_n \gg \log J_n$, for two constant C_1 and $C_2 > 0$. Then $N_j \asymp n/J^d$ uniformly for all $j \leq J$ and J . By the union bound of sub-Gaussian variables (see citevan1996weak, Section 2.2), we have $(\bar{\varepsilon}|_{I_j})^2 \lesssim (J^d \log n)/n$ with arbitrarily high probability, provided $\log J_n \lesssim \log n$. As f_0 is bounded, we have

$$|N_j^{-1} \sum_{i: \mathbf{X}_i \in I_j} f_0(\mathbf{X}_i) - \theta_{0,j}|^2 \lesssim J^d (\log n)/n$$

uniformly for all j and J with high probability. Thus we establish the claim in (2.53).

To prove (i), we observe that the $\mathbb{L}_2(G)$ -approximation rate is $J^{-1/2}$, and thus

$$\epsilon_n \asymp \bar{J}_n^{-1/2} \asymp (n/\log n)^{-1/2(d+1)},$$

so $J_n \asymp (n/\log n)^{1/(d+1)}$, and $\Pi(J > J_n | \mathbb{D}_n) \rightarrow_{\mathbb{P}_0} 0$. Since

$$\rho(f, \mathcal{M}_J) \lesssim \rho(f, \mathcal{M}) \leq \rho(f, f_0),$$

the claim follows from (2.53).

To prove (ii), we choose $\epsilon > 0$ arbitrarily small but fixed. By the martingale convergence theorem, $\|f_0 - f_{0, J_0}\|_{1, G} < \epsilon$ for any sufficiently large J_0 . Hence J_n can be chosen a sufficiently small multiple of $(n/\log n)^{1/d}$ to satisfy (2.52), and consequently, $\Pi(J > J_n | \mathbb{D}_n) \rightarrow_{\mathbb{P}_0} 0$. Let

$$\mathcal{F}_n^* = \bigcup_{J=1}^{J_n} \left\{ \sum_{j \in [1, J_n]} \theta_j \mathbb{1}_{I_j} : |\theta_j| \leq n \right\}.$$

Then $\Pi(f \notin \mathcal{F}_n^*) = o(e^{-cn})$ for some constant $c > 0$, and the $\mathbb{L}_1(G)$ -covering number of \mathcal{F}_n^* is bounded by $J_n^d (2n/\epsilon)^{J_n^d}$. Thus the ϵ -metric entropy is bounded by $J_n^d \log n \leq n\epsilon^2$. Hence the posterior distribution at f_0 is consistent with respect to the $\mathbb{L}_1(G)$ -metric, by an application of the Schwartz posterior consistency theorem (cf., Theorem 6.23 of Ghosal and van der Vaart (2017)). Therefore, as $\rho(f_0, \mathcal{M}_J)$ is bounded by a positive fixed constant from below, by (2.49), it follows that

$$\Pi(\rho(f, \mathcal{M}_J) \leq M_0 \sqrt{(J^d \log n)/n} | \mathbb{D}_n) \rightarrow_{\mathbb{P}_0} 0.$$

To prove Part (iii), we observe by Lemma A.1.2 that the approximation rate at an $f_0 \in \mathcal{H}(\alpha, L)$ is $J^{-\alpha}$, so that $\bar{J}_n \asymp \epsilon_n^{-1/\alpha}$ and $\epsilon_n \asymp (n/\log n)^{-\alpha/(2\alpha+d)}$ and $J_n \asymp (n/\log n)^{1/(2\alpha+d)}$. Using the sieve \mathcal{F}_n^* as defined above with this choice of J_n , it follows that $\Pi(f \notin \mathcal{F}_n^*) = o(e^{-Cn\epsilon_n^2})$ for a given constant $C > 0$. The ϵ_n -metric entropy is bounded by $J_n^d \log n \lesssim n\epsilon_n^2$. Hence it follows from Theorem 8.9 of Ghosal and van der Vaart (2017) that the $\mathbb{L}_1(G)$ -posterior contraction rate is $(n/\log n)^{-\alpha/(2\alpha+d)}$.

Thus, as $\rho(f_0, \mathcal{M}_J) \geq C(n/\log n)^{-\alpha/(2\alpha+d)}$ for a sufficiently large constant $C > 0$, from (2.49) and the probabilistic bound $(n/\log n)^{1/(2\alpha+d)}$ for J , the conclusion follows. \square

2.6.3 Proofs of results in Section 2.4

Proof of Theorem 2.4.1

For $t \in \mathbb{R}^d$, let $j(t) = \lceil (x_0 + t \circ r_n) \circ J \rceil$. Let

$$f_{*,c}(\mathbf{x}_0) = \max_{\substack{c^{-\gamma} \mathbf{1} \leq \mathbf{u} \leq c \mathbf{1}, \\ u_k \leq x_{0,k}, \\ s+1 \leq k \leq d}} \min_{\substack{c^{-\gamma} \mathbf{1} \leq \mathbf{v} \leq c \mathbf{1}, \\ v_k \leq 1 - x_{0,k}, \\ s+1 \leq k \leq d}} \frac{\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j \theta_j}{N_{[j(-\mathbf{u}):j(\mathbf{v})]}}, \quad (2.56)$$

where γ is a positive constant to be determined later. We also introduce the notations

$$W_n^* = \omega_n^{-1}(f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)),$$

$$W_{n,c}^* = \omega_n^{-1}(f_{*,c}(\mathbf{x}_0) - f_0(\mathbf{x}_0)),$$

and

$$\begin{aligned} W_c &= \sup_{\substack{c^{-\gamma} \mathbf{1} \leq \mathbf{u} \leq c \mathbf{1}, \\ u_k \leq x_{0,k}, \\ s+1 \leq k \leq d}} \inf_{\substack{c^{-\gamma} \mathbf{1} \leq \mathbf{v} \leq c \mathbf{1}, \\ v_k \leq 1 - x_{0,k}, \\ s+1 \leq k \leq d}} \left\{ \frac{\sigma_0 H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \\ &\quad \left. + \frac{\sigma_0 H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} + \sum_{l \in L^*} \frac{\partial^l f_0(x_0)}{(l+1)!} \prod_{k=1}^s \frac{v_k^{l_{k+1}} - (-u_k)^{l_{k+1}}}{u_k + v_k} \right\}, \\ W &= \sup_{\substack{\mathbf{u} \geq \mathbf{0}, \\ u_k \leq x_{0,k}, \\ s+1 \leq k \leq d}} \inf_{\substack{\mathbf{v} \geq \mathbf{0}, \\ v_k \leq 1 - x_{0,k}, \\ s+1 \leq k \leq d}} \left\{ \frac{\sigma_0 H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \\ &\quad \left. + \frac{\sigma_0 H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} + \sum_{l \in L^*} \frac{\partial^l f_0(x_0)}{(l+1)!} \prod_{k=1}^s \frac{v_k^{l_{k+1}} - (-u_k)^{l_{k+1}}}{u_k + v_k} \right\}. \end{aligned}$$

The proof of the theorem is carried out in several steps using Lemma A.2.1, presented as lemmas below.

Lemma 2.6.1. *Under the conditions of Theorem 2.4.1, for every $c > 0$ and $\gamma > 0$, $\mathcal{L}(W_{n,c}^* | \mathbb{D}_n)$ converges weakly to $\mathcal{L}(W_c | H_1)$ as random probability measures.*

Proof. For every $\mathbf{u}, \mathbf{v} \succeq \mathbf{0}$, we can write

$$\frac{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j \theta_j}{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j} - f_0(\mathbf{x}_0) = A_n(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) + A'_n(\mathbf{u}, \mathbf{v}) + B_n(\mathbf{u}, \mathbf{v}), \quad (2.57)$$

and then

$$W_{n,c}^* = \max_{c^{-\gamma} \mathbf{1} \leq \mathbf{u} \leq c \mathbf{1}} \min_{c^{-\gamma} \mathbf{1} \leq \mathbf{v} \leq c \mathbf{1}} \{A_n(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) + A'_n(\mathbf{u}, \mathbf{v}) + B_n(\mathbf{u}, \mathbf{v})\},$$

where

$$A_n(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) = \omega_n^{-1} \frac{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j (\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n])}{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j}, \quad (2.58)$$

$$A'_n(\mathbf{u}, \mathbf{v}) = \omega_n^{-1} \frac{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j (\mathbb{E}[\theta_j | \mathbb{D}_n] - \bar{Y}_{I_j})}{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j}, \quad (2.59)$$

$$B_n(\mathbf{u}, \mathbf{v}) = \omega_n^{-1} (\bar{Y}_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}} - f_0(\mathbf{x}_0)). \quad (2.60)$$

Since the max-min functional is continuous on the space $\mathbb{L}_\infty([c^{-\gamma} \mathbf{1}, c \mathbf{1}] \times [c^{-\gamma} \mathbf{1}, c \mathbf{1}])$, it suffices to show that $A_n + A'_n + B_n$ converges weakly in $\mathbb{L}_\infty([c^{-\gamma} \mathbf{1}, c \mathbf{1}] \times [c^{-\gamma} \mathbf{1}, c \mathbf{1}])$, conditional on the data \mathbb{D}_n . By Lemma A.2.2 and Lemma 2.6.2, we prove the weak convergence of A_n . We show that A'_n converges to zero uniformly in Lemma 2.6.3. The convergence of B_n is completed by combining Lemma A.2.2, Lemma 2.6.4 and Lemma 2.6.5. \square

Lemma 2.6.2. *Under the conditions of Theorem 2.4.1, for every $c > 0$, let*

$$\mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) = \omega_n \sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j (\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n]).$$

Then $\mathbb{H}_{2,n}$ converges weakly to a centered Gaussian process H_2 in $\mathbb{L}_\infty([0, c \mathbf{1}] \times [0, c \mathbf{1}])$ for every $c > 0$ in \mathbb{P}_0 -probability.

Proof. By (2.4), Lemmas A.2.2, A.2.4, and A.2.5, the covariance kernel of $\mathbb{H}_{2,n}$ given $(\mathbb{D}_n, \sigma_n^2)$, is

given by

$$\omega_n^2 \sigma_n^2 \sum_{j \in [j(-\mathbf{u} \wedge \mathbf{u}'), j(\mathbf{v} \wedge \mathbf{v}')] } N_j^2 / (N_j + \lambda_j^{-2}),$$

which converges in P_0 -probability to

$$\sigma_0^2 \prod_{k=1}^s (u_k \wedge u'_k + v_k \wedge v'_k) D_s(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}').$$

Thus finite-dimensional distributions of $\mathbb{H}_{2,n}$ converge weakly to those of a centered Gaussian process $\sigma_0 H_2$ in P_0 -probability.

Next we need to show that $\mathcal{L}(\mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) : (\mathbf{u}, \mathbf{v}) \in [0, c \mathbf{1}] \times [0, c \mathbf{1}])$ is tight on $\mathbb{L}_\infty([0, c \mathbf{1}] \times [0, c \mathbf{1}])$ for any $c > 0$ in P_0 -probability. In view of Theorem 18.14 of van der Vaart (1998), we need to verify that, for every $\epsilon > 0$ and $\eta > 0$, there exists a finite partition $\{T_p : p \leq K\}$ of $[0, c \mathbf{1}] \times [0, c \mathbf{1}]$ with K depending only on ϵ and η such that

$$P\left(\sup_{(\mathbf{u}_1, \mathbf{v}_1), (\mathbf{u}_2, \mathbf{v}_2) \in T_p} \{|\mathbb{H}_{2,n}(\mathbf{u}_1, \mathbf{v}_1) - \mathbb{H}_{2,n}(\mathbf{u}_2, \mathbf{v}_2)| : 1 \leq p \leq K\} > \epsilon | \mathbb{D}_n\right) < \eta$$

with P_0 -probability tending to 1. Let $\delta > 0$, to be determined later, which depends only on ϵ and η . Let $0 = s_0 < s_1 < \dots < s_l = c$ with $(s_{r-1}, s_r]$ of equal length at least δ and $l \leq 2c/\delta$. We choose a partition $\{T_p : p \leq K\}$ of $[0, c \mathbf{1}] \times [0, c \mathbf{1}]$ to be

$$\mathcal{P}(\boldsymbol{\delta}) = \left\{ \prod_{k=1}^d (s_{t_k-1}, s_{t_k}] \times \prod_{k=1}^d (s_{r_k-1}, s_{r_k}] : t_k, r_k \in \{1, \dots, l\} \right\}, \quad (2.61)$$

with cardinality $K = \#\mathcal{P}(\boldsymbol{\delta}) = l^{2d}$. It suffices to verify that, for any $p \leq K$,

$$P\left(\sup_{(\mathbf{u}_1, \mathbf{v}_1), (\mathbf{u}_2, \mathbf{v}_2) \in T_p} \{|\mathbb{H}_{2,n}(\mathbf{u}_1, \mathbf{v}_1) - \mathbb{H}_{2,n}(\mathbf{u}_2, \mathbf{v}_2)|\} > \epsilon | \mathbb{D}_n\right) < \eta \left(\frac{\delta}{2c}\right)^{2d}.$$

Let $\mathcal{J}(\mathbf{u}, \mathbf{v}) = [j(-\mathbf{u}) : j(\mathbf{v})]$. For $(\mathbf{u}_1, \mathbf{v}_1), (\mathbf{u}_2, \mathbf{v}_2)$, we write $\mathbb{H}_{2,n}(\mathbf{u}_1, \mathbf{v}_1) - \mathbb{H}_{2,n}(\mathbf{u}_2, \mathbf{v}_2)$ as the difference of the sums of $\omega_n N_j(\theta_j - E[\theta_j | \mathbb{D}_n])$ over the sets $\mathcal{J}(\mathbf{u}_1, \mathbf{v}_1) \setminus \mathcal{J}(\mathbf{u}_1 \wedge \mathbf{u}_2, \mathbf{v}_1 \wedge \mathbf{v}_2)$ and $\mathcal{J}(\mathbf{u}_2, \mathbf{v}_2) \setminus \mathcal{J}(\mathbf{u}_1 \wedge \mathbf{u}_2, \mathbf{v}_1 \wedge \mathbf{v}_2)$, after canceling out the common terms. Thus its absolute value

can be bounded by the sum of the corresponding absolute values over these two index sets. To verify tightness, it then suffices to show that

$$\mathbb{P}\left(\max\{\omega_n \mid \sum_{\mathcal{J}(\mathbf{u}, \mathbf{v}) \setminus \mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1})} N_j(\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n])\} : (\mathbf{u}, \mathbf{v}) \in T_p\} > \frac{\epsilon}{4} \mid \mathbb{D}_n\right)$$

is bounded by $\eta(\delta/(2c))^{2d}/4$, with

$$T_p = \prod_{k=1}^d (s_{t_{k-1}}, s_{t_k}] \times \prod_{k=1}^d (s_{r_{k-1}}, s_{r_k}],$$

for any $\mathbf{s}_t = (s_{t_1}, \dots, s_{t_d})$ and $\mathbf{s}_r = (s_{r_1}, \dots, s_{r_d})$.

Let

$$S_{-j(-\mathbf{u}), j(\mathbf{v})} = \sum_{\mathcal{J}(\mathbf{u}, \mathbf{v}) \setminus \mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1})} N_j(\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n]),$$

a collection of random variables indexed by a $2d$ -dimensional vector in a finite index set. The negative sign in front of $j(-\mathbf{u})$ in the subscript of S is to make the σ -fields

$$\mathcal{F}_j^{(k)} = \begin{cases} \sigma\langle N_j(\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n]) : -(j(\mathbf{s}_{t-1}))_k < -(j(-\mathbf{u}))_k \leq j \rangle, & \text{if } k \leq d, \\ \sigma\langle N_j(\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n]) : (j(\mathbf{s}_{r-1}))_{k-d} < (j(\mathbf{v}))_{k-d} \leq j \rangle, & \text{if } k > d. \end{cases}$$

increase with respect to each of the first d components in the subscript. In the sum above, all j are in $\mathcal{J}(\mathbf{s}_t, \mathbf{s}_r) \setminus \mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1})$. We note that for every $k \leq 2d$, the random sequence $\{S_{(j_1, \dots, j_{k-1}, j, j_{k+1}, \dots, j_{2d})}, \mathcal{F}_j^{(k)}\}$ is a martingale. Applying Lemma A.2.6 with $p = 4d + 2$, we can get an upper bound of the probability of the maximal deviation needed to verify tightness to be a constant multiple of

$$(\omega_n/\epsilon)^{(4d+2)} \mathbb{E}\left(\left| \sum_{\mathcal{J}(\mathbf{s}_t, \mathbf{s}_r) \setminus \mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1})} N_j(\theta_j - \mathbb{E}[\theta_j | \mathbb{D}_n]) \right|^{4d+2} \mid \mathbb{D}_n\right). \quad (2.62)$$

Observe that

$$\begin{aligned}\#\mathcal{J}(\mathbf{s}_t, \mathbf{s}_r) &\leq \prod_k (r_{n,k} J_k(s_{t_k} + s_{r_k}) + 2), \\ \#\mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1}) &\geq \prod_k r_{n,k} J_k(s_{t_k} + s_{r_k} - 2\delta).\end{aligned}$$

As $\delta \leq s_{t_k}, s_{r_k} \leq c$ and $J_k \gg r_{n,k}^{-1}$, it follows that the cardinality of the index set $\mathcal{J}(\mathbf{s}_t, \mathbf{s}_r) \setminus \mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1})$ is bounded by a multiple of

$$\prod_{k=1}^d r_{n,k} J_k \left(\prod_{k=1}^d (s_{t_k} + s_{r_k}) - \prod_{k=1}^d (s_{t_k} + s_{r_k} - 2\delta) \right) \leq (2d\delta)(2c)^{d-1} \prod_{k=1}^d r_{n,k} J_k,$$

where the last inequality follows from Lemma A.2.7.

The variance

$$\sigma_n^2 \frac{N_j^2}{(N_j + \lambda_j^{-2})} \lesssim \frac{n}{\prod_{k=1}^d J_k}$$

with P_0 -probability tending to 1 by Lemma A.2.4. Hence (2.62) is bounded by a constant multiple of

$$\begin{aligned}&\epsilon^{-(4d+2)} \omega_n^{4d+2} \left(\frac{n \#(\mathcal{J}(\mathbf{s}_t, \mathbf{s}_r) \setminus \mathcal{J}(\mathbf{s}_{t-1}, \mathbf{s}_{r-1}))}{\prod_{k=1}^d J_k} \right)^{2d+1} \\ &\lesssim \epsilon^{-(4d+2)} \omega_n^{4d+2} \left(\prod_{k=1}^d r_{n,k} \right)^{2d+1} n^{2d+1} \delta^{2d+1},\end{aligned}$$

which simplifies to $\epsilon^{-(4d+2)} \delta^{2d+1}$. With δ chosen a sufficiently small constant multiple of $\eta \epsilon^{4d+2}$, the tightness condition is verified. \square

Lemma 2.6.3. *Under the conditions of Theorem 2.4.1, $A'_n(\mathbf{u}, \mathbf{v})$ converges to 0 in P_0 -probability uniformly in (\mathbf{u}, \mathbf{v}) .*

Proof. Let

$$E_n = \{a_1 n / (2 \prod_{k=1}^d J_k) \leq N_j \leq 2a_2 n / (\prod_{k=1}^d J_k)\}$$

for some $a_1, a_2 > 0$ and $\bar{\varepsilon}|_{I_j} = \sum_{i \in I_j} \varepsilon_i / N_j$. By Lemma A.2.4, we have, for every $T > 0$,

$$P_0(\max_j |\bar{\varepsilon}|_{I_j}| > T) \leq \sum_j P_0(|\bar{\varepsilon}|_{I_j}| > T | E_n) + P_0(E_n^c). \quad (2.63)$$

By Assumption 5 and Marcinkiewicz–Zygmund inequality,

$$E(|\bar{\varepsilon}|_{I_j}|^{2(\sum_{k=1}^s \beta_k^{-1} + 1)} | E_n) \lesssim (a_1 n / (2 \prod_{k=1}^d J_k))^{-(\sum_{k=1}^s \beta_k^{-1} + 1)}.$$

Then (2.63) is bounded by a constant multiple of

$$\left(\prod_{k=1}^d J_k \right)^{\sum_{k=1}^s \beta_k^{-1} + 2} n^{-(\sum_{k=1}^s \beta_k^{-1} + 1)} + o(1),$$

which tends to zero because

$$\prod_{k=1}^d J_k \ll n \omega_n = n^{(\sum_{k=1}^s \beta_k^{-1} + 1) / (\sum_{k=1}^s \beta_k^{-1} + 2)}.$$

On the other hand, $\max_j |\overline{f_0(\mathbf{X}_i)}|_{I_j} \leq f_0(\mathbf{1})$. Thus $\max_j |\bar{Y}|_{I_j} = O_{P_0}(1)$. Because

$$E[\theta_j | \mathbb{D}_n] = (N_j \bar{Y}|_{I_j} + \zeta_j \lambda_j^{-2}) / (N_j + \lambda_j^{-2}),$$

on the event E_n ,

$$|A'_n(\mathbf{u}, \mathbf{v})| \quad (2.64)$$

$$= \omega_n^{-1} \left| \frac{\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} \lambda_j^{-2} N_j (N_j + \lambda_j^{-2})^{-1} (\zeta_j - \bar{Y}|_{I_j})}{\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j} \right|$$

$$\lesssim \omega_n^{-1} (\max_j |\bar{Y}|_{I_j} + \zeta_j) (\min_j N_j)^{-1}, \quad (2.65)$$

which is of the order of $(n \omega_n)^{-1} \prod_{k=1}^d J_k$ in P_0 -probability. As $\prod_{k=1}^d J_k \ll n \omega_n$ and $P_0(E_n) \rightarrow 1$, we can conclude $A'(\mathbf{u}, \mathbf{v}) \rightarrow_{P_0} 0$ uniformly for any $\mathbf{u} \geq \mathbf{0}$ and $\mathbf{v} \geq \mathbf{0}$ provided that $\mathbf{x}_0 - \mathbf{u} \circ \mathbf{r}_n$ and

$\mathbf{x}_0 + \mathbf{v} \circ \mathbf{r}_n$ in $[0, 1]^d$.

□

To establish the weak convergence of B_n in $\mathbb{L}_\infty([0, c \mathbf{1}] \times [0, c \mathbf{1}])$, write

$$B_n(\mathbf{u}, \mathbf{v}) = \omega_n^{-1} (\bar{\varepsilon}|_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}} + \overline{f_0(\mathbf{X})}|_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}} - f_0(\mathbf{x}_0)). \quad (2.66)$$

Lemma 2.6.4. *Let*

$$Z_{ni}(\mathbf{u}, \mathbf{v}) = \omega_n \varepsilon_i \mathbb{1}_{\{\mathbf{X}_i \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}\}},$$

and

$$\mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n Z_{ni}(\mathbf{u}, \mathbf{v}).$$

Under the conditions of Theorem 2.4.1, $\mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v}) \rightsquigarrow \sigma_0 H_1(\mathbf{u}, \mathbf{v})$ in $\mathbb{L}_\infty([0, c \mathbf{1}] \times [0, c \mathbf{1}])$.

Proof. We first verify the finite-dimensional convergence. For any pair (\mathbf{u}, \mathbf{v}) and $(\mathbf{u}', \mathbf{v}')$, as \mathbf{X} and ε are mutually independent, the covariance of $\mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v})$ and $\mathbb{H}_{1,n}(\mathbf{u}', \mathbf{v}')$ is

$$n \omega_n^2 \mathbb{E}(\varepsilon^2 \mathbb{1}_{\{\mathbf{X} \in I_{[j(-\mathbf{u} \wedge \mathbf{u}'); j(\mathbf{v} \wedge \mathbf{v}')]\}}}) = \sigma_0^2 n \omega_n^2 \int_{I_{[j(-\mathbf{u} \wedge \mathbf{u}'); j(\mathbf{v} \wedge \mathbf{v}')]}} g(\mathbf{x}) d\mathbf{x}.$$

Write $g(\mathbf{x}) = g(\mathbf{x}_s) + (g(\mathbf{x}) - g(\mathbf{x}_s))$ where $\mathbf{x}_s = (x_{0,1}, \dots, x_{0,s}, x_{s+1}, \dots, x_d)$. By the continuity of g around \mathbf{x}_0 , the last display is reduced to

$$\sigma_0^2 n \omega_n^2 \prod_{k=1}^s ((u_k \wedge u'_k + v_k \wedge v'_k) r_{n,k} + O(j_k^{-1})) D_s^J(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}')(1 + o(1)).$$

From the proof of Lemma A.2.2, $D_s^J(\mathbf{u}, \mathbf{v}) \rightarrow D_s(\mathbf{u}, \mathbf{v})$. Since $\mathbf{u}, \mathbf{v}, \mathbf{u}', \mathbf{v}'$ are bounded and $J_k \gg r_{n,k}^{-1}$, it follows that the limit of the expression in the last display converges to

$$\sigma_0^2 \prod_{k=1}^s (u_k \wedge u'_k + v_k \wedge v'_k) D_s(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}').$$

To establish the asymptotic tightness of \mathbb{H}_1 in $\mathbb{L}_\infty([0, c \mathbf{1}] \times [0, c \mathbf{1}])$, we apply Lemma A.2.3

with $\mathcal{F} = [0, c\mathbf{1}]^2$ and $\rho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) = \|\mathbf{u} - \mathbf{u}'\| + \|\mathbf{v} - \mathbf{v}'\|$, where $\|\cdot\|$ is the Euclidean norm.

To verify the first conditions in Lemma A.2.3, note that

$$\|Z_{ni}\|_{\mathcal{F}} = \omega_n |\boldsymbol{\varepsilon}_i| \mathbb{1}_{\{\mathbf{X}_i \in I_{[j(-c\mathbf{1}); j(c\mathbf{1})]}\}}.$$

For any $\eta > 0$,

$$\begin{aligned} & \sum_{i=1}^n \mathbb{E} \|Z_{ni}\|_{\mathcal{F}}^2 \mathbb{1}_{\{\|Z_{ni}\|_{\mathcal{F}} > \eta\}} \\ & \leq \omega_n^2 \sum_{i=1}^n \mathbb{E} [\boldsymbol{\varepsilon}_i^2 \mathbb{1}_{\{\mathbf{X}_i \in I_{[j(-c\mathbf{1}); j(c\mathbf{1})]}\}} \mathbb{1}_{\{\omega_n |\boldsymbol{\varepsilon}_i| > \eta\}}] \\ & = n\omega^2 \int_{I_{[j(-c\mathbf{1}); j(c\mathbf{1})]}} \mathbf{g}(\mathbf{x}) d\mathbf{x} \times \mathbb{E} [\boldsymbol{\varepsilon}^2 \mathbb{1}_{\{|\boldsymbol{\varepsilon}| > \eta\omega_n^{-1}\}}], \end{aligned}$$

which is bounded by a constant multiple of $(2c)^s D_s^J(c\mathbf{1}, c\mathbf{1}) \mathbb{E} [\boldsymbol{\varepsilon}^2 \mathbb{1}_{\{|\boldsymbol{\varepsilon}| > \eta\omega_n^{-1}\}}]$, and hence goes to zero.

To check the second condition, note that

$$\begin{aligned} & \sum_{i=1}^n \mathbb{E} |Z_{ni}(\mathbf{u}, \mathbf{v}) - Z_{ni}(\mathbf{u}', \mathbf{v}')|^2 \\ & \leq n\omega_n^2 \mathbb{E} \boldsymbol{\varepsilon}^2 \mathbb{E} |\mathbb{1}_{\{\mathbf{X} \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}\}} - \mathbb{1}_{\{\mathbf{X} \in I_{[j(-\mathbf{u}'); j(\mathbf{v}')]}\}}|^2. \end{aligned} \tag{2.67}$$

The last factor can be bounded by a constant multiple of

$$\begin{aligned} & \left(\prod_{k=1}^s r_{n,k} \right) \left[\prod_{k=1}^s (u_k + v_k + O(J_k^{-1} r_{n,k}^{-1})) D_s^J(\mathbf{u}, \mathbf{v}) \right. \\ & + \prod_{k=1}^s (u'_k + v'_k + O(J_k^{-1} r_{n,k}^{-1})) D_s^J(\mathbf{u}', \mathbf{v}') \\ & \left. - 2 \prod_{k=1}^s (u_k \wedge u'_k + v_k \wedge v'_k + O(J_k^{-1} r_{n,k}^{-1})) D_s^J(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}') \right]. \end{aligned}$$

Note that $\prod_{k=1}^s r_{n,k} = (n\omega_n^2)^{-1}$. This gives a bound for (2.67) a constant multiple of

$$\begin{aligned} & \prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v}) + \prod_{k=1}^s (u'_k + v'_k) D_s(\mathbf{u}', \mathbf{v}') \\ & - 2 \prod_{k=1}^s (u_k \wedge u'_k + v_k \wedge v'_k) D_s(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}'). \end{aligned}$$

If $\rho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) \leq \delta_n$, using Lemma A.2.7, this expression is bounded by a constant multiple of δ_n . Hence the assertion is verified for every $\delta_n \rightarrow 0$.

It remains to verify the third condition of Lemma A.2.3. For any $\epsilon > 0$, we consider the partition $\mathcal{P}(\delta')$ given by (6.6) with some $\delta' > 0$ depending on ϵ , to be determined later. Let $0 = s_0 < s_1 < \dots < s_l = c$ with $(s_{t-1}, s_t]$ of equal length at most δ' and $l \leq 2c/\delta'$. Then \mathcal{F} is covered by

$$\{(\mathbf{u}, \mathbf{v}) \in [0, c\mathbf{1}]^2 : s_{t_{k-1}} < u_k \leq s_{t_k}, s_{r_{k-1}} < v_k \leq s_{r_k}, 1 \leq k \leq d\}, \mathbf{t}, \mathbf{r} \in \{1, \dots, l\}^d.$$

Let $\mathcal{F}_{\epsilon j}^n$ stand for the elements of the partition indexed by j arranged in a certain order. Then for every j ,

$$\begin{aligned} & \sum_{i=1}^n \mathbb{E} \left[\sup_{f, g \in \mathcal{F}_{\epsilon j}^n} |Z_{ni}(f) - Z_{ni}(g)|^2 \right] \\ & \leq n\omega_n^2 \mathbb{E} \mathbb{E}^2 \mathbb{E} \left| \mathbb{1}_{\{\mathbf{X} \in I_{[j(-s_t):j(s_r)]}\}} - \mathbb{1}_{\{\mathbf{X} \in I_{[j(-s_{t-1}):j(s_{r-1})]\}} \right|^2 \\ & \lesssim \mathbb{E} \mathbb{E}^2 \left(\prod_{k=1}^d (s_{t_k} + s_{r_k} + O(J_k^{-1} r_{n,k}^{-1})) - \prod_{k=1}^d (s_{t_{k-1}} + s_{r_{k-1}} + O(J_k^{-1} r_{n,k}^{-1})) \right) \\ & \lesssim \mathbb{E} \mathbb{E}^2 \left(\prod_{k=1}^d (s_{t_k} + s_{r_k}) - \prod_{k=1}^d (s_{t_{k-1}} + s_{r_{k-1}}) \right), \end{aligned}$$

as s_{t_k}, s_{r_k} are bounded by c and $J_k \gg r_{n,k}^{-1}$ for $k = 1, \dots, d$. By Lemma A.2.7, the above expression is bounded by a constant multiple of δ' . Thus δ' can be set to a suitable multiple of ϵ^2 to satisfy the partitioning condition, while the bracketing number $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_n)$ with respect to the

empirical \mathbb{L}_2 -metric is bounded by $N_\epsilon = l^{2d} \leq (2c/\delta')^{2d} = C\epsilon^{-4d}$ for some constant $C > 0$. Then

$$\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[\cdot]}(\epsilon, \mathcal{F}, \|\cdot\|_n)} d\epsilon \leq \int_0^{\delta_n} \sqrt{\log(C\epsilon^{-4d})} d\epsilon \rightarrow 0,$$

for any $\delta_n \rightarrow 0$. □

Lemma 2.6.5. *Under the conditions of Theorem 2.4.1, for any $c > 0$, uniformly in $(\mathbf{u}, \mathbf{v}) \in [0, c\mathbf{1}] \times [0, c\mathbf{1}]$, we have*

$$\omega_n^{-1}(\overline{f_0(\mathbf{X})}|_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}} - f_0(\mathbf{x}_0)) \rightarrow_{\mathbb{P}_0} \sum_{l \in L^*} \frac{\partial^l f_0(\mathbf{x}_0)}{l+1!} \prod_{k=1}^s \frac{v_k^{l_k+1} - (-u_k)^{l_k+1}}{u_k + v_k}.$$

Proof. By Assumption 1, for $(\mathbf{u}, \mathbf{v}) \in [0, c\mathbf{1}]^2$,

$$\overline{f_0(\mathbf{X})}|_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}} - f_0(\mathbf{x}_0) = \frac{\sum_{i: \mathbf{X}_i \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}} \sum_{l \in L^*} \partial^l f_0(\mathbf{x}_0) (\mathbf{X}_i - \mathbf{x}_0)^l / l!}{N_{[j(-\mathbf{u}), j(\mathbf{v})]}} + o(\omega_n),$$

as u_k, v_k are bounded by c . We observe that

$$\mathbb{E}[\omega_n \sum_{i: \mathbf{X}_i \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}} (\mathbf{X}_i - \mathbf{x}_0)^l] = n\omega_n \int_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}} \prod_{k=1}^s (\mathbf{x} - \mathbf{x}_0)_k^{l_k} g(\mathbf{x}) d\mathbf{x}.$$

Again, by writing $g(\mathbf{x}) = g(\mathbf{x}_s) + (g(\mathbf{x}) - g(\mathbf{x}_s))$ and using the continuity of $g(\mathbf{x})$ at \mathbf{x}_0 , the right-hand side of the last display is reduced to

$$n\omega_n \prod_{k=1}^s \frac{r_{n,k}^{l_k+1}}{l_k+1} [v_k^{l_k+1} - (-u_k)^{l_k+1} + O((J_k r_{n,k})^{-1})] (D_s^J(\mathbf{u}, \mathbf{v}) + o(1)).$$

As $l \in L^*$ with $\sum_{k=1}^s l_k / \beta_k = 1$, $n\omega_n \prod_{k=1}^s r_{n,k}^{l_k+1} = 1$. Since $J_k \gg r_{n,k}^{-1}$ and that u_k and v_k are all bounded for every k , the expression converges to

$$\prod_{k=1}^s (l_k + 1)^{-1} [v_k^{l_k+1} - (-u_k)^{l_k+1}] D_s(\mathbf{u}, \mathbf{v}).$$

Further,

$$\begin{aligned}
& \text{Var}\left(\omega_n \sum_{i: \mathbf{X}_i \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}} (\mathbf{X}_i - \mathbf{x}_0)^l\right) \\
&= n\omega_n^2 \text{Var}\left((\mathbf{X} - \mathbf{x}_0)^l \mathbb{1}_{\{\mathbf{X} \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}\}}\right) \\
&\leq n\omega_n^2 \text{E}\left((\mathbf{X} - \mathbf{x}_0)^{2l} \mathbb{1}_{\{\mathbf{X} \in I_{[j(-\mathbf{u}); j(\mathbf{v})]}\}}\right)
\end{aligned}$$

is bounded by a constant multiple of

$$n\omega_n^{2+\sum_{k=1}^s (2l_k+1)/\beta_k} = n^{-2/(2+\sum_{k=1}^s \beta_k^{-1})}.$$

Together these two imply the assertion. \square

Lemma 2.6.6. *Under the conditions of Theorem 2.4.1, for any $M_n \uparrow \infty$, $\Pi(|f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)| > M_n \omega_n | \mathbb{D}_n) \rightarrow 0$ in P_0 -probability.*

Proof. We first prove $\Pi(f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0) \leq M_n \omega_n | \mathbb{D}_n) \rightarrow 1$ in P_0 -probability. For the ease of notation, we show this for the case $s = d$ only. By the max-min formula, we have

$$\begin{aligned}
& f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0) \\
&\leq \max_{\mathbf{u} \geq \mathbf{0}} \frac{\sum_{[j(-\mathbf{u}); j(\mathbf{1})]} N_j \theta_j}{\sum_{[j(-\mathbf{u}); j(\mathbf{1})]} N_j} - f_0(\mathbf{x}_0) \\
&= \omega_n \max_{\mathbf{u} \geq \mathbf{0}} \{A_n(\mathbf{u}, \mathbf{1}; \boldsymbol{\theta}) + A'_n(\mathbf{u}, \mathbf{1}) + B_n(\mathbf{u}, \mathbf{1})\},
\end{aligned}$$

where A_n , A'_n and B_n are defined in (6.3)–(6.5).

Let

$$E_n = \{a_1 n / (2 \prod_{k=1}^d J_k) \leq \min_j N_j \leq \max_j N_j \leq 2a_2 n / (\prod_{k=1}^d J_k)\}.$$

Then by Lemma A.2.6, writing $\psi_j = \theta_j - \text{E}[\theta_j | \mathbb{D}_n]$, we can bound

$$\text{E}\left[\omega_n \sup_{\mathbf{u} \geq \mathbf{0}} |A_n(\mathbf{u}, \mathbf{1}; \boldsymbol{\theta})| | \mathbb{D}_n\right] = \text{E}\left[\sup_{\mathbf{u} \geq \mathbf{1}} \left| \frac{\sum_{[j(-\mathbf{u}+\mathbf{1}); j(\mathbf{1})]} N_j \psi_j}{\sum_{[j(-\mathbf{u}+\mathbf{1}); j(\mathbf{1})]} N_j} \right| | \mathbb{D}_n\right]$$

by the sum of the supremums over subregions $\prod_{k=1}^d [2^{h_k} \leq u_k \leq 2^{h_k+1}]$ as

$$\begin{aligned}
& \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \mathbb{E} \left[\sup_{\substack{2^{h_k} \leq u_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \frac{\sum_{[j(-u+1); j(1)]} N_j \psi_j}{\sum_{[j(-u+1); j(1)]} N_j} \right| \middle| \mathbb{D}_n \right] \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{\sum_{[j(-2^{h+1}); j(1)]} N_j} \mathbb{E} \left[\sup_{2^{h_k} \leq u_k \leq 2^{h_k+1}} \left| \sum_{[j(-u+1); j(1)]} N_j \psi_j \right| \middle| \mathbb{D}_n \right] \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{\sum_{[j(-2^{h+1}); j(1)]} N_j} \left(\mathbb{E} \left[\left| \sum_{[j(-2^{h+1}+1); j(1)]} N_j \psi_j \right|^2 \middle| \mathbb{D}_n \right] \right)^{1/2}.
\end{aligned}$$

Using (3.7), on the event E_n , this can be bounded by

$$\begin{aligned}
& \sigma_n \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{\left(\sum_{[j(-2^{h+1}+1); j(1)]} N_j \right)^{1/2}}{\sum_{[j(-2^{h+1}); j(1)]} N_j} \\
& \lesssim \sigma_n \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{\left(n \left(\prod_{k=1}^d J_k \right)^{-1} \cdot \left(\prod_k r_k \cdot 2^{h_k+1} J_k \right) \right)^{1/2}}{n \left(\prod_{k=1}^d J_k \right)^{-1} \cdot \left(\prod_k r_k \cdot 2^{h_k} J_k \right)}.
\end{aligned}$$

This is clearly bounded by a constant multiple of $\omega_n \sigma_0 \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} 2^{-\sum_k h_k/2}$, and hence

$$\Pi(\max\{A_n(\mathbf{u}, \mathbf{1}; \boldsymbol{\theta}) : \mathbf{u} \geq \mathbf{0}\} > M_n \omega_n | \mathbb{D}_n) \rightarrow 0,$$

in P_0 -probability.

We have shown in (6.9) that

$$\sup_{\mathbf{u} \geq \mathbf{0}} |A'_n(\mathbf{u}, \mathbf{1})| = O_{P_0} \left(\omega_n^{1 + \sum_{k=1}^s \beta_k^{-1}} \prod_{k=1}^d J_k \right) \rightarrow 0,$$

by the choice of J_k in P_0 -probability.

To bound B_n , we decompose B_n into an approximation part and an error part, and bound these two parts separately. Using the similar calculation for the expectation of

$$\omega_n \sup_{\mathbf{u} \geq \mathbf{0}} |A_n(\mathbf{u}, \mathbf{1}; \boldsymbol{\theta})|,$$

restricted on the event E_n , we obtain

$$\mathbb{E}\left[\sup_{\mathbf{u} \geq \mathbf{0}} |\tilde{\varepsilon}|_{I_{[j(-\mathbf{u}); j(\mathbf{1})]}} \mid \mathbb{1}_{A_n}\right] = O(\omega_n).$$

By the monotonicity of f_0 and Assumption 1,

$$\overline{f_0(\mathbf{X})}_{I_{[j(-\mathbf{u}); j(\mathbf{1})]}} - f_0(\mathbf{x}_0) \leq f_0(\mathbf{x}_0 + \mathbf{r}_n + \mathbf{J}^{-1}) - f_0(\mathbf{x}_0),$$

which can be expanded as

$$\sum_{l \in L^*} \partial^l f_0(\mathbf{x}_0) \mathbf{r}_n^l / l! + o(\omega_n) = O(\omega_n).$$

Combining these bounds, the claim follows. For the other side, we note that

$$\begin{aligned} & -(f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)) \\ & \leq -\omega_n \min_{\mathbf{v} \geq \mathbf{0}} \{A_n(\mathbf{1}, \mathbf{v}; \boldsymbol{\theta}) + A'_n(\mathbf{1}, \mathbf{v}) + B_n(\mathbf{1}, \mathbf{v})\} \\ & = \omega_n \max_{\mathbf{v} \geq \mathbf{0}} \{|A_n(\mathbf{1}, \mathbf{v}; \boldsymbol{\theta})| + |A'_n(\mathbf{1}, \mathbf{v})| + |B_n(\mathbf{1}, \mathbf{v})|\} \end{aligned}$$

and apply the same line of arguments. □

With the aid of Lemma 2.6.6, the second condition of Lemma A.2.1 is verified by Lemma 2.6.7 in the following.

Lemma 2.6.7. *Let \mathbf{u}^* and \mathbf{v}^* be any pair indexes such that*

$$f_*(\mathbf{x}_0) = \max_{\mathbf{u} \geq \mathbf{0}} \min_{\mathbf{v} \geq \mathbf{0}} \frac{\sum_{[j(-\mathbf{u}); j(\mathbf{v})]} N_j \theta_j}{\sum_{[j(-\mathbf{u}); j(\mathbf{v})]} N_j} = \frac{\sum_{[j(-\mathbf{u}^*); j(\mathbf{v}^*)]} N_j \theta_j}{\sum_{[j(-\mathbf{u}^*); j(\mathbf{v}^*)]} N_j}. \quad (2.68)$$

Let $\omega_n = n^{-1/(2+\sum_{k=1}^s \beta_k^{-1})}$ and let $\mathbf{r}_n = (\omega_n^{1/\beta_1}, \dots, \omega_n^{1/\beta_s}, 1, \dots, 1)^\top$. Suppose that \mathbf{J} satisfies $J_k \gg r_{n,k}^{-1}$, for each $k = 1, \dots, d$, and $\prod_{k=1}^d J_k \ll n \omega_n$. Under Assumptions 4 and 5, there exists $\gamma > 0$ such

that

$$\lim_{c \rightarrow \infty} \limsup_{n \rightarrow \infty} \Pi(c^{-\gamma} \leq \min_{1 \leq k \leq d} \{v_k^*\} \leq \max_{1 \leq k \leq d} \{v_k^*\} \leq c | \mathbb{D}_n) = 1,$$

in P_0 -probability.

Proof. We write

$$f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0) = \omega_n \max_{\mathbf{u} \geq \mathbf{0}} \min_{\mathbf{v} \geq \mathbf{0}} \{A_n(\mathbf{u}, \mathbf{v}) + A'_n(\mathbf{u}, \mathbf{v}) + B_n(\mathbf{u}, \mathbf{v})\}.$$

Furthermore, we write $B_n(\mathbf{u}, \mathbf{v}) = B_{1n}(\mathbf{u}, \mathbf{v}) + B_{2n}(\mathbf{u}, \mathbf{v})$, where

$$B_{1n}(\mathbf{u}, \mathbf{v}) = \omega_n^{-1} (\bar{\varepsilon} |_{I_{[j(-\mathbf{u}); j(\mathbf{v})]}},$$

$$B_{2n}(\mathbf{u}, \mathbf{v}) = \omega_n^{-1} (\overline{f_0(\mathbf{X})} |_{I_{[j(-\mathbf{u}); j(\mathbf{v})]} - f_0(\mathbf{x}_0)).$$

For $c > \max\{(1 - x_{0,k}) : s+1 \leq k \leq d\}$, we only need to consider the event $\{\max\{v_k^* : 1 \leq k \leq s\} > c\}$. By the monotonicity of f , we have

$$\overline{f_0(\mathbf{X})} |_{I_{[j(-\mathbf{u}^*); j(\mathbf{v}^*)]}} - f_0(\mathbf{x}_0) \geq \overline{f_0(\mathbf{X})} |_{I_{[j(-1); j(\mathbf{v}^*)]}} - f_0(\mathbf{x}_0).$$

By Lemma A.2.8, on an event with P_0 -probability tending to 1, up to $o(\omega_n \max\{v_k^{*\beta_k} : 1 \leq k \leq s\})$, this can be expressed as

$$\begin{aligned} & \sum_{l \in L^*} \frac{\partial^l f_0(\mathbf{x}_0)}{l!} \frac{\sum_{i: \mathbf{X}_i \in I_{[j(-1); j(\mathbf{v}^*)]}} (\mathbf{X}_i - \mathbf{x}_0)^l}{\sum_{[j(-1); j(\mathbf{v}^*)]} N_j} \\ &= \sum_{l \in L^*} \frac{\partial^l f_0(\mathbf{x}_0)}{l!} \frac{\int_{I_{[j(-1); j(\mathbf{v}^*)]}} (\mathbf{x} - \mathbf{x}_0)^l g(\mathbf{x}) d\mathbf{x}}{\int_{I_{[j(-1); j(\mathbf{v}^*)]}} g(\mathbf{x}) d\mathbf{x}} \\ &= \sum_{l \in L^*} \frac{\partial^l f_0(\mathbf{x}_0)}{l!} \prod_{k=1}^s \frac{\int_{[(x_{0,k} - r_{n,k}) J_k] / J_k}^{[(x_{0,k} + r_{n,k} v_k^*) J_k] / J_k} (\mathbf{x} - \mathbf{x}_0)^{l_k} g(\mathbf{x}) d\mathbf{x}}{\int_{[(x_{0,k} - r_{n,k}) J_k] / J_k}^{[(x_{0,k} + r_{n,k} v_k^*) J_k] / J_k} g(\mathbf{x}) d\mathbf{x}} \\ &\lesssim \sum_{l \in L^*} \frac{\partial^l f_0(\mathbf{x}_0)}{(l+1)!} \prod_{k=1}^s \frac{(v_k^* r_{n,k})^{l_k+1} - (-r_{n,k})^{l_k+1}}{(1 + v_k^*) r_{n,k}}, \end{aligned}$$

which is bounded above by a constant multiple of $\omega_n \max\{v_k^{*\beta_k} : 1 \leq k \leq s\}$. As $\Pi(|f_*(\mathbf{x}_0) - f_0(\mathbf{x}_0)| > M_n \omega_n |D_n) \xrightarrow{P_0} 0$, in view of Lemma 6.6, this gives that $\Pi(\max\{v_k^* : 1 \leq k \leq s\} \leq c |D_n) \xrightarrow{P_0} 1$ when $n, c \rightarrow \infty$.

Define $\Lambda^{(0)} = \{v_d^* < c^{-\gamma}\}$ and $\Lambda^{(1)} = \{\max(v_k^* : 1 \leq k \leq d) \leq c\}$. By the previous proof, for every $\eta, \epsilon > 0$, we have $P_0(\Pi(\Lambda^{(1)} | D_n) \geq 1 - \eta) \geq 1 - \epsilon$ when n and c are large enough. We consider $s < d$ for simplicity. The case $s = d$ follows with a slightly different bound by the same argument. For some $a, b, \gamma > 0$ to be determined later, define a subset $R_{a,b,\gamma}(c) \subset \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$ by

$$\left\{ (\mathbf{u}, \mathbf{v}) : \begin{array}{ll} 0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq s} + x_{0,k} \mathbb{1}_{s+1 \leq k \leq d}, & 0 \leq u_d \leq c^{-b} \\ 0 \leq v_k \leq c \mathbb{1}_{1 \leq k \leq s} + (1 - x_{0,k}) \mathbb{1}_{s+1 \leq k \leq d}, & 0 \leq v_d \leq c^{-\gamma} \end{array} \right\}.$$

Define these two events, for some $C_1 > 0$ and $C'_1 > 0$,

$$\begin{aligned} \Lambda^{(2)} &= \left\{ \sup_{(\mathbf{u}, \mathbf{v}) \in R_{a,b,\gamma}} |\mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v}) - \mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]})| \leq (C_1/\eta) \sqrt{c^{as-\gamma} \log c} \right\}, \\ \Lambda'^{(2)} &= \left\{ \sup_{(\mathbf{u}, \mathbf{v}) \in R_{a,b,\gamma}} |\mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v}) - \mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]})| \leq (C'_1/\epsilon) \sqrt{c^{as-\gamma} \log c} \right\}. \end{aligned}$$

Since $\mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v}) \rightsquigarrow \sigma_0 H_2(\mathbf{u}, \mathbf{v})$ in P_0 -probability in $\mathbb{L}_\infty([0, c] \times [0, c])$ and

$$\mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v}) \rightsquigarrow \sigma_0 H_1(\mathbf{u}, \mathbf{v}) \text{ in } \mathbb{L}_\infty([0, c] \times [0, c]),$$

for any $c > 0$, it follows from Lemma A.2.9 that, when n and c are large enough, there exist constants C_1 and C'_1 depending on σ_0^2, d, a only, such that $P_0(\Pi(\Lambda^{(2)} | D_n) \geq 1 - \eta) \geq 1 - \epsilon$ and $P_0(\Lambda'^{(2)}) \geq 1 - \epsilon$.

By Lemma A.2.8, given any $\eta, \epsilon > 0$, there exists $\rho_{\eta\epsilon} > 0$ such that, when $a > 1, c > 1$ and n

large enough, we have

$$\begin{aligned}
& P_0 \times \Pi \left(\min_{\substack{0 \leq v_k \leq c \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq v_d \leq c^{-\gamma}}} \max_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} \right. \\
& \quad \left. \mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]}) + \mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]}) \leq \sqrt{c^{as-b/x_{0,d}}} \rho_{\eta\epsilon} \right) \\
& \lesssim P \left(\min_{\substack{0 \leq v_k \leq c \mathbb{1}_{s+1 \leq k \leq d} \\ v_d=0}} \max_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} H_2(\mathbf{u}, \mathbf{v}) + H_1(\mathbf{u}, \mathbf{v}) \leq \sqrt{c^{as-b/x_{0,d}}} \rho_{\eta\epsilon} \right) \\
& \leq P \left(\min_{\substack{0 \leq v_k \leq c \mathbb{1}_{s+1 \leq k \leq d} \\ v_d=0}} \max_{\substack{0 \leq u_k \leq \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq x_{0,d}}} H_2(\mathbf{u}, \mathbf{v}) + H_1(\mathbf{u}, \mathbf{v}) \leq \rho_{\eta\epsilon} \right) \leq \eta\epsilon.
\end{aligned}$$

Hence, $P_0 \times \Pi(\Lambda^{(3)}) \geq 1 - \eta\epsilon$ for sufficiently large n , where $\Lambda^{(3)}$ stands for the event that for any $0 \leq v_k \leq c \mathbb{1}_{\{1 \leq k \leq d\}}$, $0 \leq v_d \leq c^{-\gamma}$, there exists $0 \leq u_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d\}}$, $0 \leq u_d \leq c^{-b}$ such that

$$\mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]}) + \mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]}) > C_2 \sqrt{c^{as-b/x_{0,d}}}$$

for some constant $C_2 > 0$. Therefore, we have $P_0(\Pi(\Lambda^{(3)}) \geq 1 - \eta) \geq 1 - \epsilon$.

Let

$$u(c) = (c^a \mathbb{1}_{\{1 \leq k \leq s\}} + x_{0k} \mathbb{1}_{\{s+1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}} : k = 1, \dots, d)$$

and

$$v(c) = (c \mathbb{1}_{1 \leq k \leq s} + (1 - x_{0,k}) \mathbb{1}_{s+1 \leq k \leq d-1} + c^{-\gamma} \mathbb{1}_{k=d} : k = 1, \dots, d).$$

By Bernstein's inequality (cf. Lemma 2.2.9 of van der Vaart and Wellner (1996)),

$$P_0 \left(\left| \sum_{[j(-u(c)):j(v(c))]} N_j - E \left[\sum_{[j(-u(c)):j(v(c))]} N_j \right] \right| \geq n\sigma_c^2 \right) \leq C_3 \exp\{-C_3^{-1} n\sigma_c^2\},$$

where $\sigma_c^2 = \text{Var}(\mathbb{1}_{\{\mathbf{X} \in I_{[j(-u(c)):j(v(c))]\}}}) \leq E(\mathbb{1}_{\{\mathbf{X} \in I_{[j(-u(c)):j(v(c))]\}}})$. Then for

$$\Lambda^{(4)} = \left\{ \sum_{[j(-u(c)):j(v(c))]} N_j \leq 2E \left[\sum_{[j(-u(c)):j(v(c))]} N_j \right] \right\},$$

it follows that $P_0(\Lambda^{(4)}) \rightarrow 1$.

On $\Lambda^{(0)} \cap \Lambda^{(1)} \cap \Lambda^{(1)} \cap \Lambda^{(2)} \cap \Lambda^{(3)} \cap \Lambda^{(4)}$, it holds that

$$\begin{aligned}
& \max_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} A_n(\mathbf{u}, \mathbf{v}^*) + B_{1,n}(\mathbf{u}, \mathbf{v}^*) \\
&= \max_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} \frac{\mathbb{H}_{2,n}(\mathbf{u}, \mathbf{v}^*) + \mathbb{H}_{1,n}(\mathbf{u}, \mathbf{v}^*)}{\omega_n^2 \sum_{[j(-\mathbf{u}); j(\mathbf{v}^*)]} N_j} \\
&\geq \frac{C_2 \sqrt{c^{as-b}} - C_1 \sqrt{c^{as-\gamma} \log c / \eta}}{(c^a + c)^s (c^{-b} + c^{-\gamma})},
\end{aligned}$$

which is greater than or equal to $C_3 c^{(b-as)/2}$ for some positive constant C_3 .

On the other hand, for some positive constant C_4 , we have

$$\begin{aligned}
& \min_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} B_{2n}(\mathbf{u}, \mathbf{v}^*) \\
&= \omega_n^{-1} \min_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} \overline{f_0(\mathbf{X})}|_{I_{[j(-\mathbf{u}); j(\mathbf{v}^*)]}} - f_0(\mathbf{x}_0) \\
&\geq \omega_n^{-1} (f_0(\mathbf{x}_0 - (c^a + o(1))\mathbf{r}_n) - f_0(\mathbf{x}_0)),
\end{aligned}$$

which is greater than or equal to $-C_4 c^{a \max_{k \leq s} \beta_k}$. In view of (6.9), we conclude that, on $\Lambda^{(0)} \cap \Lambda^{(1)} \cap \Lambda^{(1)} \cap \Lambda^{(2)} \cap \Lambda^{(3)} \cap \Lambda^{(4)}$, $f^*(\mathbf{x}_0) - f(\mathbf{x}_0)$ is bounded below by

$$\begin{aligned}
& \omega_n \max_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} \{A_n(\mathbf{u}, \mathbf{v}^*) + B_{1n}(\mathbf{u}, \mathbf{v}^*)\} \\
&+ \min_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} B_{2n}(\mathbf{u}, \mathbf{v}^*) - \max_{\substack{0 \leq u_k \leq c^a \mathbb{1}_{1 \leq k \leq d} \\ 0 \leq u_d \leq c^{-b}}} |A'_n(\mathbf{u}, \mathbf{v}^*)| \\
&\geq \omega_n c^{a \max\{\beta_k : k \leq s\}} (C_3 c^{(b-as)/2 - a \max_{k \leq s} \beta_k} - C_4 + o(1)).
\end{aligned}$$

Take $a = 3$, $b \geq 2(1 + a \max\{\beta_k : k \leq s\}) + as$ and $\gamma = b + 1$. Hence the intersection of these events can only occur with arbitrarily small posterior probability in P_0 -probability in view of Lemma 6.6, for large enough n and c . As $\Pi(\Lambda^{(0)} \cap \Lambda^{(1)} \cap \Lambda^{(2)} \cap \Lambda^{(3)} | \mathbb{D}_n) \rightarrow_{P_0} 0$ while $\Pi(\Lambda^{(1)} \cap \Lambda^{(2)} \cap \Lambda^{(3)} | \mathbb{D}_n) \rightarrow_{P_0} 1$ when $n, c \rightarrow \infty$, thus we can conclude $\Pi([\Lambda^{(0)}]^c \cap \Lambda^{(1)} | \mathbb{D}_n) \rightarrow 1$ in P_0 -probability. \square

The proof of Theorem 2.4.1 can now be completed. Using arguments similar to Proposition 7

of Han and Zhang (2020), it can be verified that $P(W_c \neq W) \rightarrow 0$ as $c \rightarrow \infty$. Hence the proof follows by an application of Lemma A.2.1.

Proof of Proposition 2.4.1

This can be shown by the self-similarity property of Gaussian processes H_1 and H_2 : for $\mathbf{t} \in \mathbb{R}_{>0}^d$ such that $t_{s+1} = \dots = t_d = 1$, we have that

$$H_i(\mathbf{t} \circ \mathbf{u}, \mathbf{t} \circ \mathbf{v}) =_d \left(\prod_{j=1}^s t_j \right)^{1/2} H_i(\mathbf{u}, \mathbf{v}),$$

$i = 1, 2$. By the choice of \mathbf{t} , multiplying a vector coordinatewise by \mathbf{t} does not change the last $d - s$ coordinates and thus $D_s(\mathbf{t} \circ \mathbf{u}, \mathbf{t} \circ \mathbf{v}) = D_s(\mathbf{u}, \mathbf{v})$. Then, since a scaling of the domain does not alter suprema and infima, the expression in the limiting distribution is equal to

$$\begin{aligned} & \sup_{\mathbf{u} \geq \mathbf{0}} \inf_{\mathbf{v} \geq \mathbf{0}} \left\{ \frac{\sigma_0 H_1(\mathbf{t} \circ \mathbf{u}, \mathbf{t} \circ \mathbf{v}) + \sigma_0 H_2(\mathbf{t} \circ \mathbf{u}, \mathbf{t} \circ \mathbf{v})}{\prod_{k=1}^s (t_k u_k + t_k v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \\ & \quad \left. + \sum_{k=1}^s \left[\frac{\partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \cdot \frac{(t_k v_k)^{\beta_k + 1} - (-t_k u_k)^{\beta_k + 1}}{t_k u_k + t_k v_k} \right] \right\} \\ & =_d \sup_{\mathbf{u} \geq \mathbf{0}} \inf_{\mathbf{v} \geq \mathbf{0}} \left\{ \left(\sigma_0^{-2} \prod_{j=1}^s t_j \right)^{-1/2} \frac{H_1(\mathbf{u}, \mathbf{v}) + H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v})} \right. \\ & \quad \left. + \sum_{k=1}^s \left[\frac{t_k^{\beta_k} \partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \cdot \frac{v_k^{\beta_k + 1} - (-u_k)^{\beta_k + 1}}{u_k + v_k} \right] \right\}. \end{aligned}$$

By equating $(\sigma_0^{-2} \prod_{j=1}^s t_j)^{-1/2}$ to $t_k^{\beta_k} \partial_k^{\beta_k} f_0(\mathbf{x}_0) / (\beta_k + 1)!$ for each $k = 1, \dots, s$, we can find the solution t_k to the system of equations, and also the common factor A_β as stated in the proposition.

If $s = d$, then $D_d(\mathbf{u}, \mathbf{v}) = g(\mathbf{x}_0)$ and $H_i(\mathbf{u}, \mathbf{v}) =_d \sqrt{g(\mathbf{x}_0)} \tilde{H}_i(\mathbf{u}, \mathbf{v})$. For $\mathbf{t} \in \mathbb{R}_{>0}^d$,

$$\tilde{H}_i(\mathbf{t} \circ \mathbf{u}, \mathbf{t} \circ \mathbf{v}) =_d \left(\prod_{j=1}^d t_j \right)^{1/2} \tilde{H}_i(\mathbf{u}, \mathbf{v}),$$

$i = 1, 2$. Hence by self-similarity, the last expression reduces to

$$\begin{aligned} & \sup_{u \geq 0} \inf_{v \geq 0} \left\{ \frac{\sigma_0}{\sqrt{g(\mathbf{x}_0)}} \left(\frac{\tilde{H}_1(t \circ \mathbf{u}, t \circ \mathbf{v})}{\prod_{k=1}^d (t_k u_k + t_k v_k)} + \frac{\tilde{H}_2(t \circ \mathbf{u}, t \circ \mathbf{v})}{\prod_{k=1}^d (t_k u_k + t_k v_k)} \right) \right. \\ & \quad \left. + \sum_{k=1}^d \left[\frac{\partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \cdot \frac{(t_k v_k)^{\beta_k + 1} - (-t_k u_k)^{\beta_k + 1}}{t_k u_k + t_k v_k} \right] \right\} \\ & = {}_d \sup_{u \geq 0} \inf_{v \geq 0} \left\{ \sqrt{\frac{\sigma_0^2}{g(\mathbf{x}_0) \prod_{j=1}^d t_j}} \left(\frac{\tilde{H}_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\tilde{H}_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} \right) \right. \\ & \quad \left. + \sum_{k=1}^d \left[\frac{t_k^{\beta_k} \partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!} \cdot \frac{v_k^{\beta_k + 1} - (-u_k)^{\beta_k + 1}}{u_k + v_k} \right] \right\}. \end{aligned}$$

By exploring the equation system for t_k as follows,

$$\sqrt{\frac{\sigma_0^2}{g(\mathbf{x}_0) \prod_{j=1}^d t_j}} = \frac{t_k^{\beta_k} \partial_k^{\beta_k} f_0(\mathbf{x}_0)}{(\beta_k + 1)!}, \text{ for } k = 1, \dots, d,$$

we can find the common factor \tilde{A}_β in a similar way of solving a set of equations.

Proof of Proposition 2.4.2. For $0 \leq z \leq 1$,

$$\begin{aligned} \mathbb{P}(Z_B^{(1)} \leq z) &= \mathbb{P}(1 - Z_B^{(1)} \geq 1 - z) \\ &= \mathbb{P}(\mathbb{P}(-\sup_{u \geq 0} \inf_{v \geq 0} \tilde{U}(\mathbf{u}, \mathbf{v}) \leq 0 | \tilde{H}_1) \geq 1 - z) \\ &= \mathbb{P}(\mathbb{P}(\inf_{u \geq 0} \sup_{v \geq 0} [-\tilde{U}(\mathbf{u}, \mathbf{v})] \leq 0 | \tilde{H}_1) \geq 1 - z). \end{aligned}$$

Note that $\tilde{H}_i(\mathbf{u}, \mathbf{v}) =_d \tilde{H}_i(\mathbf{v}, \mathbf{u})$ and $\tilde{H}_i =_d -\tilde{H}_i$ for $i = 1, 2$. Denote $\tilde{H}_1^* = -\tilde{H}_1$. Then we have

$$\begin{aligned}
& \mathbb{P}\left(\inf_{\mathbf{u} \geq \mathbf{0}} \sup_{\mathbf{v} \geq \mathbf{0}} [-\tilde{U}(\mathbf{u}, \mathbf{v})] \leq 0 \mid \tilde{H}_1\right) \\
&= {}_d \mathbb{P}\left(\inf_{\mathbf{u} \geq \mathbf{0}} \sup_{\mathbf{v} \geq \mathbf{0}} \left\{ \frac{\tilde{H}_1^*(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{-\tilde{H}_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} \right. \right. \\
&\quad \left. \left. + \sum_{k=1}^d \frac{-v_k^{\beta_k+1} + (-u_k)^{\beta_k+1}}{u_k + v_k} \right\} \leq 0 \mid -\tilde{H}_1^*\right) \\
&= {}_d \mathbb{P}\left(\inf_{\mathbf{u} \geq \mathbf{0}} \sup_{\mathbf{v} \geq \mathbf{0}} \left\{ \frac{\tilde{H}_1^*(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\tilde{H}_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \sum_{k=1}^d \frac{u_k^{\beta_k+1} - v_k^{\beta_k+1}}{u_k + v_k} \right\} \leq 0 \mid \tilde{H}_1^*\right) \\
&= {}_d \mathbb{P}\left(\inf_{\mathbf{u} \geq \mathbf{0}} \sup_{\mathbf{v} \geq \mathbf{0}} \left\{ \frac{\tilde{H}_1^*(\mathbf{v}, \mathbf{u})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\tilde{H}_2(\mathbf{v}, \mathbf{u})}{\prod_{k=1}^d (u_k + v_k)} + \sum_{k=1}^d \frac{u_k^{\beta_k+1} - v_k^{\beta_k+1}}{u_k + v_k} \right\} \leq 0 \mid \tilde{H}_1^*\right) \\
&= \mathbb{P}\left(\inf_{\mathbf{u} \geq \mathbf{0}} \sup_{\mathbf{v} \geq \mathbf{0}} \tilde{U}(\mathbf{v}, \mathbf{u}) \leq 0 \mid \tilde{H}_1^*\right).
\end{aligned}$$

Hence $\mathbb{P}(Z_B^{(1)} \leq z) = \mathbb{P}(Z_B^{(2)} \geq 1 - z)$. The symmetry of the distribution of $Z_B^{(3)}$ holds by similar arguments. □

CHAPTER

3

MULTIVARIATE MONOTONE DENSITIES: CONTRACTION RATES, TESTS, AND FREQUENTIST COVERAGE

We aim to make inferences on the unknown multivariate probability density function g based on the independent and identically distributed (i.i.d.) sample of size n . Specifically we have n d -dimensional i.i.d. observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ from an unknown probability distribution G with probability density function g supported on $[0, 1]^d$. We shall use \mathbb{D}_n to denote the data. The probability density function g is presumed to be nonincreasing with respect to each of its variate when fixing the other ones. Like in Chapter 2, we define the natural partial ordering on \mathbb{R}^d in the following way. We say $\mathbf{x} \succeq \mathbf{y}$ if and only if $x_k \geq y_k$ for all $1 \leq k \leq d$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and

similarly define $x \preceq y$. Then g is supposed to be monotonically nonincreasing with respect to the natural partial ordering. We denote the set of multivariate nonincreasing probability density functions as follows,

$$\mathcal{G}^* = \{g : [0, 1]^d \rightarrow \mathbb{R}_{\geq 0} : g(\mathbf{x}) \leq g(\mathbf{y}) \text{ if } \mathbf{x} \succeq \mathbf{y}; \int g = 1\}.$$

Let

$$\mathcal{F}^* = \{f : [0, 1]^d \rightarrow \mathbb{R} : f(\mathbf{x}) \leq f(\mathbf{y}) \text{ if } \mathbf{x} \succeq \mathbf{y}\}$$

stand for the set of multivariate nonincreasing functions. Let $\mathcal{H}(\alpha, L)$ stand for the set of Hölder continuous function class on $[0, 1]^d$ with Hölder smoothness index α and Hölder constant L ; see Definition A.1.1.

3.1 Prior and posterior

The class of piece-wise constant functions with an increasing number of hyperrectangles approximates the class of all integrable functions. Therefore, to put a prior on the density function g , we can consider a distribution on piece-wise constant functions. To this end, partition the domain $[0, 1]^d$ by splitting the k th direction into J_k equal length subintervals, $\{[0, 1/J_k], (1/J_k, 2/J_k], \dots, ((J_k - 1)/J_k, 1]\}$, $k = 1, \dots, d$, resulting in $\prod_{k=1}^d J_k$ pieces in total. Let $I_1 = \prod_{k=1}^d [0, 1/J_k]$ and $I_j = \prod_{k=1}^d ((j_k - 1)/J_k, j_k/J_k]$ for $\mathbf{j} \in [1 : \mathbf{J}] \setminus \{\mathbf{1}\}$. Then a typical function in the support of the prior can be represented as

$$g = \left(\prod_{k=1}^d J_k \right) \sum_{\mathbf{j} \in [1 : \mathbf{J}]} \theta_{\mathbf{j}} \mathbb{1}_{I_{\mathbf{j}}},$$

with a suitable prior on the scaled vector of step-heights $\boldsymbol{\theta} := (\theta_{\mathbf{j}} : \mathbf{j} \in [1 : \mathbf{J}])$. Let the set of multivariate functions taking constant value on $I_{\mathbf{j}}$, $\mathbf{j} \in [1 : \mathbf{J}]$, be denoted by $\mathcal{F}_{\mathbf{J}}$, and the set of multivariate piece-wise constant probability densities by $\mathcal{G}_{\mathbf{J}}$. These sets of functions with

the constraint of multivariate monotonicity are respectively $\mathcal{F}_J^* = \mathcal{F}^* \cap \mathcal{F}_J$ and $\mathcal{G}_J^* = \mathcal{G}^* \cap \mathcal{G}_J$. For most results in this paper, J_k , $k = 1, \dots, d$, can be taken to be deterministic with value dependent on n .

A conjugate prior Π on g is given by a Dirichlet distribution on θ : for some array of positive numbers $(\alpha_j : j \in [1 : J])$,

$$(\theta_j : j \in [1 : J]) \sim \text{Dir}\left(\prod_{k=1}^d J_k; (\alpha_j : j \in [1 : J])\right), \quad (3.1)$$

leading to the posterior measure $\Pi(\cdot | \mathbb{D}_n)$ given by

$$(\theta_j : j \in [1 : J]) | \mathbb{D}_n \sim \text{Dir}\left(\prod_{k=1}^d J_k; (\alpha_j + N_j : j \in [1 : J])\right), \quad (3.2)$$

where $N_j = \sum_{i=1}^n \mathbb{1}_{I_j}(\mathbf{X}_i)$.

Clearly, neither the prior nor the posterior is supported within the desirable space \mathcal{G}^* of multivariate monotone densities. To comply with the shape restriction, we make an inference based on the posterior distribution $\Pi_n^* := \Pi_n \circ \iota^{-1}$ induced by an immersion map ι that transforms a density to a multivariate monotone density. The choice of the immersion map will depend on the application. More specifically, to obtain the posterior contraction rate, we use an immersion map a composition of the \mathbb{L}_1 -projection on the space of monotone functions \mathcal{F}^* followed by a renormalization taking in \mathcal{G}^* . Indeed, since the immersion map is applied to a sample from the posterior distribution of g which is supported with \mathcal{G}_J , it will be observed that the image under the immersion map belongs to \mathcal{G}_J^* . Thus the immersion posterior is supported within the space of piece-wise constant multivariate monotone densities. This property has the significant implication that the computation of the immersion posterior consists entirely of a finite-dimensional sampling and a finite-dimensional optimization algorithm. To study the asymptotic frequentist coverage of a Bayesian credible interval, we use a min-max or max-min block operation, to be precisely defined in Section 3.3. In both cases, the choice of the immersion maps is motivated by the desired asymptotic properties.

3.2 Contraction rates and testing for multivariate monotonicity

We first study the posterior contraction rate of the immersion posterior induced by the map ι mapping $g \in \mathcal{G}_J$ to $g^* \in \mathcal{G}_J^*$ given by

$$g^* = \frac{\tilde{g}}{\int \tilde{g}}, \text{ where } \tilde{g} \in \arg \min\{\|g - h\|_1 : h \in \mathcal{F}^*\}.$$

We observe an important fact below that for any $p \geq 1$, the \mathbb{L}_p -projection of a piece-wise constant function belongs to \mathcal{F}_J^* , and hence $\iota(g) \in \mathcal{G}_J^*$ for any g sampled from the posterior.

Lemma 3.2.1. *For $p \geq 1$ and $s \in \mathbb{L}_p([0, 1]^d)$, let $s_J = (\prod_{k=1}^d J_k) \sum_{j \in [1:J]} b_j \mathbb{1}_{I_j}$ where $b_j = \int_{I_j} s$. Then,*

- (i) *for any $h \in \mathcal{F}_J$, $\|s_J - h\|_p \leq \|s - h\|_p$;*
- (ii) *if $s \in \mathcal{F}^*$, then $s_J \in \mathcal{F}_J^*$;*
- (iii) *if $s \in \mathcal{G}^*$, then $s_J \in \mathcal{G}_J^*$.*

In view of Lemma 3.2.1, to obtain the monotone \mathbb{L}_p -projection of a piece-wise function $h = \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j}$, it suffices to isotone the coefficient vector $\theta = (\theta_j : j \in [1:J])$ by minimizing $\sum_{j \in [1:J]} |c_j - \theta_j|^p$ over $(c_j : j \in [1:J]) \in \mathcal{C}_J$, where $\mathcal{C}_J = \{(c_j : j \in [1:J]) : c_{j_1} \leq c_{j_2}, \text{ if } j_1 \geq j_2\}$, the closed and convex monotone cone. The solution to this isotone problem exists by the convexity of the loss function and the convexity and closeness of the feasible set. It is not hard to see that when $\theta_j \geq 0$ for all $j \in [1:J]$, the nonnegativity of the isotone is ensured. However, for $p \neq 2$, the condition $\sum_{j \in [1:J]} \theta_j = 1$ does not automatically ensure that the isotone coefficients add up to one. For the case $d = 1$ and $p = 2$, from the description of the Pool Adjacent Violation Algorithm (cf. Ayer et al. (1955)), it is clear that the sum of the coefficients does not change, and hence the isotone of a density $g \in \mathcal{G}_J$ is automatically a density. This can be shown to be true for the multivariate case $d \geq 2$ and $p = 2$ and thus makes

the renormalization step redundant. Nevertheless, we work with a monotone \mathbb{L}_1 -projection as the approximation rate at a monotone function by piece-wise constant functions is not sufficiently strong in terms of the \mathbb{L}_p -metric ($p > 1$), and hence a renormalization step following the \mathbb{L}_1 -isotonization of the step-heights is necessary to make the monotone projection a valid probability density. An alternative would be to directly consider an \mathbb{L}_1 -projection of $g \in \mathcal{G}_J$ onto \mathcal{G}^* , but a convenient computational algorithm for that does not seem to be readily available. This necessitates the use of a more general immersion map instead of a projection used for the multivariate monotone regression problem studied in 2.2.

Let $\tilde{\theta} = (\tilde{\theta}_j : j \in [1 : J])$ stand for the isotonization of θ with respect to the \mathbb{L}_1 -distance. There may be more than one solutions, in which case we may choose any of them. Then the immersion posterior sample will be given by $\theta_j^* = \tilde{\theta}_j / \sum_{j \in [1:J]} \tilde{\theta}_j$ for all $j \in [1 : J]$.

For the rest of the section, for all $k = 1, \dots, d$, we let $J_k = J$ for some J . The following result gives the posterior contraction rate of the resulting immersion posterior.

Theorem 3.2.2. *Let the true density $g_0 \in \mathcal{G}^*$. If $1 \ll J^d \ll n$ and $a_0 \leq \alpha_j \leq a_1$ for all $j \in [1 : J]$ and some $a_0, a_1 > 0$, then $E_0 \Pi(\|g^* - g_0\|_1 > M_n \epsilon_n | \mathbb{D}_n) \rightarrow 0$ for any $M_n \rightarrow \infty$, where $\epsilon_n = \max\{J^{-1}, \sqrt{J^d/n}\}$.*

The optimal rate $n^{-1/(2+d)}$ is achieved by choosing $J \asymp n^{1/(2+d)}$.

Next, we construct a Bayesian test for multivariate monotonicity analogous to that in 2.2 for the multivariate monotone regression problem.

Theorem 3.2.3. *Let $\phi_n = \mathbb{1}\{\Pi(d_1(g, \mathcal{F}^*) \leq M_n n^{-1/(2+d)} | \mathbb{D}_n) < \gamma\}$, where M_n is a slowly growing sequence and $\gamma \in (0, 1)$ is a predetermined constant. If $J \asymp n^{1/(2+d)}$, then we have*

- (i) *for any fixed $g_0 \in \mathcal{G}^*$, $E_0(\phi_n) \rightarrow 0$;*
- (ii) *for any fixed g_0 not in the \mathbb{L}_1 -closure of \mathcal{G}^* , $E_0(1 - \phi_n) \rightarrow 0$;*
- (iii) *for any $\alpha \in (0, 1]$ and any $L > 0$, we have $\sup\{E_0(1 - \phi_n) : g_0 \in \mathcal{H}(\alpha, L), d_1(g_0, \mathcal{G}^*) > \rho_n(\alpha)\} \rightarrow$*

0, where

$$\rho_n(\alpha) = \begin{cases} C n^{-\alpha/(2+d)}, & \text{for some } C > 0 \text{ if } \alpha < 1, \\ C M_n n^{-1/(2+d)}, & \text{for any } C > 2 \text{ if } \alpha = 1. \end{cases}$$

The universal consistency against any fixed non-monotone alternative density is appealing. Nevertheless, the separation rate $n^{-\alpha/(2+d)}$ of the alternative density from the null hypothesis of multivariate monotonicity assumed above for power to approach one is slower than the optimal posterior contraction rate $n^{-\alpha/(2\alpha+d)}$ of Hölder functions of smoothness index α . This is because the optimal posterior contraction rate for this class using piece-wise constant functions is obtained only by choosing $J \asymp n^{1/(2\alpha+d)}$, while the choice $J \asymp n^{1/(2+d)}$ used to construct the test leads to the suboptimal contraction rate $n^{-\alpha/(2+d)}$. However, the assumed choice is essential to obtain the optimal posterior contraction rate at multivariate monotone functions since otherwise, at some monotone true density, the posterior will not concentrate within its $M_n n^{-1/(2+d)}$ -neighborhood, and hence the size of the Bayesian test will tend to one. The problem can be rectified by also putting an appropriate prior on J , but at the expense of higher computational complexity. Therefore, fixed J can not give rise to the optimal separation rate for some classes of smooth functions. Then the required separation rate adapts optimally with α within a logarithmic factor.

Theorem 3.2.4. *Let J be endowed with a prior π such that*

$$e^{-b_1 J^d \log J} \leq \pi(J) \leq e^{-b_2 J^d \log J}$$

for some b_1 and $b_2 > 0$, and suppose that $a_2 n^{-a_3} \leq \alpha_j \leq a_1$, for some constants a_1, a_2 , and $a_3 > 0$.

Consider the test

$$\phi_n = \mathbb{1}\{\Pi(d_1(g, \mathcal{F}^*) \leq M_0 \sqrt{(J^d \log n)/n} | \mathbb{D}_n) < \gamma\}$$

for testing the hypothesis of multivariate monotonicity of g , where M_0 is a sufficiently large constant and $\gamma \in (0, 1)$ is predetermined. Then

- (i) for fixed $g_0 \in \mathcal{G}^*$ and g_0 bounded away from zero, $E_0(\phi_n) \rightarrow 0$;
- (ii) for fixed density g_0 bounded away from zero and not belonging to the \mathbb{L}_1 -closure of \mathcal{G}^* , $E_0(1 - \phi_n) \rightarrow 0$;
- (iii) there exists a sufficiently large constant $C > 0$ depending only on α , L , and d , such that $\sup\{E_0(1 - \phi_n) : g_0 \in \mathcal{H}(\alpha, L), g_0 \geq l > 0, d_1(g_0, \mathcal{G}^*) > C(n/\log n)^{-\alpha/(2\alpha+d)}\} \rightarrow 0$.

3.3 Coverage of pointwise credible intervals

To obtain a Bayesian credible interval for $g(x_0)$ at an interior point $x_0 \in (0, 1)^d$ with an asserted frequentist coverage, we use quantiles of immersion posteriors induced by the immersion maps via the max-min (or min-max) operation over blocks containing x_0 . The immersion map is partly inspired by the operation used in the estimator of the monotone function value at a point proposed by Fokianos et al. (2020). The asymptotic distribution of such block estimators was studied by Han and Zhang (2020) for the multivariate monotone regression problem. We drew on their methodology in the proof of the following theorem.

Throughout this section, \mathbf{J} is taken as deterministic and changing with the sample size. We allow J_k to take different values adapting to the different local smoothness levels along each coordinate. Let $j_0 = \lceil x_0 \circ \mathbf{J} \rceil$ so that $x_0 \in I_{j_0}$. Since x_0 is arbitrary so far, we can define an immersion map ι by the value of $\iota(g)$ at x_0 for any $g = (\prod_{k=1}^d J_k) \sum_{j \in [1:J]} \theta_j \mathbb{1}_{I_j}$ with θ from the unit simplex in $\mathbb{R}^{\prod_{k=1}^d J_k}$. Specifically, we consider the min-max immersion map $\bar{\iota}$ defined by $\bar{\iota}(g)(x_0) = \theta_{j_0}^*$, where

$$\theta_{j_0}^* = \min_{j_1 \preceq j_0} \max_{j_0 \preceq j_2} \frac{\sum_{j \in [j_1:j_2]} \theta_j}{\prod_{k=1}^d (j_{2,k} - j_{1,k} + 1)}. \quad (3.3)$$

It is clear that $\theta_{j_0}^*$ is uniquely defined although there can be multiple pairs of j_2 and j_1 maximizing and minimizing the average of θ_j in the last display. It is not hard to see that $\theta^* := (\theta_j^* : j \in [1:J]) \in \mathcal{C}_J$. If we switch the order of maximization and minimization in the transformation

in (3.3), we obtain the max-min immersion map $\underline{\iota}$ defined by $\underline{\iota}(g)(\mathbf{x}_0) = \theta_{j_0}^\dagger$, where

$$\theta_j^\dagger = \max_{j \preceq j_2} \min_{j_1 \preceq j} \frac{\sum_{j \in [j_1:j_2]} \theta_j}{\prod_{k=1}^d (j_{2,k} - j_{1,k} + 1)}. \quad (3.4)$$

It again follows that $\theta^\dagger := (\theta_j^\dagger : j \in [1 : J]) \in \mathcal{C}_J$, but θ^\dagger is possibly different from θ^* . As both operations result in monotone outcomes, the average immersion map $(\bar{\iota} + \underline{\iota})/2$ will also map piece-wise constant functions to monotone piece-wise constant functions. For a given \mathbf{x}_0 , we shall obtain asymptotic frequentist coverage of Bayesian credible intervals of $g(\mathbf{x}_0)$ based on the quantiles of its immersion posterior distribution using the immersion maps $\bar{\iota}$, $\underline{\iota}$ and $(\bar{\iota} + \underline{\iota})/2$.

We shall make the following assumptions on the prior and the true probability density function.

Assumption 6. The parameters of the prior distribution satisfy $\max_j \alpha_j \leq a_1$ for some positive constant a_1 .

Assumption 7. The true density g_0 is a coordinate-wise nonincreasing function on $[0, 1]^d$. g_0 is continuously differentiable in a small neighborhood of $\mathbf{x}_0 \in (0, 1)^d$. For every $1 \leq k \leq d$, there exists some $\eta_k \in \mathbb{Z}_{>0}$ such that $\partial_k^{m_k} g_0(\mathbf{x}_0) = 0$ for $m_k = 1, \dots, \eta_k - 1$ and $\partial_k^{\eta_k} g_0(\mathbf{x}_0) \neq 0$, and for any $t > 0$, with $M_0 = \{\mathbf{m} \in \mathbb{Z}_{\geq 0}^d : \sum_{k=1}^d m_k / \eta_k \leq 1\}$,

$$\sup \left\{ \left| g_0(\mathbf{x}) - \sum_{\mathbf{m} \in M_0} \frac{\partial^{\mathbf{m}} g_0(\mathbf{x}_0)}{m!} (\mathbf{x} - \mathbf{x}_0)^{\mathbf{m}} \right| : |\mathbf{x} - \mathbf{x}_0| \leq t \mathbf{r}_n \right\} = o(\omega_n),$$

where $\omega_n \downarrow 0$ and $\mathbf{r}_n = (\omega_n^{1/\eta_1}, \dots, \omega_n^{1/\eta_d})^\top$.

Assumption 7 is adapted from Han and Zhang (2020) and some unique features, essential for the proof, of multivariate monotone functions follow from this assumption, see Lemma 1 of Han and Zhang (2020). Assumption 7 holds generally when g_0 has smoothness of order $\max\{\eta_k : 1 \leq k \leq d\}$ at \mathbf{x}_0 .

To state the theorems on limiting coverage, we introduce some stochastic processes. Let $H_1(\mathbf{u}, \mathbf{v})$ and $H_2(\mathbf{u}, \mathbf{v})$ be two independent centered Gaussian processes indexed by $(\mathbf{u}, \mathbf{v}) \in$

$\mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$ with the covariance structure

$$\text{Cov}(H_i(\mathbf{u}, \mathbf{v}), H_i(\mathbf{u}', \mathbf{v}')) = \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k),$$

for $(\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}') \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$ and $i = 1, 2$. Then define a Gaussian process by the relation

$$\begin{aligned} U(\mathbf{u}, \mathbf{v}) &= \frac{\sqrt{g_0(\mathbf{x}_0)} H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\sqrt{g_0(\mathbf{x}_0)} H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} \\ &\quad + \sum_{\mathbf{m} \in M} \frac{\partial_{\mathbf{m}}^{m_k} g_0(\mathbf{x}_0)}{(\mathbf{m} + 1)!} \prod_{k=1}^d \frac{v_k^{m_k+1} - (-u)^{m_k}}{u_k + v_k}, \end{aligned}$$

where $M = \{\mathbf{m} \in \mathbb{Z}_{\geq 0}^d : \sum_{k=1}^d m_k / \eta_k = 1\}$. For each of the three immersion maps, let the images be denoted by $g^* = \bar{\iota}(g)$, $g^\dagger = \underline{\iota}(g)$ and $\bar{g} = ((\bar{\iota} + \underline{\iota})/2)(g)$.

Theorem 3.3.1. *Let Assumptions 6 and 7 hold and*

$$\omega_n = n^{-1/(2+\sum_{k=1}^d \eta_k^{-1})} \text{ and } \mathbf{r}_n = (\omega_n^{1/\eta_1}, \dots, \omega_n^{1/\eta_d})^\top.$$

If $J_k \gg r_{n,k}^{-1}$ and $\prod_{k=1}^d J_k \ll n \omega_n$, then for every $z \in \mathbb{R}$, we have

$$\Pi(\omega_n^{-1}(g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0)) \leq z | \mathbb{D}_n) \rightsquigarrow \mathbb{P}\left(\inf_{\mathbf{u} \geq 0} \sup_{\mathbf{v} \geq 0} U(\mathbf{u}, \mathbf{v}) \leq z | H_1\right),$$

$$\Pi(\omega_n^{-1}(g^\dagger(\mathbf{x}_0) - g_0(\mathbf{x}_0)) \leq z | \mathbb{D}_n) \rightsquigarrow \mathbb{P}\left(\sup_{\mathbf{v} \geq 0} \inf_{\mathbf{u} \geq 0} U(\mathbf{u}, \mathbf{v}) \leq z | H_1\right),$$

$$\Pi(\omega_n^{-1}(\bar{g}(\mathbf{x}_0) - g_0(\mathbf{x}_0)) \leq z | \mathbb{D}_n) \rightsquigarrow \mathbb{P}\left(\frac{1}{2}(\inf_{\mathbf{u} \geq 0} \sup_{\mathbf{v} \geq 0} U(\mathbf{u}, \mathbf{v}) + \sup_{\mathbf{v} \geq 0} \inf_{\mathbf{u} \geq 0} U(\mathbf{u}, \mathbf{v})) \leq z | H_1\right).$$

The corresponding immersion-posterior $(1 - \gamma)$ -quantiles are

$$Q_{n,\gamma}^{(1)} = \inf\{z : \Pi(g^*(\mathbf{x}_0) \leq z | \mathbb{D}_n) \geq 1 - \gamma\},$$

$$Q_{n,\gamma}^{(2)} = \inf\{z : \Pi(g^\dagger(\mathbf{x}_0) \leq z | \mathbb{D}_n) \geq 1 - \gamma\},$$

$$Q_{n,\gamma}^{(3)} = \inf\{z : \Pi(\bar{g}(\mathbf{x}_0) \leq z | \mathbb{D}_n) \geq 1 - \gamma\}.$$

Theorem 3.3.1 gives the coverage of the corresponding immersion-posterior credible intervals in terms of the distributions of the following random variables:

$$Z_B^{(1)} = P(\inf_{u \geq 0} \sup_{v \geq 0} U(\mathbf{u}, \mathbf{v}) \leq 0 | H_1)$$

$$Z_B^{(2)} = P(\sup_{v \geq 0} \inf_{u \geq 0} U(\mathbf{u}, \mathbf{v}) \leq 0 | H_1)$$

$$Z_B^{(3)} = P(\frac{1}{2}(\inf_{u \geq 0} \sup_{v \geq 0} U(\mathbf{u}, \mathbf{v}) \sup_{v \geq 0} \inf_{u \geq 0} U(\mathbf{u}, \mathbf{v})) \leq 0 | H_1).$$

Corollary 3.3.2. *The asymptotic coverage of one-sided immersion-posterior credible intervals is given by*

$$P_0(g_0(\mathbf{x}_0) \leq Q_{n,\gamma}^{(l)}) \rightarrow P(Z_B^{(l)} \leq 1 - \gamma), \quad (3.5)$$

and that of two-sided immersion-posterior credible intervals are

$$P_0(g_0(\mathbf{x}_0) \in [Q_{n,(1-\gamma)/2}^{(l)}, Q_{n,\gamma/2}^{(l)}]) \rightarrow P(\gamma/2 \leq Z_B^{(l)} \leq 1 - \gamma/2), \quad (3.6)$$

for $l = 1, 2, 3$.

Thus the coverage depends only on the distributions of $Z_B^{(l)}$, $l = 1, 2, 3$, which do not involve any model parameters. The distribution functions of $Z_B^{(l)}$ at a set of most commonly used points under several local smoothness levels of g_0 at \mathbf{x}_0 were tabulated in 2.5.1. From their tables, we can conclude that if $\boldsymbol{\eta} = (1, 1)$, the asymptotic coverage of the three 95% one-sided credible intervals considered here are 96.6%, 97.5%, and 96.8%, respectively. To target a specific asymptotic frequentist coverage, we tabulated values of the inverted distribution functions of $Z_B^{(1)}$, $Z_B^{(2)}$ and $Z_B^{(3)}$ at several points and back-calculated the credibility level necessary for certain standard confidence levels in 2.5.1. For example, if we use the symmetrized immersion-posterior quantile intervals with the targeted coverage of 95%, the required two endpoints are $Q_{n,0.959}^{(3)}$ and $Q_{n,0.041}^{(3)}$, resulting in a shorter interval than the nominal 95% credible interval.

3.4 Simulation

In this section, we shall carry out two simulation studies to investigate the approach we proposed in previous sections in finite sample settings. We shall look at both the global performance with respect to the \mathbb{L}_1 -metric and the pointwise inference in terms of the frequentist coverage of posterior quantile-based credible interval.

3.4.1 Global deviation of immersion posteriors

We consider the following four data-generating probability density functions throughout this section.

1. $g_1(x, y) = 9(1-x)^2(1-y)^2$;
2. $g_2(x, y) = 2.25\sqrt{(1-x)(1-y)}$;
3. $g_3(x, y) = 4[(1 + e^{12(x-0.5)})(1 + e^{12(y-0.5)})]^{-1}$;
4. $g_4(x, y) = 4[(1 + e^{4(x-0.5)})(1 + e^{4(y-0.5)})]^{-1}$.

Their function graphs are pictured in Figure 3.1. For each density function, we generate i.i.d. samples of sizes, $n = 500, 1000, 2000, 5000$, and 10000 . For density functions g_1 and g_2 , we obtain the data by independently sampling from Beta distributions for each of the two coordinates x and y . For density functions g_3 and g_4 , we use the rejection sampling to generate the samples. To implement the Bayesian procedure, we set $J = \lceil 2n^{1/4} \rceil$, which is the rate-optimal choice for the \mathbb{L}_1 -posterior contraction rates according to Theorem 3.2.2. All hyperparameters α_j in the Dirichlet priors are set to 1 without any prior information. The unrestricted posterior is immediate by conjugacy. Next, we calculate an \mathbb{L}_1 -projection of the unrestricted posterior samples onto the multivariate monotone decreasing function class. We implement the \mathbb{L}_1 -projection algorithm from Section 4 of Stout (2013). This algorithm can isotone a sample of size s within $O(s \log s)$ time for data on a 2-dimensional grid. The immersion posterior sample

of a multivariate probability density function can then be obtained by a renormalization step following the \mathbb{L}_1 -projection.

From the proof of Theorem 3.2.2, we know that the immersion posterior inherits the same contraction rate as the unrestricted posterior. In addition to evaluating the proposed methods when the sample size is finite, we also compare the \mathbb{L}_1 -metric between the immersion posterior samples and the true data generating density function and the \mathbb{L}_1 -metric between the unrestricted posterior samples and the true data generating density function in the simulation studies.

For each data set, we generate 1000 posterior samples and calculate 1000 immersion posterior samples correspondingly. For each posterior density function, we numerically calculate the \mathbb{L}_1 -metric to the true probability density. We repeat this process for 100 times.

We summarize our computation results in Table 3.1. We reported the average of posterior mean and posterior standard deviation of \mathbb{L}_1 -metric in terms of unrestricted posterior and immersion posterior. In Table 3.1, the average of posterior mean is denoted by L_1 and L_1^* respectively for unrestricted and immersion posterior, and the average of posterior standard deviation is denoted by SD and SD^* . Marked in parentheses are standard deviations of posterior means and posterior standard deviations over 100 replicates for each setting.

From our simulation results, immersion posterior improves the performance of the conventional unrestricted posterior by leveraging the functional shape information at a later stage by a monotone mapping. The improvement is consistent over all functions concerned here and all sample sizes in terms of the \mathbb{L}_1 -metric and its posterior variation.

It may be noted that as no other test for multivariate monotone density, Bayesian or frequentist, exists in the literature, we do not compare our test with any other procedure.

3.4.2 Coverage of credible intervals

In this part, we conduct simulation studies to investigate the frequentist coverage of credible intervals based on the immersion posterior quantiles. The data is generated from the same

Table 3.1: \mathbb{L}_1 -metric of unrestricted posterior and immersion posterior.

g_0	n	L_1	SD	L_1^*	SD^*
g_1	500	0.397(0.015)	0.029(1.091)	0.264(0.011)	0.022(0.002)
	1000	0.337(0.011)	0.021(0.720)	0.220(0.008)	0.016(0.001)
	2000	0.280(0.009)	0.015(0.483)	0.182(0.006)	0.011(0.001)
	5000	0.217(0.005)	0.010(0.284)	0.143(0.003)	0.007(0.000)
	10000	0.180(0.004)	0.007(0.188)	0.119(0.002)	0.005(0.000)
g_2	500	0.429(0.014)	0.028(0.785)	0.185(0.012)	0.026(0.004)
	1000	0.375(0.012)	0.021(0.566)	0.156(0.011)	0.019(0.003)
	2000	0.320(0.007)	0.015(0.422)	0.130(0.007)	0.013(0.002)
	5000	0.254(0.006)	0.010(0.260)	0.104(0.004)	0.008(0.001)
	10000	0.214(0.004)	0.007(0.160)	0.087(0.003)	0.006(0.000)
g_3	500	0.393(0.014)	0.030(1.120)	0.256(0.015)	0.031(0.003)
	1000	0.336(0.012)	0.022(0.603)	0.208(0.012)	0.022(0.002)
	2000	0.277(0.007)	0.016(0.381)	0.166(0.008)	0.015(0.001)
	5000	0.213(0.006)	0.010(0.289)	0.125(0.005)	0.010(0.001)
	10000	0.177(0.004)	0.007(0.187)	0.101(0.004)	0.007(0.001)
g_4	500	0.418(0.017)	0.029(0.986)	0.199(0.011)	0.024(0.002)
	1000	0.364(0.009)	0.021(0.546)	0.168(0.009)	0.017(0.002)
	2000	0.311(0.008)	0.015(0.377)	0.142(0.006)	0.013(0.001)
	5000	0.244(0.005)	0.010(0.252)	0.112(0.004)	0.008(0.001)
	10000	0.207(0.004)	0.007(0.180)	0.094(0.003)	0.006(0.000)

set of probability density functions considered in the last section. We consider five different sample sizes, $n = 500, 1000, 2000$, and 5000 . The point on which we will make inference is $x_0 = (0.5, 0.5)$. According to the sample size, we take $J = \lceil n^{1/4} \sqrt{\log n} \rceil$. The hyperparameters in the Dirichlet priors are taken to be 1. All three immersion maps are considered. They are indicated as "min-max", "max-min", and "ave" in the following tables. We considered the two-sided credible interval for $g_0(x_0)$ with four credibilities, 0.99, 0.95, 0.90, and 0.80. To recalibrate the credible interval to the right asymptotic coverage, we use the corresponding quantiles of $Z_B^{(3)}$ from Table 2.4. Note that the distribution function of $Z_B^{(3)}$ is symmetric about 1/2. Thus it is possible to get both the upper and lower quantiles from the table. For example, we use 0.990 and 0.010, instead of 0.995 and 0.005, immersion quantiles to construct the credible sets targeting the coverage 99%.

We repeated generating the data in each setting for 1000 times and calculate the frequency

of credible intervals including the true parameter. The results are summarized in Table 3.2–3.5. When the sample size is larger, the recalibrated credible intervals behave very well, with coverage staying closer to the target while the credible intervals based on raw quantiles are more conservative. However, when the sample size is moderate, the performance of all methods varies among different probability density functions. For instance, for function g_1 , as the function value at x_0 is relatively closer to 0, the proposed methods result in poor coverage when $n = 500$, but soon the coverage gets much better when $n = 1000$. In all the tables, C and L denote the coverage and average length of credible intervals, rounded to two digits.

Table 3.2: Coverage and length of credible intervals for $g_1(x_0)$

n	immersion maps	credibility					
		.99		.95		.90	
		C	L	C	L	C	L
500	min-max	0.91	0.86(0.12)	0.70	0.66(0.10)	0.56	0.56(0.09)
	max-min	0.92	0.85(0.13)	0.74	0.66(0.10)	0.60	0.56(0.09)
	average	0.91	0.85(0.12)	0.71	0.66(0.10)	0.58	0.56(0.09)
	adjusted	0.85	0.83(0.12)	0.61	0.63(0.10)	0.48	0.53(0.09)
1000	min-max	1.00	0.62(0.09)	0.97	0.48(0.07)	0.94	0.41(0.06)
	max-min	1.00	0.62(0.09)	0.98	0.48(0.08)	0.95	0.41(0.07)
	average	1.00	0.62(0.09)	0.97	0.48(0.08)	0.94	0.41(0.07)
	adjusted	0.99	0.60(0.09)	0.95	0.46(0.07)	0.90	0.39(0.06)
2000	min-max	1.00	0.54(0.08)	0.98	0.42(0.06)	0.93	0.36(0.06)
	max-min	1.00	0.54(0.08)	0.98	0.42(0.06)	0.95	0.35(0.06)
	average	1.00	0.54(0.08)	0.98	0.42(0.06)	0.94	0.35(0.06)
	adjusted	0.99	0.52(0.08)	0.95	0.40(0.06)	0.89	0.34(0.05)
5000	min-max	1.00	0.44(0.06)	0.98	0.35(0.05)	0.93	0.29(0.04)
	max-min	1.00	0.44(0.06)	0.98	0.34(0.05)	0.95	0.29(0.04)
	average	1.00	0.44(0.06)	0.98	0.34(0.05)	0.94	0.29(0.04)
	adjusted	0.99	0.43(0.06)	0.95	0.33(0.05)	0.90	0.27(0.04)

Table 3.3: Coverage and length of credible intervals for $g_2(\mathbf{x}_0)$

n	immersion maps	credibility					
		.99		.95		.90	
		C	L	C	L	C	L
500	min-max	1.00	0.59(0.08)	0.99	0.45(0.07)	0.97	0.38(0.06)
	max-min	1.00	0.59(0.08)	0.99	0.45(0.07)	0.97	0.38(0.06)
	average	1.00	0.58(0.08)	0.99	0.45(0.07)	0.97	0.38(0.06)
	adjusted	1.00	0.56(0.08)	0.98	0.43(0.06)	0.94	0.36(0.06)
1000	min-max	0.99	0.50(0.07)	0.98	0.39(0.06)	0.94	0.33(0.05)
	max-min	0.99	0.50(0.07)	0.96	0.39(0.06)	0.92	0.33(0.05)
	average	0.99	0.50(0.07)	0.97	0.39(0.06)	0.92	0.33(0.05)
	adjusted	0.98	0.48(0.07)	0.94	0.37(0.05)	0.88	0.31(0.05)
2000	min-max	0.99	0.43(0.06)	0.98	0.34(0.05)	0.94	0.29(0.04)
	max-min	0.99	0.43(0.06)	0.96	0.34(0.05)	0.91	0.28(0.04)
	average	0.99	0.43(0.06)	0.97	0.33(0.05)	0.93	0.28(0.04)
	adjusted	0.99	0.41(0.06)	0.94	0.32(0.05)	0.89	0.27(0.04)
5000	min-max	0.99	0.36(0.05)	0.97	0.28(0.04)	0.92	0.23(0.03)
	max-min	0.99	0.36(0.05)	0.95	0.28(0.04)	0.91	0.23(0.03)
	average	0.99	0.35(0.05)	0.96	0.27(0.04)	0.92	0.23(0.03)
	adjusted	0.99	0.34(0.05)	0.93	0.26(0.04)	0.87	0.22(0.03)

3.5 Proofs

Proof of Lemma 3.2.1. For part (i), let $\mathbf{x} \in I_j$. Then, as h is constant over I_j , by the definition of s_J

$$|s_J(\mathbf{x}) - h(\mathbf{x})|^p = \left| \left(\prod_{k=1}^d J_k \right) \int_{I_j} (s - h) \right|^p \leq \left(\prod_{k=1}^d J_k \right) \int_{I_j} |s - h|^p, \quad (3.7)$$

by Jensen's inequality. By taking integral on both sides of (3.7) over I_j , we have $\int_{I_j} |s_J - h|^p \leq \int_{I_j} |s - h|^p$. Part (i) now follows then by summing over all j on both sides of the inequality.

For parts (ii) and (iii), if s is a multivariate monotone nonincreasing function, then $\{b_j\}$ is a monotonically decreasing array in terms of the natural partial order on the indices since, by

Table 3.4: Coverage and length of credible intervals for $g_3(\mathbf{x}_0)$

n	immersion maps	credibility					
		.99		.95		.90	
		C	L	C	L	C	L
500	min-max	0.95	1.27(0.16)	0.85	1.00(0.14)	0.74	0.85(0.12)
	max-min	0.96	1.26(0.16)	0.86	0.99(0.14)	0.76	0.84(0.13)
	average	0.96	1.26(0.16)	0.86	0.99(0.14)	0.74	0.84(0.13)
	adjusted	0.93	1.22(0.16)	0.78	0.95(0.14)	0.67	0.80(0.12)
1000	min-max	1.00	1.00(0.13)	0.98	0.78(0.11)	0.93	0.66(0.10)
	max-min	1.00	0.99(0.13)	0.97	0.78(0.11)	0.93	0.66(0.10)
	average	1.00	0.99(0.13)	0.98	0.78(0.11)	0.93	0.66(0.10)
	adjusted	0.99	0.96(0.13)	0.94	0.75(0.11)	0.89	0.63(0.10)
2000	min-max	1.00	0.89(0.11)	0.97	0.70(0.10)	0.94	0.59(0.09)
	max-min	1.00	0.88(0.11)	0.97	0.69(0.10)	0.93	0.59(0.09)
	average	1.00	0.88(0.11)	0.97	0.69(0.10)	0.93	0.59(0.09)
	adjusted	0.99	0.86(0.11)	0.94	0.66(0.09)	0.90	0.56(0.08)
5000	min-max	1.00	0.76(0.09)	0.98	0.59(0.08)	0.93	0.50(0.07)
	max-min	1.00	0.75(0.09)	0.97	0.59(0.08)	0.92	0.50(0.07)
	average	1.00	0.75(0.09)	0.98	0.59(0.08)	0.93	0.50(0.07)
	adjusted	0.99	0.73(0.09)	0.95	0.56(0.08)	0.87	0.47(0.07)

the translation of s ,

$$b_{j_1} = \int_{I_{j_1}} s(\mathbf{x})d\mathbf{x} \geq \int_{I_{j_1}} s(\mathbf{x} + (j_2 - j_1)/J)d\mathbf{x} = \int_{I_{j_2}} s(\mathbf{x})d\mathbf{x} = b_{j_2},$$

when $j_2 \succeq j_1$. Additionally, it is clear that if $s \geq 0$, $b_j \geq 0$ for all j , and $\sum_{j \in [1:J]} (\prod_{k=1}^d J_k) b_j = \int_{[0,1]^d} s$. \square

The main ideas of the proofs of Theorems 3.2.2, 3.2.3 and 3.2.4 are very similar to their univariate counterpart in Chakraborty and Ghosal (2022), but the associated bounds will have to be reestablished in the multidimensional case. For the sake of self-contentedness, we present the proofs in brief.

Proof of Theorem 3.2.2. Let \tilde{g} be an \mathbb{L}_1 -projection of g to \mathcal{F}^* , from which we obtain the nor-

Table 3.5: Coverage and length of credible intervals for $g_4(\mathbf{x}_0)$

n	immersion maps	credibility					
		.99		.95		.90	
		C	L	C	L	C	L
500	min-max	0.99	0.76(0.10)	0.96	0.59(0.09)	0.92	0.50(0.08)
	max-min	1.00	0.76(0.11)	0.96	0.59(0.09)	0.94	0.50(0.08)
	average	1.00	0.76(0.10)	0.96	0.59(0.09)	0.93	0.50(0.08)
	adjusted	0.99	0.73(0.10)	0.94	0.56(0.08)	0.87	0.47(0.07)
1000	min-max	1.00	0.63(0.09)	0.98	0.49(0.07)	0.94	0.42(0.06)
	max-min	1.00	0.63(0.09)	0.97	0.49(0.07)	0.93	0.41(0.06)
	average	1.00	0.63(0.09)	0.97	0.49(0.07)	0.94	0.41(0.06)
	adjusted	0.99	0.60(0.08)	0.95	0.46(0.07)	0.89	0.39(0.06)
2000	min-max	1.00	0.55(0.07)	0.98	0.43(0.06)	0.95	0.36(0.05)
	max-min	1.00	0.55(0.07)	0.97	0.42(0.06)	0.94	0.36(0.06)
	average	1.00	0.54(0.07)	0.98	0.42(0.06)	0.94	0.36(0.05)
	adjusted	1.00	0.52(0.07)	0.96	0.40(0.06)	0.90	0.34(0.05)
5000	min-max	1.00	0.46(0.06)	0.99	0.36(0.05)	0.97	0.30(0.05)
	max-min	1.00	0.46(0.06)	0.98	0.36(0.05)	0.96	0.30(0.05)
	average	1.00	0.46(0.06)	0.99	0.36(0.05)	0.96	0.30(0.04)
	adjusted	1.00	0.44(0.06)	0.97	0.34(0.05)	0.93	0.28(0.04)

malized g^* . Since $\|g^* - g_0\|_1 \leq \|\tilde{g} - g_0\|_1 + \|\tilde{g} - g^*\|_1$, and

$$\|\tilde{g} - g^*\|_1 = \int \left| \tilde{g} - \frac{\tilde{g}}{\int \tilde{g}} \right| = \int \tilde{g} \left| 1 - \frac{1}{\int \tilde{g}} \right| = \left| \int \tilde{g} - \int g_0 \right| \leq \|\tilde{g} - g_0\|_1,$$

we have $\|g^* - g_0\|_1 \leq 2\|\tilde{g} - g_0\|_1$. Moreover, $\|\tilde{g} - g_0\|_1 \leq \|\tilde{g} - g\|_1 + \|g - g_0\|_1 \leq 2\|g - g_0\|_1$ by the triangle inequality and the definition of projection. Combining $\|g^* - g_0\|_1 \leq 4\|g - g_0\|_1$, hence the immersion posterior contraction rate is inherited from the unrestricted posterior contraction rate.

Let $g_{0,J} = J^d \sum_{j \in [1:J]} \theta_{0,j} \mathbb{1}_{I_j}$, where $\theta_{0,j} = P_0(I_j)$. For every j and $\mathbf{x} \in I_j$, $|g_{0,J}(\mathbf{x}) - g_0(\mathbf{x})| = |J^d \int_{I_j} g_0 - g_0(\mathbf{x})| \leq g_0((j-1)/J) - g_0(j/J)$ as g_0 is coordinate-wise decreasing. Then, splitting the integral over each I_j ,

$$\|g_{0,J} - g_0\|_1 \leq J^{-d} \sum_{j \in [1:J]} [g_0((j-1)/J) - g_0(j/J)] \leq d J^{-1} (g_0(\mathbf{0}) - g_0(\mathbf{1})),$$

by the telescoping property of the series corresponding to the points $\{\dots, j, j+1, j+2 \cdot 1, \dots\}$, and there are no more than dJ^{d-1} such series in the above summation.

Since $\|g - g_0\|_1 \leq \|g - g_{0,j}\|_1 + \|g_{0,j} - g_0\|_1$, we only have to show that for every $M_n \rightarrow \infty$,

$$E_0\Pi(\|g - g_{0,j}\|_1 \geq M_n \sqrt{J^d/n} | \mathbb{D}_n) \rightarrow 0. \quad (3.8)$$

As the \mathbb{L}_2 -norm dominates the \mathbb{L}_1 -norm, it suffices to show that

$$E_0\Pi(\|g - g_{0,j}\|_2 \geq M_n \sqrt{J^d/n} | \mathbb{D}_n) \rightarrow 0.$$

Observing that $\|g - g_{0,j}\|_2^2 = J^d \sum_{j \in [1:J]} |\theta_j - \theta_{0,j}|^2$, it is enough to verify the bounds

$$\sum_{j \in [1:J]} \text{Var}(\theta_j | \mathbb{D}_n) = O_p(n^{-1}), \quad \sum_{j \in [1:J]} |E(\theta_j | \mathbb{D}_n) - \theta_{0,j}|^2 = O_p(n^{-1})$$

in view of the Markov inequality and a standard variance-bias decomposition. Since

$$\text{Var}(\theta_j | \mathbb{D}_n) \leq (\alpha_j + N_j) / (\alpha + n)^2,$$

the first part of the assertion is immediately obtained as $\alpha_j \leq a_1 < \infty$, $J^d \leq n$ and $\sum_{j \in [1:J]} N_j = n$. For the second claim, note that $E(\theta_j | \mathbb{D}_n) = (\alpha_j + N_j) / (\alpha + n)$ and $N_j \sim \text{Bin}(n, \theta_{0,j})$, where $\alpha = \sum_{j \in [1:J]} \alpha_j \leq J^d a_1$. Thus

$$\begin{aligned} \text{Var}(E(\theta_j | \mathbb{D}_n)) &= n \theta_{0,j} (1 - \theta_{0,j}) / (\alpha_0 + n)^2 \leq \theta_{0,j} / n \leq J^{-d} g_0(\mathbf{0}) / n, \\ (E_0(E(\theta_j | \mathbb{D}_n)) - \theta_{0,j})^2 &= (\alpha_j - \alpha_0 \theta_{0,j})^2 / (\alpha_0 + n)^2 \leq 2a_1^2 (1 + g_0^2(\mathbf{0})) / n^2, \end{aligned}$$

as $\theta_{0,j} \leq J^{-d} g_0(\mathbf{0})$. Now summing over $j \in [1 : J]$, we obtain the result. \square

Proof of Theorem 3.2.3. (i) Since $d_1(g, \mathcal{F}) = \|g - \tilde{g}\|_1 \leq \|g - g_0\|_1$, the first assertion follows from Theorem 3.2.2.

(ii) We claim that $d_1(g_0, \mathcal{F}^*) > 0$. If not, for every $\eta > 0$, there exists $f \in \mathcal{F}^*$ such that $\|f - g_0\| \leq \eta/2$. Then $f_+ \in \mathcal{F}^*$ and $\|f_+ - g_0\|_1 \leq \|f - g_0\|_1$, where f_+ is the positive part of f . Further, $\bar{f} = f_+/\|f_+\|_1 \in \mathcal{G}^*$ and, since $\bar{f} - f_+ = f_+(1 - \|f_+\|_1)/\|f_+\|_1$ and $\|g_0\|_1 = 1$,

$$\|\bar{f} - g_0\|_1 \leq \|\bar{f} - f_+\|_1 + \|f_+ - g_0\|_1 \leq |\|g_0\|_1 - \|f_+\|_1| + \|f_+ - g_0\|_1$$

is bounded by $2\|f_+ - g_0\|_1 \leq 2\|f - g_0\|_1 \leq \eta$, contradicting the assumption that g_0 does not belong to the closure of \mathcal{G}^* . Thus $d_1(g_0, \mathcal{F}^*) > 0$. Hence by the triangle inequality and the definition of the \mathbb{L}_1 -projection,

$$d_1(g, \mathcal{F}^*) \geq \|g_0 - \bar{g}\|_1 - \|g - g_0\|_1 \geq d_1(g_0, \mathcal{F}^*) - \|g - g_0\|_1. \quad (3.9)$$

To conclude the consistency of the test, it now suffices to show that the posterior distribution of g is consistent at g_0 with respect to the \mathbb{L}_1 -metric. By the martingale convergence theorem, we have $\|g_0 - g_{0,J}\|_1 \rightarrow 0$ as $J \rightarrow \infty$. Observing (3.8) does not require monotonicity, posterior consistency at g_0 follows.

(iii) Uniformly over $g_0 \in \mathcal{H}(\alpha, L)$, we have $\|g_0 - g_{0,J}\|_1 \lesssim J^{-\alpha}$. Combined with (3.8), this gives the posterior contraction rate $\max\{J^{-\alpha}, \sqrt{J^d/n}\}$ at g_0 , which is optimized to $n^{-\alpha/(2+d)}$ for the choice $J \asymp n^{1/(2+d)}$. As in the proof of part (ii), we can conclude that $d_1(g_0, \mathcal{F}^*) \geq \rho_n(\alpha)/2$ if $d_1(g_0, \mathcal{G}^*) > \rho_n(\alpha)$. By (3.9), $\Pi(d_1(g, \mathcal{F}^*) \leq M_n n^{-1/(2+d)} | \mathbb{D}_n)$ is bounded by $\Pi(\|g - g_0\|_1 \geq d_1(g_0, \mathcal{F}^*) - M_n n^{-1/(2+d)} | \mathbb{D}_n)$. When $\alpha < 1$, $n^{-\alpha/(d+2)} \gg M_n n^{-1/(d+2)}$ and hence the bound goes to 0 in P_0 -probability with the choice of $\rho_n(\alpha) = C n^{-\alpha/(2+d)}$ for any fixed $C > 0$. When $\alpha = 1$, the bound is at most $\Pi(\|g - g_0\|_1 \geq (C/2 - 1)M_n n^{-1/(2+d)} | \mathbb{D}_n)$, which converges to zero in P_0 -probability. \square

Proof of Theorem 3.2.4. (i) By the definition of the piece-wise constant approximation, $g_{0,J} \in \mathcal{G}^* \subset \mathcal{F}^*$ if $g_0 \in \mathcal{G}^*$. Then we have $d_1(g, \mathcal{F}^*) \leq \|g - g_{0,J}\|_1$ for every $J > 0$. Then for part (i), we

assert that

$$P_0(\Pi(\|g - g_{0,J}\|_1 > M_0 \sqrt{(J^d \log n)/n} | \mathbb{D}_n) \leq 1 - \gamma) \rightarrow 1. \quad (3.10)$$

To this end, we shall show that for $J_n^d \lesssim n$,

$$(a) \quad \Pi(\|g - g_{0,J}\|_1 > M_0 \sqrt{J^d \log n/n}, J \leq J_n | \mathbb{D}_n) \rightarrow_{P_0} 0$$

$$(b) \quad \Pi(J > J_n | \mathbb{D}_n) \rightarrow_{P_0} 0 \text{ for } J_n \asymp (n/\log n)^{1/(d+1)}.$$

As the \mathbb{L}_2 -metric dominates the \mathbb{L}_1 -metric on $[0, 1]^d$, the posterior probability in (a) is bounded by $\sum_{J=1}^{J_n} A_{n,J}(M_0) \pi(J | \mathbb{D}_n)$, where $A_{n,J}(M_0) = \Pi(\int |g - g_{0,J}|^2 > M_0^2 (J^d \log n)/n | J, \mathbb{D}_n)$. By Markov's inequality

$$A_{n,J}(M_0) \leq \frac{n}{M_0^2 J^d \log n} \left[\sum_{j \in [1:J]} \text{Var}(\theta_j | \mathbb{D}_n) + \sum_{j \in [1:J]} (\mathbb{E}(\theta_j | \mathbb{D}_n) - \theta_{0,j})^2 \right]. \quad (3.11)$$

Using the bounds established in the proof of (3.8), the first term inside the bracket is $O(n^{-1})$ uniformly for all J , while we can decompose $\mathbb{E}(\theta_j | \mathbb{D}_n) - \theta_{0,j}$ as the sum of $(\alpha_j - \alpha N_j/n)/(\alpha + n)$ and $N_j/n - \theta_{0,j}$ with $\alpha := \sum_{j \in [1:J]} \alpha_j \leq a_1 J^d$. The sum of squares of the first term over $j \in [1:J]$ is bounded by

$$\frac{2[\sum_{j \in [1:J]} \alpha_j^2 + \alpha^2 \sum_{j \in [1:J]} N_j^2/n^2]}{(\alpha + n)^2} \leq \frac{2[a_1^2 J^d + (a_1 J^d)^2 n \sum_{j \in [1:J]} N_j/n^2]}{(\alpha + n)^2} \lesssim \frac{J^d}{n}$$

uniformly for all j and all J . Using Bennett's inequality (cf. Proposition A.6.2 of van der Vaart and Wellner (1996)) and summing over j , we conclude that simultaneously for all $J \leq J_n$, $\max\{|N_j/n - \theta_{0,j}| : j \in [1:J]\} \lesssim \sqrt{(\log n)/(nJ^d)}$ if $J_n^d \log n \lesssim n$. This gives that the expression in (3.11) is bounded by a constant multiple of M_0^{-2} , and hence can be made arbitrarily small by choosing M_0 sufficiently large. Thus claim (a) follows.

To establish (b), we apply the general theory of posterior contraction in Ghosal and van der Vaart (2017). Because g_0 is bounded below, any g sufficiently uniformly close is also bounded

below, and so g_0/g is uniformly bounded. Using standard relations between Kullback-Leibler divergences, the Hellinger distance and the \mathbb{L}_1 -distance, and recalling that the piece-wise constant approximation $g_{0,J}$ is $O(J^{-1})$ close to g_0 in \mathbb{L}_1 , to estimate the prior ϵ_n -ball probability necessary for the application of the general theory, it suffices to lower bound $\Pi(J = J_n^*)\Pi(\sum_{j \in [1:J_n^*]} |\theta_j - \theta_{0,j}| \leq \epsilon_n)$, where $J_n^* \asymp \epsilon_n^{-1}$. By the assumption on the prior on J and the Dirichlet small ball probability estimate (cf. Lemma G.13 of Ghosal and van der Vaart (2017)), a lower bound is $e^{-c(J_n^*)^d \log n}$ for some $c > 0$ provided that ϵ_n is lower bounded by some negative power of n . Thus (b) holds if J_n is chosen a sufficiently large constant multiple of J_n^* by Theorem 8.20 of Ghosal and van der Vaart (2017). Now, with these estimates, proceeding as in the proof of Theorem 3.2.3, the conclusion in (i) can be established.

For part (ii), in view of the martingale convergence theorem, for any given $\epsilon > 0$, $\|g_0 - g_{0,J}\|_1 < \epsilon$ when $J \geq J_0$, say. Then we proceed as in part (i) with a fixed small ϵ , $J_n^* = J_0$ and establish that (a) and (b) hold for J_n a multiple of $(n/\log n)^{1/d}$, where the constant of proportionality is taken to be sufficiently small depending on ϵ .

The proof of part (iii) also proceeds similarly after establishing the posterior contraction rate $\epsilon_n \asymp (n/\log n)^{-\alpha/(2\alpha+d)}$ by standard arguments for Hölder functions. The estimates used in the proof imply that (a) and (b) hold for J_n a sufficiently large constant multiple of $(n/\log n)^{1/(2\alpha+d)}$. □

The proof of Theorem 3.3.1 is long, so we separate it into several key steps. These are stated as separate lemmas below, under the setup and assumptions of Theorem 3.3.1.

For $\mathbf{t} \in \mathbb{R}^d$, let $j(\mathbf{t}) = [(\mathbf{x}_0 + \mathbf{t} \circ \mathbf{r}_n) \circ \mathbf{J}]$. Then we can rewrite $g^*(\mathbf{x}_0)$ as

$$g^*(\mathbf{x}_0) = \min_{\mathbf{u} \geq \mathbf{0}} \max_{\mathbf{v} \geq \mathbf{0}} \frac{\sum_{j_0 \in [j(-\mathbf{u}):j(\mathbf{v})]} \theta_j}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}},$$

and the localized version of $g^*(\mathbf{x}_0)$ with some positive constant c and γ as

$$g_c^*(\mathbf{x}_0) = \min_{c^{-\gamma} \mathbf{1} \leq \mathbf{u} \leq c \mathbf{1}} \max_{c^{-\gamma} \mathbf{1} \leq \mathbf{v} \leq c \mathbf{1}} \frac{\sum_{j_0 \in [j(-\mathbf{u}):j(\mathbf{v})]} \theta_j}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}.$$

We denote

$$\begin{aligned}
W_n^* &= \omega_n^{-1}(\mathbf{g}^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0)), \\
W_{n,c}^* &= \omega_n^{-1}(\mathbf{g}_c^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0)), \\
W_c &= \inf_{c^{-\gamma} \mathbf{1} \leq \mathbf{u} \leq c \mathbf{1}} \sup_{c^{-\gamma} \mathbf{1} \leq \mathbf{v} \leq c \mathbf{1}} \left\{ \frac{\sqrt{\mathbf{g}_0(\mathbf{x}_0)} H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\sqrt{\mathbf{g}_0(\mathbf{x}_0)} H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} \right. \\
&\quad \left. + \sum_{m \in M} \frac{\partial^m \mathbf{g}_0(\mathbf{x}_0)}{(m+1)!} \prod_{k=1}^d \frac{v_k^{m_k+1} - (-u_k)^{m_k+1}}{u_k + v_k} \right\}, \\
W &= \inf_{\mathbf{u} \geq \mathbf{0}} \sup_{\mathbf{v} \geq \mathbf{0}} \left\{ \frac{\sqrt{\mathbf{g}_0(\mathbf{x}_0)} H_1(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} + \frac{\sqrt{\mathbf{g}_0(\mathbf{x}_0)} H_2(\mathbf{u}, \mathbf{v})}{\prod_{k=1}^d (u_k + v_k)} \right. \\
&\quad \left. + \sum_{m \in M} \frac{\partial^m \mathbf{g}_0(\mathbf{x}_0)}{(m+1)!} \prod_{k=1}^d \frac{v_k^{m_k+1} - (-u_k)^{m_k+1}}{u_k + v_k} \right\}.
\end{aligned}$$

The desired weak convergence of random measures on \mathbb{R} follows from Lemma A.2.1.

We apply the Gamma representation for the Dirichlet distribution to the unrestricted posterior $\boldsymbol{\theta}$ given the data \mathbb{D}_n . Let $V_j | \mathbb{D}_n \sim \text{Ga}(\alpha_j + N_j, 1)$, mutually independent. Then θ_j given the data \mathbb{D}_n is distributed as $V_j / \sum_{l \in [1:J]} V_l$. Let $\boldsymbol{\alpha} = \sum_{l \in [1:J]} \alpha_l$. It follows immediately that $\sum_{l \in [1:J]} V_l \sim \text{Ga}(\boldsymbol{\alpha} + n, 1)$, and $\mathbb{E}(\sum_{l \in [1:J]} V_l) = \text{Var}(\sum_{l \in [1:J]} V_l) = \boldsymbol{\alpha} + n$.

We decompose

$$\begin{aligned}
&\omega_n^{-1}(\mathbf{g}^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0)) \\
&= \min_{\mathbf{u} \geq \mathbf{0}} \max_{\mathbf{v} \geq \mathbf{0}} \{A_{n,1}(\mathbf{u}, \mathbf{v}) + A_{n,2}(\mathbf{u}, \mathbf{v}) + B_{n,1}(\mathbf{u}, \mathbf{v}) + B_{n,2}(\mathbf{u}, \mathbf{v})\},
\end{aligned}$$

where

$$\begin{aligned}
A_{n,1}(\mathbf{u}, \mathbf{v}) &= \omega_n^{-1} \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} (V_j - \mathbb{E}(V_j | \mathbb{D}_n)) / \sum_l V_l}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}, \\
A_{n,2}(\mathbf{u}, \mathbf{v}) &= \omega_n^{-1} \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} (\mathbb{E}(V_j | \mathbb{D}_n) / \sum_l V_l - N_j/n)}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}, \\
B_{n,1}(\mathbf{u}, \mathbf{v}) &= \omega_n^{-1} \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} (N_j - \mathbb{E}_0(N_j)) / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}, \\
B_{n,2}(\mathbf{u}, \mathbf{v}) &= \omega_n^{-1} \left(\frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} \mathbb{E}_0(N_j) / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} - g_0(\mathbf{x}_0) \right).
\end{aligned}$$

Lemma 3.5.1. *The conditional distribution of stochastic process $Y_n(\mathbf{u}, \mathbf{v}; V) = \omega_n \sum_{[j(-\mathbf{u}):j(\mathbf{v})]} (V_j - \mathbb{E}(V_j | \mathbb{D}_n))$ given \mathbb{D}_n converges weakly to the distribution of $\sqrt{g_0(\mathbf{x}_0)} H_2(\mathbf{u}, \mathbf{v})$ in $\mathbb{L}_\infty([0, c] \times [0, c])$ in \mathbb{P}_0 -probability.*

Proof. Conditional on \mathbb{D}_n , the expectation of Y_n is zero and the variance is given by

$$\text{Var}(Y_n(\mathbf{u}, \mathbf{v}; V) | \mathbb{D}_n) = \omega_n^2 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} (\alpha_j + N_j).$$

By Assumption 1, we have

$$\begin{aligned}
\omega_n^2 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} \alpha_j &\leq a_1 \cdot \omega_n^2 \prod_{k=1}^d (r_{n,k} u_k + r_{n,k} v_k + 2J_k^{-1}) J_k \\
&= a_1 \cdot n^{-1} \prod_{k=1}^d J_k \prod_{k=1}^d (u_k + v_k + 2r_{n,k}^{-1} J_k^{-1}),
\end{aligned}$$

as $\omega_n^2 \prod_{k=1}^d r_{n,k} = n^{-1}$. By noting that $\prod_{k=1}^d J_k \ll n \omega_n \ll n$, $r_{n,k}^{-1} \ll J_k$, and $u_k, v_k \leq c$ for every $1 \leq k \leq d$, it follows that $\omega_n^2 \sum_{[j(-\mathbf{u}):j(\mathbf{v})]} \alpha_j \rightarrow 0$.

Since g_0 is differentiable, and hence is continuous at \mathbf{x}_0 ,

$$\begin{aligned} \mathbb{E}_0(\omega_n^2 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j) &= n \omega_n^2 \int_{\cup \{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}} g_0(\mathbf{x}) d\mathbf{x} \\ &= n \omega_n^2 (g_0(\mathbf{x}_0) + o(1)) \prod_{k=1}^d (r_{n,k} u_k + r_{n,k} v_k + O(J_k^{-1})) \\ &\rightarrow g_0(\mathbf{x}_0) \prod_{k=1}^d (u_k + v_k), \end{aligned}$$

as $\omega_n^2 \prod_{k=1}^d r_{n,k} = n^{-1}$ and $r_{n,k}^{-1} \ll J_k$. Since $\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j$ is binomially distributed, we similarly have

$$\begin{aligned} \text{Var}(\omega_n^2 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j) &\leq n \omega_n^4 \int_{\cup \{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}} g_0(\mathbf{x}) d\mathbf{x} \\ &\lesssim \omega_n^2 g_0(\mathbf{x}_0) \prod_{k=1}^d \max\{u_k + v_k, 2J_k^{-1} r_{n,k}^{-1}\} \rightarrow 0. \end{aligned}$$

Thus $\text{Var}(Y_n(\mathbf{u}, \mathbf{v}; V) | \mathbb{D}_n) \rightarrow_{P_0} g_0(\mathbf{x}_0) \prod_{k=1}^d (u_k + v_k)$.

We shall apply the Lyapunov's central limit theorem using bounds for the fourth moment to show the marginal convergence. Using the moment bound $\mathbb{E}((V_j - \mathbb{E}(V_j | \mathbb{D}_n))^4 | \mathbb{D}_n) \lesssim (\alpha_j + N_j)^2$, and observing that

$$\begin{aligned} \omega_n^4 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} (\alpha_j + N_j)^2 &\lesssim \omega_n^4 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j^2 \\ &\lesssim \omega_n^4 \prod_{k=1}^d (u_k r_{n,k} + v_k r_{n,k}) J_k \left(\frac{n}{\prod_{k=1}^d J_k} \right)^2 \\ &\lesssim \frac{n \omega_n^2}{\prod_{k=1}^d J_k} = \frac{\prod_{k=1}^d r_{n,k}^{-1}}{\prod_{k=1}^d J_k}, \end{aligned}$$

with P_0 -probability tending to one, as $r_{n,k}^{-1} \ll J_k$ and $u_k, v_k \leq c$ for every k . Comparing this with the asymptotic variance, it is clear that the Lyapunov's condition holds. Thus for a fixed (\mathbf{u}, \mathbf{v}) ,

the conditional distribution of $Y_n(\mathbf{u}, \mathbf{v}, V)$ given the data converges weakly to

$$\sqrt{g_0(\mathbf{x}_0) \prod_{k=1}^d (u_k + v_k)} \mathbf{N}(0, 1)$$

in P_0 -probability. The conclusion easily extends to finite dimensional joint distributions by observing that for (\mathbf{u}, \mathbf{v}) and $(\mathbf{u}', \mathbf{v}') \in [0, c\mathbf{1}] \times [0, c\mathbf{1}]$,

$$\begin{aligned} & \mathbb{E}(Y_n(\mathbf{u}, \mathbf{v}; V) \cdot Y_n(\mathbf{u}', \mathbf{v}'; V) | \mathbb{D}_n) \\ &= \omega_n^2 \sum_{j \in [j(-\mathbf{u} \wedge \mathbf{u}'); j(\mathbf{v} \wedge \mathbf{v}')] } \text{Var}(V_j | \mathbb{D}_n) \\ &\rightarrow g_0(\mathbf{x}_0) \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k), \end{aligned}$$

in P_0 -probability.

We now show the asymptotic tightness of the conditional law of the process given by $Y_n(\mathbf{u}, \mathbf{v}; V)$ given the data in $L_\infty([0, c\mathbf{1}] \times [0, c\mathbf{1}])$. We apply Theorem 2.2.4 of van der Vaart and Wellner (1996) with the function $\psi : x \mapsto x^{4d+2}$, and use the metric $\rho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) = \sqrt{\|\mathbf{u} - \mathbf{u}'\| + \|\mathbf{v} - \mathbf{v}'\|}$ on $[0, c\mathbf{1}] \times [0, c\mathbf{1}]$. We first calculate the $(4d+2)$ -th moment of the increments, conditional on the data. By Lemma A.3.4, we have that

$$\begin{aligned} & \mathbb{E}(|Y_n(\mathbf{u}, \mathbf{v}; V) - Y_n(\mathbf{u}', \mathbf{v}'; V)|^{4d+2} | \mathbb{D}_n) \\ &= \omega_n^{4d+2} \mathbb{E}\left(\left(\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})] \Delta [j(-\mathbf{u}'):j(\mathbf{v}')]} (V_j - \mathbb{E}(V_j | \mathbb{D}_n))\right)^{4d+2} | \mathbb{D}_n\right) \\ &\lesssim \omega_n^{4d+2} \left(\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})] \Delta [j(-\mathbf{u}'):j(\mathbf{v}')]} N_j\right)^{2d+1}, \end{aligned} \tag{3.12}$$

with P_0 -probability tending to one, by noting that $\min\{N_j : j \in [1 : \mathbf{J}]\} \rightarrow_{P_0} \infty$ and $\{\alpha_j\}$ is

uniformly bounded. With P_0 -probability tending to one, we have

$$\begin{aligned}
& \sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})] \Delta [j(-\mathbf{u}'); j(\mathbf{v}')] } N_j \\
& \lesssim n \prod_{k=1}^d r_{n,k} \left(\prod_{k=1}^d (u_k \vee u'_k + v_k \vee v'_k) - \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k) \right) \\
& = \omega_n^{-2} \left(\prod_{k=1}^d (u_k \vee u'_k + v_k \vee v'_k) - \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k) \right).
\end{aligned}$$

Thus by Lemma A.3.1, (3.12) is bounded by a constant multiple of $(\|\mathbf{u} - \mathbf{u}'\| + \|\mathbf{v} - \mathbf{v}'\|)^{2d+1}$ with P_0 -probability tending to one. The ϵ -packing number of $[0, c\mathbf{1}] \times [0, c\mathbf{1}]$ with respect to the metric ρ is bounded by a constant multiple of ϵ^{-4d} . Then by Theorem 2.2.4 of van der Vaart and Wellner (1996), by taking $\eta = \delta^{(2d+1)/(4d+1)}$,

$$\begin{aligned}
& \|\sup\{|Y_n(\mathbf{u}, \mathbf{v}; V) - Y_n(\mathbf{u}', \mathbf{v}'; V)| : \rho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) \leq \delta\}\|_{4d+2} \\
& \lesssim \int_0^\eta (\epsilon^{-4d})^{1/(4d+2)} d\epsilon + \delta(\eta^{-8d})^{1/(4d+2)},
\end{aligned}$$

which is of the order of $\delta^{1/(4d+1)}$. Hence the process $Y_n(\mathbf{u}, \mathbf{v}; V)$ is asymptotically uniformly equicontinuous. This concludes the proof of asymptotic tightness. \square

Lemma 3.5.2. *Let*

$$P_n(\mathbf{u}, \mathbf{v}) = \omega_n \sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} (N_j - E_0(N_j))$$

and $J_k \gg r_{n,k}^{-1}$. Then

$$P_n(\mathbf{u}, \mathbf{v}) \rightsquigarrow \sqrt{g_0(\mathbf{x}_0)} H_1(\mathbf{u}, \mathbf{v})$$

in $\mathbb{L}_\infty([0, c\mathbf{1}] \times [0, c\mathbf{1}])$.

Proof. Let $Z_{n,i}(\mathbf{u}, \mathbf{v}) = \omega_n \mathbb{1}_{\{I_j; j \in [j(-\mathbf{u}); j(\mathbf{v})]\}}(\mathbf{X}_i)$ for $i = 1, \dots, n$. We rewrite

$$P_n(\mathbf{u}, \mathbf{v}) = \omega_n \sum_{i=1}^n Z_{n,i}(\mathbf{u}, \mathbf{v}) - E_0 Z_{n,i}(\mathbf{u}, \mathbf{v}).$$

First, we verify the finite-dimensional convergence. For (\mathbf{u}, \mathbf{v}) and $(\mathbf{u}', \mathbf{v}') \in [0, c\mathbf{1}] \times [0, c\mathbf{1}]$,

$$\text{Cov}(P_n(\mathbf{u}, \mathbf{v}), P_n(\mathbf{u}', \mathbf{v}')) = n\omega_n^2 \mathbb{E}_0 Z_{n,1}(\mathbf{u} \wedge \mathbf{u}', \mathbf{v} \wedge \mathbf{v}') - n\omega_n^2 \mathbb{E}_0 Z_{n,1}(\mathbf{u}, \mathbf{v}) \cdot \mathbb{E}_0 Z_{n,1}(\mathbf{u}', \mathbf{v}').$$

We observe that

$$\begin{aligned} n\omega_n^2 \mathbb{E}_0 Z_{n,1}(\mathbf{u}, \mathbf{v}) &= n\omega_n^2 \int_{\bigcup_{\{I_j: j \in [j(-\mathbf{u}): j(\mathbf{v})]\}} \mathbf{g}_0(\mathbf{x}) d\mathbf{x}} \\ &= (\mathbf{g}_0(\mathbf{x}_0) + o(1)) \prod_{k=1}^d (u_k + v_k + O((J_k r_{n,k})^{-1})) \\ &\rightarrow \mathbf{g}_0(\mathbf{x}_0) \prod_{k=1}^d (u_k + v_k), \end{aligned}$$

uniformly for all $(\mathbf{u}, \mathbf{v}) \in [0, c\mathbf{1}] \times [0, c\mathbf{1}]$ as $J_k \gg r_{n,k}^{-1}$. Thus

$$\text{Cov}(P_n(\mathbf{u}, \mathbf{v}), P_n(\mathbf{u}', \mathbf{v}')) \rightarrow \mathbf{g}_0(\mathbf{x}_0) \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k).$$

Next, we apply Lemma A.2.3 to show the asymptotic tightness. Take $m_n = n$ and $\mathcal{F} = \{(\mathbf{u}, \mathbf{v}) \in [0, c\mathbf{1}] \times [0, c\mathbf{1}]\}$. By noting that $\|Z_{n,i}\|_{\mathcal{F}} = \omega_n \mathbb{1}_{\bigcup_{\{I_j: j \in [j(-c\mathbf{1}): j(c\mathbf{1})]\}} (\mathbf{X}_i)}$ converges to 0, and hence for every $\eta > 0$,

$$\sum_{i=1}^n \mathbb{E}_0 \|Z_{n,i}\|_{\mathcal{F}}^2 \mathbb{1}_{\{\|Z_{n,i}\|_{\mathcal{F}} > \eta\}} = 0$$

for sufficiently large n , trivially verifying the first condition of their theorem.

Let $\varrho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) = \|\mathbf{u} - \mathbf{u}'\| + \|\mathbf{v} - \mathbf{v}'\|$. Then

$$\begin{aligned}
& \sum_{i=1}^n \mathbb{E}_0(Z_{n,i}(\mathbf{u}, \mathbf{v}) - Z_{n,i}(\mathbf{u}', \mathbf{v}'))^2 \\
&= n\omega_n^2 \mathbb{E}_0 |Z_{n,1}(\mathbf{u}, \mathbf{v}) - Z_{n,1}(\mathbf{u}', \mathbf{v}')| \\
&= n\omega_n^2 \int_{\cup\{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})] \Delta [j(-\mathbf{u}'):j(\mathbf{v}')]\}} \mathbf{g}_0(\mathbf{x}) d\mathbf{x} \\
&\lesssim \prod_{k=1}^d (u_k + v_k + 2(J_k r_{n,k})^{-1}) + \prod_{k=1}^d (u'_k + v'_k + 2(J_k r_{n,k})^{-1}) \\
&\quad - 2 \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k) \\
&\lesssim \varrho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) + \max_{1 \leq k \leq d} (J_k r_{n,k})^{-1},
\end{aligned}$$

where the last inequality follows from Lemma A.3.1. Hence,

$$\sum_{i=1}^n \mathbb{E}_0(Z_{n,i}(\mathbf{u}, \mathbf{v}) - Z_{n,i}(\mathbf{u}', \mathbf{v}'))^2 \rightarrow 0$$

when $\varrho((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) \rightarrow 0$ and $n \rightarrow \infty$.

Next, for any $\epsilon > 0$, we construct a partition of \mathcal{F} as follows. Choose a $\delta > 0$, to be determined later, which depends only on ϵ . For the interval $[0, c]$, with equispaced grid points $0 = s_0 < s_1 < \dots < s_l = c$, the partition of $(0, c]^d \times (0, c]^d$ is given by $\{\prod_{k=1}^d (s_{t_k-1}, s_{t_k}] \times \prod_{k=1}^d (s_{r_k-1}, s_{r_k}]: t_k, r_k \in \{1, \dots, l\}\}$. Then $\mathcal{F} \subset \bigcup_{t, r \in \{1, \dots, l\}^d} \mathcal{F}_{t, r}$, where

$$\mathcal{F}_{t, r} = \{(\mathbf{u}, \mathbf{v}) \in (0, c]^d \times (0, c]^d : s_{t_k-1} < u_k \leq s_{t_k}, s_{r_k-1} < v_k \leq s_{r_k}\}.$$

Now

$$\begin{aligned}
& \sum_{i=1}^n \mathbb{E}_0 \sup_{f, g \in \mathcal{F}_{t,r}} |Z_{n,i}(f) - Z_{n,i}(g)|^2 \\
& \leq n \omega_n^2 \mathbb{E}_0 \left(\mathbb{1}_{\cup\{I_j: j \in [j(-s_t): j(s_r)]\}}(\mathbf{X}_1) - \mathbb{1}_{\cup\{I_j: j \in [j(-s_{t-1}): j(s_{r-1})]\}}(\mathbf{X}_1) \right)^2 \\
& \lesssim \prod_{k=1}^d (s_{t_k} + s_{r_k} + 2(J_k r_{n,k})^{-1}) - \prod_{k=1}^d (s_{t_{k-1}} + s_{r_{k-1}}) \\
& \lesssim \delta + \max_{1 \leq k \leq d} (J_k r_{n,k})^{-1},
\end{aligned}$$

where the last inequality follows from Lemma A.3.1. As $J_k \gg r_{n,k}^{-1}$ for $k = 1, \dots, d$, the second term in the last display will be eventually smaller than a prescribed $\delta > 0$. Thus δ can be set to be a multiple of ϵ^2 to meet the second condition. Meanwhile, the bracketing number with respect to the semimetric \mathbb{L}_2^n defined therein, $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2^n) \leq l^{2d} \leq (2c/\delta)^{2d} \lesssim \epsilon^{-4d}$. Then there exists a positive constant C such that

$$\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, L_2^n)} d\epsilon \leq \int_0^{\delta_n} \sqrt{\log(C \epsilon^{-4d})} d\epsilon \rightarrow 0,$$

as $\delta_n \rightarrow 0$. This verifies all three conditions in Lemma A.2.3, and hence concludes the weak convergence of P_n in $\mathbb{L}_\infty([\mathbf{0}, c\mathbf{1}], [\mathbf{0}, c\mathbf{1}])$. \square

Lemma 3.5.3. *Under the conditions of Theorem 3.3.1, for any $\tau > 0$, it holds that*

$$\mathbb{E}_0 \sup_{\substack{\mathbf{u} \geq \tau \mathbf{1}, \\ \mathbf{v} \geq \tau \mathbf{1}}} \left| \frac{\sum_{[j(-\mathbf{u}): j(\mathbf{v})]} (N_j - \mathbb{E}_0(N_j))/n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} \right| \lesssim \omega_n.$$

Proof. Without loss of generality, we shall take $\tau = 1$.

$$\begin{aligned}
& \mathbb{E}_0 \sup_{\mathbf{u} \geq \mathbf{1}, \mathbf{v} \geq \mathbf{1}} \left| \frac{\sum_{j \in [j(-\mathbf{u}): j(\mathbf{v})]} (N_j - \mathbb{E}_0(N_j))/n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} \right| \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \sum_{\substack{h'_k \geq 0 \\ 1 \leq k \leq d}} \mathbb{E}_0 \max_{\substack{2^{h_k} \leq u_k \leq 2^{h_k+1} \\ 2^{h'_k} \leq v_k \leq 2^{h'_k+1} \\ 1 \leq k \leq d}} \left| \frac{\sum_{j \in [j(-\mathbf{u}+\mathbf{1}): j(\mathbf{v}-\mathbf{1})]} (N_j - \mathbb{E}_0(N_j))}{n \prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} \right|. \quad (3.13)
\end{aligned}$$

Let $\mathbf{h} = (h_1, \dots, h_d)^\top$ and $\mathbf{h}' = (h'_1, \dots, h'_d)^\top$. First, we have

$$\begin{aligned}
& \min_{\substack{2^{h_k} \leq u_k \leq 2^{h_k+1} \\ 2^{h'_k} \leq v_k \leq 2^{h'_k+1} \\ 1 \leq k \leq d}} n \prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1} \\
&= n \prod_{k=1}^d (j(2^{\mathbf{h}'})_k - j(-2^{\mathbf{h}})_k + 1) J_k^{-1} \\
&\geq n \prod_{k=1}^d [r_{n,k} (2^{h_k} + 2^{h'_k})].
\end{aligned} \tag{3.14}$$

By Jensen's inequality,

$$\mathbb{E}_0 \max_{\substack{2^{h_k} \leq u_k \leq 2^{h_k+1} \\ 2^{h'_k} \leq v_k \leq 2^{h'_k+1} \\ 1 \leq k \leq d}} \left| \sum_{[j(-\mathbf{u}); j(\mathbf{v})]} (N_j - \mathbb{E}_0(N_j)) \right| \leq \left[\mathbb{E}_0 \max_{\substack{2^{h_k} \leq u_k \leq 2^{h_k+1} \\ 2^{h'_k} \leq v_k \leq 2^{h'_k+1} \\ 1 \leq k \leq d}} \left| \sum_{[j(-\mathbf{u}); j(\mathbf{v})]} (N_j - \mathbb{E}_0(N_j)) \right|^2 \right]^{1/2} \tag{3.15}$$

By Lemma A.2.6, the right-hand side of (3.15) is further bounded, up to a constant multiple depending only on d , by

$$\left[\mathbb{E}_0 \left| \sum_{j \in [j(-2^{\mathbf{h}+1}); j(2^{\mathbf{h}' + 1})]} (N_j - \mathbb{E}(N_j)) \right|^2 \right]^{1/2}.$$

As $\sum_{j \in [j(-2^{\mathbf{h}+1}); j(2^{\mathbf{h}' + 1})]} N_j \sim \text{Bin}(n, \int_{\cup \{I_j: j \in [j(-2^{\mathbf{h}+1}); j(2^{\mathbf{h}' + 1})]\}} \mathbf{g}_0(\mathbf{x}) d\mathbf{x})$, we have

$$\begin{aligned}
& \left[\mathbb{E}_0 \left| \sum_{[j(-2^{\mathbf{h}+1}); j(2^{\mathbf{h}' + 1})]} (N_j - \mathbb{E}(N_j)) \right|^2 \right]^{1/2} \\
&\leq \left[n \int_{\cup \{I_j: j \in [j(-2^{\mathbf{h}+1}); j(2^{\mathbf{h}' + 1})]\}} \mathbf{g}_0(\mathbf{x}) d\mathbf{x} \right]^{1/2} \\
&\lesssim \left[n \prod_{k=1}^d r_{n,k} (2^{h_k+1} + 2^{h'_k+1}) \right]^{1/2},
\end{aligned}$$

as \mathbf{g}_0 is bounded and $J_k \gg r_{n,k}^{-1}$. Combining with (3.14), we conclude that (3.13) is bounded, up

to some positive constant, by

$$\omega_n \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \sum_{\substack{h'_k \geq 0 \\ 1 \leq k \leq d}} \prod_{k=1}^d (2^{h_k} + 2^{h'_k})^{-1/2}.$$

Then it remains to show the convergence of the sum of the series in the last display. To see this, using the fact that $\sum_{h \geq 0} \sum_{h' \geq 0} (2^h + 2^{h'})^{-1/2}$ converges, we have

$$\sum_{\substack{0 \leq h_k \leq n_k \\ 1 \leq k \leq d}} \sum_{\substack{0 \leq h'_k \leq n'_k \\ 1 \leq k \leq d}} \prod_{k=1}^d (2^{h_k} + 2^{h'_k})^{-1/2} = \prod_{k=1}^d \left(\sum_{0 \leq h_k \leq n_k} \sum_{0 \leq h'_k \leq n'_k} (2^{h_k} + 2^{h'_k})^{-1/2} \right),$$

and it will converges to some positive constant when n_k and n'_k go to infinity. Thus we conclude the proof of the lemma. \square

Lemma 3.5.4. *Under the conditions of Theorem 3.3.1, for any $M_n \uparrow \infty$, we have $\Pi(|g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0)| \geq M_n \omega_n |\mathbb{D}_n|) \rightarrow_{\mathbb{P}_0} 0$.*

Proof. We shall prove one side of the claim that

$$\mathbb{P}_0(\Pi(g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0) \geq M_n \omega_n |\mathbb{D}_n|) \geq \eta) \rightarrow 0 \quad (3.16)$$

for any $\eta > 0$. The other side follows by the similar arguments.

By the min-max formula, we have

$$\begin{aligned} \omega_n^{-1} (g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0)) &\leq \omega_n^{-1} \left(\max_{\mathbf{v} \geq 0} \frac{\sum_{[j(-1):j(\mathbf{v})]} \theta_j}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-1)_k + 1) J_k^{-1}} - g_0(\mathbf{x}_0) \right) \\ &\leq \max_{\mathbf{v} \geq 0} |A_{n,1}(\mathbf{1}, \mathbf{v}; V)| + \max_{\mathbf{v} \geq 0} |A_{n,2}(\mathbf{1}, \mathbf{v}; V)| \\ &\quad + \max_{\mathbf{v} \geq 0} |B_{n,1}(\mathbf{1}, \mathbf{v})| + \max_{\mathbf{v} \geq 0} |B_{n,2}(\mathbf{1}, \mathbf{v})|. \end{aligned}$$

Thus it follows that

$$\begin{aligned}
& \Pi(\mathbf{g}^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0) \geq M_n \omega_n | \mathbb{D}_n) \\
& \leq \Pi\left(\max_{\mathbf{v} \geq \mathbf{0}} |A_{n,1}(\mathbf{1}, \mathbf{v}; V)| > M_n/4 | \mathbb{D}_n\right) + \Pi\left(\max_{\mathbf{v} \geq \mathbf{0}} |A_{n,2}(\mathbf{1}, \mathbf{v}; V)| > M_n/4 | \mathbb{D}_n\right) \\
& \quad + \mathbb{1}\left\{\max_{\mathbf{v} \geq \mathbf{0}} |B_{n,1}(\mathbf{1}, \mathbf{v})| > M_n/4\right\} + \mathbb{1}\left\{\max_{\mathbf{v} \geq \mathbf{0}} |B_{n,2}(\mathbf{1}, \mathbf{v})| > M_n/4\right\}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
& P_0(\Pi(\mathbf{g}^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0) \geq M_n \omega_n | \mathbb{D}_n) \leq \eta) \\
& \leq P_0\left(\Pi\left(\max_{\mathbf{v} \geq \mathbf{0}} |A_{n,1}(\mathbf{1}, \mathbf{v}; V)| > M_n/4\right) > \eta/2 | \mathbb{D}_n\right) \tag{3.17}
\end{aligned}$$

$$+ P_0\left(\Pi\left(\max_{\mathbf{v} \geq \mathbf{0}} |A_{n,2}(\mathbf{1}, \mathbf{v}; V)| > M_n/4 | \mathbb{D}_n\right) > \eta/2\right) \tag{3.18}$$

$$+ P_0\left(\max_{\mathbf{v} \geq \mathbf{0}} |B_{n,1}(\mathbf{1}, \mathbf{v})| > M_n/4\right) \tag{3.19}$$

$$+ P_0\left(\max_{\mathbf{v} \geq \mathbf{0}} |B_{n,2}(\mathbf{1}, \mathbf{v})| > M_n/4\right). \tag{3.20}$$

It suffices to show that each term (3.17)–(3.20) converges to zero.

To show that (3.17) converges to zero, it is enough to show that

$$\mathbb{E}\left(\max_{\mathbf{v} \geq \mathbf{0}} \left| \frac{\sum_{[j(-1); j(\mathbf{v})]} (V_j - \mathbb{E}(V_j))/n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-1)_k + 1) J_k^{-1}} \right| \middle| \mathbb{D}_n \right) = O_{P_0}(\omega_n). \tag{3.21}$$

We use arguments similar to those in the proof of Lemma 3.5.3. By splitting the domain into

smaller rectangles, we can bound

$$\begin{aligned}
& \mathbb{E} \left(\max_{v \geq 0} \left| \frac{\sum_{[j(-1):j(v)]} (V_j - \mathbb{E}(V_j)) / n}{\prod_{k=1}^d (j(v)_k - j(-1)_k + 1) J_k^{-1}} \right| \middle| \mathbb{D}_n \right) \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \mathbb{E} \left(\max_{\substack{2^{h_k} \leq v_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \frac{\sum_{[j(-1):j(v-1)]} (V_j - \mathbb{E}(V_j)) / n}{\prod_{k=1}^d (j(v-1)_k - j(-1)_k + 1) J_k^{-1}} \right| \middle| \mathbb{D}_n \right) \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \mathbb{E} \left(\max_{\substack{2^{h_k} \leq v_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \sum_{[j(-1):j(v-1)]} (V_j - \mathbb{E}(V_j)) \right| \middle| \mathbb{D}_n \right) \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \left[\mathbb{E} \left(\max_{\substack{2^{h_k} \leq v_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \sum_{[j(-1):j(v-1)]} (V_j - \mathbb{E}(V_j)) \right|^2 \middle| \mathbb{D}_n \right) \right]^{1/2} \\
& \leq C_d \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \left[\mathbb{E} \left(\left| \sum_{[j(-1):j(2^{h+1}-1)]} (V_j - \mathbb{E}(V_j)) \right|^2 \middle| \mathbb{D}_n \right) \right]^{1/2} \\
& = C_d \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \left[\sum_{[j(-1):j(2^{h+1}-1)]} (\alpha_j + N_j) \right]^{1/2} \tag{3.22}
\end{aligned}$$

for some positive constant C_d depending only on d . As, by Lemma A.2.4, $\{\alpha_j\}$ is bounded uniformly by a constant $a_1 > 0$ and $N_j \asymp n / \prod_{k=1}^d J_k$ uniformly for all j with \mathbb{P}_0 -probability tending to one, (3.22) is bounded, up to some constant multiple, by

$$\frac{(n \prod_{k=1}^d r_{n,k})^{1/2}}{n \prod_{k=1}^d r_{n,k}} \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} 2^{-\sum_{k=1}^d h_k/2} = C \omega_n,$$

for some constant $C > 0$ with \mathbb{P}_0 -probability tending to one, since the sum of the series in the last display converges. Thus (3.21) holds.

To bound (3.18), we follow the arguments used in bounding (6.7) by replacing u by 1 and lower-bounding v to 0.

To obtain a bound for (3.19), we proceed as in the proof of Lemma 3.5.3, and observe that

$$\begin{aligned}
& \mathbb{E}_0 \max_{\mathbf{v} \geq \mathbf{0}} \left| \frac{\sum_{[j(-1):j(\mathbf{v})]} (N_j - \mathbb{E}_0(N_j))/n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-1)_k + 1) J_k^{-1}} \right| \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \mathbb{E}_0 \max_{\substack{2^{h_k} \leq v_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \frac{\sum_{[j(-1):j(\mathbf{v}-1)]} (N_j - \mathbb{E}_0(N_j))/n}{\prod_{k=1}^d (j(\mathbf{v}-1)_k - j(-1)_k + 1) J_k^{-1}} \right| \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \mathbb{E}_0 \max_{\substack{2^{h_k} \leq v_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \sum_{[j(-1):j(\mathbf{v}-1)]} (N_j - \mathbb{E}_0(N_j)) \right| \\
& \leq \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \left[\mathbb{E}_0 \max_{\substack{2^{h_k} \leq v_k \leq 2^{h_k+1} \\ 1 \leq k \leq d}} \left| \sum_{[j(-1):j(\mathbf{v}-1)]} (N_j - \mathbb{E}_0(N_j)) \right|^2 \right]^{1/2} \\
& \leq C_d \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{1}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})} \left[\mathbb{E}_0 \left| \sum_{[j(-1):j(2^{h_k+1}-1)]} (N_j - \mathbb{E}_0(N_j)) \right|^2 \right]^{1/2} \\
& \leq C_{d, g_0} \sum_{\substack{h_k \geq 0 \\ 1 \leq k \leq d}} \frac{(n \prod_{k=1}^d (r_{n,k} 2^{h_k+1}))^{1/2}}{n \prod_{k=1}^d (r_{n,k} 2^{h_k})},
\end{aligned}$$

which is bounded by a constant multiple of ω_n . Thus $\max\{|B_{n,1}(\mathbf{1}, \mathbf{v})| : \mathbf{v} \geq \mathbf{0}\} = O_{\mathbb{P}_0}(1)$.

We observe that $B_{n,2}$ is deterministic and

$$|B_{n,2}(\mathbf{1}, \mathbf{v})| = \omega_n^{-1} \left| \frac{\int_{\bigcup\{I_j : j \in [j(-1):j(\mathbf{0})] \setminus j(\mathbf{0})\}} g_0(\mathbf{x}) d\mathbf{x} + \int_{I_{j(\mathbf{0})}} g_0(\mathbf{x}) d\mathbf{x}}{\int_{\bigcup\{I_j : j \in [j(-1):j(\mathbf{0})]\}} d\mathbf{x}} - g_0(\mathbf{x}_0) \right|,$$

As g_0 is nonincreasing, it is clear that $g_0(\mathbf{x}) > g_0(\mathbf{x}_0)$ for all $\mathbf{x} \in \bigcup\{I_j : j \in [j(-1):j(\mathbf{0})] \setminus j(\mathbf{0})\}$.

Thus, $|B_{n,2}(\mathbf{1}, \mathbf{v})|$ is bounded by

$$\omega_n^{-1} \frac{\int_{\bigcup\{I_j : j \in [j(-1):j(\mathbf{0})] \setminus j(\mathbf{0})\}} g_0(\mathbf{x}) - g_0(\mathbf{x}_0) d\mathbf{x}}{\int_{\bigcup\{I_j : j \in [j(-1):j(\mathbf{0})]\}} d\mathbf{x}} + \omega_n^{-1} \left| \frac{\int_{I_{j(\mathbf{0})}} g_0(\mathbf{x}) - g_0(\mathbf{x}_0) d\mathbf{x}}{\int_{\bigcup\{I_j : j \in [j(-1):j(\mathbf{0})]\}} d\mathbf{x}} \right|,$$

which is further bounded by

$$\omega_n^{-1} [g_0((\|\mathbf{x}_0 - \mathbf{r}_n\| - 1)/J) - g_0(\mathbf{x}_0)] + o(1) \lesssim \left| \sum_{m \in M} \frac{\partial^m g_0(\mathbf{x}_0)}{m!} \right|,$$

as $\prod_{k=1}^d r_{n,k}^{m_k} = \omega_n^{\sum_{k=1}^d m_k/\eta_k} = \omega_n$ for $\mathbf{m} \in M$. Thus, $\max_{\mathbf{v} \geq 0} |B_n(\mathbf{1}, \mathbf{v})| \leq M_n/4$ when n large enough.

Piecing these together, (3.16) follows. \square

Lemma 3.5.5. *Let \mathbf{u}^* and \mathbf{v}^* be such that*

$$\frac{\sum_{[j(-\mathbf{u}^*); j(\mathbf{v}^*)]} \theta_j}{\prod_{k=1}^d (j(\mathbf{v}^*)_k - j(-\mathbf{u}^*)_k + 1) J_k^{-1}} = \min_{\mathbf{u} \geq 0} \max_{\mathbf{v} \geq 0} \frac{\sum_{[j(-\mathbf{u}); j(\mathbf{v})]} \theta_j}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}.$$

Then there exist $c > 1$ and $\gamma > 0$ such that

$$\lim_{c \rightarrow \infty} \limsup_{n \rightarrow \infty} \Pi(c^{-\gamma} \leq \min_{1 \leq k \leq d} u_k^* \leq \max_{1 \leq k \leq d} u_k^* \leq c |D_n|) = 1,$$

in P_0 -probability.

Proof. First, we show that

$$\lim_{c \rightarrow \infty} \limsup_{n \rightarrow \infty} \Pi(\max_{1 \leq k \leq d} u_k^* \leq c |D_n|) = 1 \text{ in } P_0\text{-probability.} \quad (3.23)$$

We note that, by the min-max formula,

$$\begin{aligned} \omega_n B_{n,2}(\mathbf{u}^*, \mathbf{v}^*) &= \frac{\sum_{[j(-\mathbf{u}^*); j(\mathbf{v}^*)]} E(N_j)/n}{\prod_{k=1}^d (j(\mathbf{v}^*)_k - j(-\mathbf{u}^*)_k + 1) J_k^{-1}} - g_0(\mathbf{x}_0) \\ &\geq \frac{\int_{\cup_{\{I_j: j \in [j(-\mathbf{u}^*); j(\mathbf{1})]\}} \mathbf{g}_0(\mathbf{x}) - g_0(\mathbf{x}_0) d\mathbf{x}}{\int_{\cup_{\{I_j: j \in [j(-\mathbf{u}^*); j(\mathbf{1})]\}} d\mathbf{x}}. \end{aligned} \quad (3.24)$$

By Assumption 2, for \mathbf{x} in a small neighborhood of \mathbf{x}_0 , we have

$$g_0(\mathbf{x}) - g_0(\mathbf{x}_0) = \sum_{m \in M} \frac{\partial^m g_0(\mathbf{x}_0)}{m!} (\mathbf{x} - \mathbf{x}_0)^m + o(\max_k |x_k - x_{0,k}|^{\eta_k}).$$

Thus, by noting that $g_0(\mathbf{x})$ is monotone nonincreasing and then $\partial_k^{\eta_k} g_0(\mathbf{x}_0) < 0$, (3.24) can be

lower bounded by

$$\begin{aligned}
& \sum_{m \in M} \frac{\partial^m g_0(\mathbf{x}_0) \int_{\cup_{j \in (-\mathbf{u}^*); j(1)} I_j} (\mathbf{x} - \mathbf{x}_0)^m d\mathbf{x}}{m! \int_{\cup_{j \in (-\mathbf{u}^*); j(1)} I_j} d\mathbf{x}} + o(\omega_n \max_k u_k^{*\eta_k}) \\
& \asymp \omega_n \sum_{m \in M} \frac{\partial^m g_0(\mathbf{x}_0)}{(m+1)!} \frac{1 - \prod_{k=1}^d (-u_k^*)^{m_k+1}}{\prod_{k=1}^d (1+u_k^*)} + o(\omega_n \max_k u_k^{*\eta_k}) \\
& \gtrsim \omega_n \max_k u_k^{*\eta_k},
\end{aligned}$$

which holds when n and $\max_{1 \leq k \leq d} u_k^*$ are sufficiently large. If $\max_k u_k^* \geq c$ for a sufficiently large $c > 0$, then $|g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0)| \gtrsim c \omega_n$. In view of Lemma 3.5.4, the assertion in (3.23) follows.

Without loss of generality, it is sufficient to show that $\Pi(u_d^* \leq c^{-\gamma} | \mathbb{D}_n) \rightarrow_{\mathbb{P}_0} 0$, as $c, n \rightarrow \infty$. We shall prove the claim by showing that $\{g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0) \geq c^\delta \omega_n\}$ for some $\delta > 0$, to be determined later, if $\{\min_{1 \leq k \leq d} u_k^* \leq c^{-\gamma}\}$ happens when n and c large enough. Then the claim is concluded with the help of Lemma 3.5.4.

Recall the notations, $A_{n,1}, A_{n,2}, B_{n,1}, B_{n,2}, s_n, Y_n$, and P_n , defined in the proofs of Theorem 3.3.1, Lemma 3.5.1, and Lemma 3.5.2. We shall use the same decomposition as in this proof of Theorem 3.3.1. For some constants $a > 0$ and $b > 0$, to be determined later, we can write $\omega_n^{-1}(g^*(\mathbf{x}_0) - g_0(\mathbf{x}_0))$ as

$$\begin{aligned}
& \max_{v \geq 0} \left(\frac{n}{\sum_l V_l} \cdot \frac{Y_n(\mathbf{u}^*, \mathbf{v}; V)}{s_n(\mathbf{u}^*, \mathbf{v})} + A_{n,2}(\mathbf{u}^*, \mathbf{v}; V) + \frac{P_n(\mathbf{u}^*, \mathbf{v})}{s_n(\mathbf{u}^*, \mathbf{v})} + B_{n,2}(\mathbf{u}^*, \mathbf{v}) \right) \\
& \geq \max_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} \left(\frac{n}{\sum_l V_l} \cdot \frac{Y_n(\mathbf{u}^*, \mathbf{v}; V)}{s_n(\mathbf{u}^*, \mathbf{v})} \right. \\
& \quad \left. + A_{n,2}(\mathbf{u}^*, \mathbf{v}; V) + \frac{P_n(\mathbf{u}^*, \mathbf{v})}{s_n(\mathbf{u}^*, \mathbf{v})} + B_{n,2}(\mathbf{u}^*, \mathbf{v}) \right) \\
& \geq \max_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} \left(\frac{n}{\sum_l V_l} \cdot \frac{Y_n(\mathbf{u}^*, \mathbf{v}; V)}{s_n(\mathbf{u}^*, \mathbf{v})} + \frac{P_n(\mathbf{u}^*, \mathbf{v})}{s_n(\mathbf{u}^*, \mathbf{v})} \right) \\
& \quad + \min_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} A_{n,2}(\mathbf{u}^*, \mathbf{v}; V) \\
& \quad + \min_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} B_{n,2}(\mathbf{u}^*, \mathbf{v}).
\end{aligned}$$

Define $\mathcal{E}_0(c) = \{u_d^* < c^{-\gamma}\}$ and $\mathcal{E}_1(c) = \{\max_{1 \leq k \leq d} u_k^* \leq c\}$. By the first part of the proof, for every η and $\epsilon > 0$, we have $P_0(\Pi(\mathcal{E}_1(c)|\mathbb{D}_n) \geq 1 - \eta) \geq 1 - \epsilon$ when c and n are large enough. Define $R_{a,b,\gamma}(c) = \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d : 0 \leq u_k \leq c \mathbb{1}\{1 \leq k \leq d-1\} + c^{-\gamma} \mathbb{1}\{k = d\}, 0 \leq v_k \leq c^a \mathbb{1}\{1 \leq k \leq d-1\} + c^{-b} \mathbb{1}\{k = d\}\}$, where a, b, γ will be determined later.

By Lemma 3.5.1 and Lemma 3.5.2, we know

$$Y_n(\mathbf{u}, \mathbf{v}; V) \rightsquigarrow \sqrt{g_0(\mathbf{x}_0)} H_2(\mathbf{u}, \mathbf{v}) \text{ in } P_0\text{-probability in } \mathbb{L}_\infty([0, c] \times [0, c]),$$

$$P_n(\mathbf{u}, \mathbf{v}) \rightsquigarrow \sqrt{g_0(\mathbf{x}_0)} H_1(\mathbf{u}, \mathbf{v}) \text{ in } \mathbb{L}_\infty([0, c] \times [0, c]).$$

Then by Lemma A.3.2, when c, n are large enough, there exists a constant C_1 depending on $g_0(\mathbf{x}_0), d, a$ such that $P_0(\Pi(\mathcal{E}_2(c)|\mathbb{D}_n) \geq 1 - \eta) \geq 1 - \epsilon$ where $\mathcal{E}_2(c)$ is defined as

$$\mathcal{E}_2(c) = \left\{ \sup_{(\mathbf{u}, \mathbf{v}) \in R_{a,b,\gamma}(c)} |Y_n(\mathbf{u}, \mathbf{v}) - Y_n(\mathbf{0}, \mathbf{v})| \leq (C_1/\eta) \sqrt{c^{a(d-1)-\gamma} \log c} \right\}.$$

Similarly, there exists $C_2 > 0$ such that, for

$$E_2(c) = \left\{ \sup_{(\mathbf{u}, \mathbf{v}) \in R_{a,b,\gamma}(c)} |P_n(\mathbf{u}, \mathbf{v}) - P_n(\mathbf{0}, \mathbf{v})| \leq (C_2/\epsilon) \sqrt{c^{a(d-1)-\gamma} \log c} \right\},$$

we have $P_0(E_2(c)) \geq 1 - \epsilon$ when n and c large enough.

As H_1 and H_2 are two independent Gaussian processes, by Lemma A.3.3, there exists some $\rho_{\eta,\epsilon} > 0$ such that, when $a > 1, c > 1$ and n large enough, we have

$$\begin{aligned} & P_0 \times \Pi \left(\max_{0 \leq v_k \leq c^a \mathbb{1}\{1 \leq k \leq d-1\} + c^{-b} \mathbb{1}\{k=d\}} Y_n(\mathbf{0}, \mathbf{v}) + P_n(\mathbf{0}, \mathbf{v}) \leq \sqrt{c^{a(d-1)-b}} \rho_{\eta,\epsilon} \right) \\ & \rightarrow P \left(\max_{\substack{0 \leq v_k \leq c^a \mathbb{1}\{1 \leq k \leq d-1\} \\ 0 \leq v_d \leq c^{-b}}} H_1(\mathbf{0}, \mathbf{v}) + H_2(\mathbf{0}, \mathbf{v}) \leq \sqrt{c^{a(d-1)-b}} \rho_{\eta,\epsilon} / \sqrt{g_0(\mathbf{x}_0)} \right) \\ & \leq P \left(\max_{\substack{0 \leq v_k \leq 1 \\ 1 \leq k \leq d}} H_1(\mathbf{0}, \mathbf{v}) + H_2(\mathbf{0}, \mathbf{v}) \leq \rho_{\eta,\epsilon} / \sqrt{g_0(\mathbf{x}_0)} \right) \leq \eta \epsilon. \end{aligned}$$

Hence, there exists a constant $C_3 > 0$ such that, for the event

$$\mathcal{E}_3(c) = \{\sup\{Y_n(\mathbf{0}, \mathbf{v}) + P_n(\mathbf{0}, \mathbf{v}) : 0 \leq v_k \leq c^a, 0 \leq v_d \leq c^{-b}\} > C_3 \sqrt{c^{a(d-1)-b}}\},$$

we have $(P_0 \times \Pi)(\mathcal{E}_c^3) \geq 1 - \eta\epsilon$ for n large enough. This implies the assertion that $P_0(\Pi(\mathcal{E}_3(c)|\mathbb{D}_n) \geq 1 - \eta) \geq 1 - \epsilon$ when n large enough.

For s_n , we have on $R_{a,b,\gamma}(c)$,

$$\begin{aligned} s_n(\mathbf{u}, \mathbf{v}) &= n\omega_n^2 \prod_{k=1}^d \left(\left[(x_{0,k} + r_{n,k} v_k) J_k \right] - \left[(x_{0,k} - r_{n,k} u_k) J_k \right] + 1 \right) J_k^{-1} \\ &\leq \prod_{k=1}^d (u_k + v_k + 2J_k^{-1} r_{n,k}^{-1}) \\ &\lesssim (c + c^a)^{d-1} (c^{-b} + c^{-\gamma}), \end{aligned}$$

Hence, on the intersection of events $\mathcal{E}_0(c), \mathcal{E}_1(c), \mathcal{E}_2(c), E_2(c)$ and $\mathcal{E}_3(c)$, it holds that

$$\begin{aligned} &\max_{\substack{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} \\ 0 \leq u_d \leq c^{-b}}} \frac{Y_n(\mathbf{u}^*, \mathbf{v}; V) + P_n(\mathbf{u}^*, \mathbf{v})}{s_n(\mathbf{u}^*, \mathbf{v})} \\ &\gtrsim \frac{C_3 \sqrt{c^{a(d-1)-b}} - C_1 \sqrt{c^{a(d-1)-\gamma} \log c} / \eta - C_2 \sqrt{c^{a(d-1)-\gamma} \log c} / \epsilon}{(c^a + c)^{d-1} (c^{-b} + c^{-\gamma})} \\ &\geq C_4 \sqrt{c^{b-a(d-1)}}, \end{aligned}$$

for some $C_4 > 0$ when n and c large enough and $\gamma > b$.

To bound $A_{n,2}(\mathbf{u}, \mathbf{v})$ uniformly in (\mathbf{u}, \mathbf{v}) , we follow the arguments used in bounding $A_{n,2}$ in the proof of Theorem 3.3.1 and Lemma A.2.4, to observe that

$$\frac{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} N_j / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} \asymp \frac{\sum_{j \in [j(-\mathbf{u}); j(\mathbf{v})]} \prod_{k=1}^d J_k^{-1}}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} = 1,$$

uniformly for all (\mathbf{u}, \mathbf{v}) , with P_0 -probability tending to one.

We also observe that for n sufficiently large, by monotonicity and Assumption 2, there exists

$C_5 > 0$ such that

$$\begin{aligned}
& \min_{\substack{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} \\ 0 \leq v_d \leq c^{-b}}} B_{n,2}(\mathbf{u}^*, \mathbf{v}) \\
&= \omega_n^{-1} \left(\min_{\substack{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} \\ 0 \leq v_d \leq c^{-b}}} \frac{\int_{\cup \{I_j: j \in [j(-\mathbf{u}^*); j(\mathbf{v})]\}} \mathbf{g}_0(\mathbf{x}) d\mathbf{x}}{\int_{\cup \{I_j: j \in [j(-\mathbf{u}^*); j(\mathbf{v})]\}} d\mathbf{x}} - \mathbf{g}_0(\mathbf{x}_0) \right) \\
&\geq \omega_n^{-1} \left(\min_{\substack{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} \\ 0 \leq v_d \leq c^{-b}}} \mathbf{g}_0(\mathbf{x}_0 + r_n \mathbf{v} + \mathbf{J}^{-1}) - \mathbf{g}_0(\mathbf{x}_0) \right) \\
&= -\omega_n^{-1} \cdot \\
&\quad \max_{\substack{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} \\ 0 \leq v_d \leq c^{-b}}} \sum_{m \in M} \frac{|\partial^m \mathbf{g}_0(\mathbf{x}_0)|}{m!} \prod_{k=1}^d (v_k r_{n,k} + J_k^{-1})^{m_k} + o(c^{a \max_k \eta_k}) \\
&\geq -C_5 c^{a \max_k \eta_k}.
\end{aligned}$$

As here we assume that $1 \leq \alpha_k < \infty$ for every k , we take $a = 3$, $b \geq 2a \max_k \alpha_k + a(d-1)$ and $\gamma = b + 1$. To fulfill the conditions on a, b , and γ in Lemma A.3.2. Let $\delta = (b - a(d-1))/2 \geq a \max_k \eta_k$. On the intersection of events $\mathcal{E}_c^0, \mathcal{E}_c^1, \mathcal{E}_c^2, E_c^2$ and \mathcal{E}_c^3 , when n and c large enough, we have

$$\begin{aligned}
& \omega_n^{-1}(\mathbf{g}^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0)) \\
&\geq \max_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} \left(\frac{n}{\sum_l V_l} \cdot \frac{Y_n(\mathbf{u}^*, \mathbf{v}; V)}{s_n(\mathbf{u}^*, \mathbf{v})} + \frac{P_n(\mathbf{u}^*, \mathbf{v})}{s_n(\mathbf{u}^*, \mathbf{v})} \right) \\
&\quad + \min_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} A_{n,2}(\mathbf{u}^*, \mathbf{v}; V) \\
&\quad + \min_{0 \leq v_k \leq c^a \mathbb{1}_{\{1 \leq k \leq d-1\}} + c^{-b} \mathbb{1}_{\{k=d\}}} B_{n,2}(\mathbf{u}^*, \mathbf{v}) \\
&\gtrsim C_4 \sqrt{c^{b-a(d-1)}} - C_5 c^{a \max_k \eta_k} \gtrsim c^\delta.
\end{aligned}$$

Hence for some $C_6 > 0$, on an event with P_0 -probability tending to 1,

$$\Pi\left(\bigcap_{p=0}^3 \mathcal{E}_p(c) \cap E_2(c) \mid \mathbb{D}_n\right) \leq \Pi(\omega_n^{-1}(\mathbf{g}^*(\mathbf{x}_0) - \mathbf{g}_0(\mathbf{x}_0)) \geq C_6 c^\delta \mid \mathbb{D}_n). \quad (3.25)$$

By Lemma 3.5.4, the right hand side of (3.25) can be arbitrarily small if we take c sufficiently large in P_0 -probability. On the other hand, we know that $P_0(\Pi(\mathcal{E}_1(c)|\mathbb{D}_n) \geq 1 - \eta) \geq 1 - \epsilon$, $P_0(\Pi(\mathcal{E}_2(c)|\mathbb{D}_n) \geq 1 - \eta) \geq 1 - \epsilon$, $P_0(E_2(c)) \geq 1 - \epsilon$, and $P_0(\Pi(\mathcal{E}_3(c)|\mathbb{D}_n) \geq 1 - \eta) \geq 1 - \epsilon$. Thus we can conclude that $P_0(\Pi_n(\mathcal{E}_0(c)) \leq \eta) \geq 1 - 5\epsilon$, when n and c are taken large enough. \square

Proof of Theorem 3.3.1. Let $s_n(\mathbf{u}, \mathbf{v}) = n\omega_n^2 \prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}$. We claim that $s_n(\mathbf{u}, \mathbf{v}) \rightarrow \prod_{k=1}^d (u_k + v_k)$ uniformly for $(\mathbf{u}, \mathbf{v}) \in [0, c\mathbf{1}] \times [0, c\mathbf{1}]$ for any $c > 0$. We can bound s_n by

$$n\omega_n^2 \prod_{k=1}^d (u_k r_{n,k} + v_k r_{n,k}) \leq s_n(\mathbf{u}, \mathbf{v}) \leq n\omega_n^2 \prod_{k=1}^d (u_k r_{n,k} + v_k r_{n,k} + 2J_k^{-1}).$$

Note that $n\omega_n^2 \prod_{k=1}^d r_{n,k} = 1$. By Lemma A.3.1, we have

$$\left| \prod_{k=1}^d (u_k + v_k + 2(r_{n,k} J_k)^{-1}) - \prod_{k=1}^d (u_k + v_k) \right| \leq C_{c,d} \max_{1 \leq k \leq d} (r_{n,k} J_k)^{-1},$$

where $C_{c,d}$ is a positive constant only relative to c and d . Then the claim follows as $J_k \gg r_{n,k}^{-1}$ for all $1 \leq k \leq d$.

As $\sum_{l \in [1:J]} V_l \sim \text{Ga}(\alpha + n, 1)$, by Assumption 6, we have

$$\frac{n}{\sum_{l \in [1:J]} V_l} - 1 = O_P(\max\{n^{-1} \prod_{k=1}^d J_k, n^{-1/2}\}). \quad (3.26)$$

We see that $A_{n,1}(\mathbf{u}, \mathbf{v}) = Y_n(\mathbf{u}, \mathbf{v}; V) \cdot (n / \sum_l V_l) / s_n(\mathbf{u}, \mathbf{v})$, where Y_n is defined in Lemma 3.5.1. Combining with Lemma 3.5.1, we prove that the conditional distribution of $A_{n,1}(\mathbf{u}, \mathbf{v})$ given \mathbb{D}_n converges weakly to $\sqrt{g_0(\mathbf{x}_0)} H_2(\mathbf{u}, \mathbf{v}) / \prod_{k=1}^d (u_k + v_k)$ in $\mathbb{L}_\infty([0, c\mathbf{1}] \times [0, c\mathbf{1}])$ in P_0 -probability.

Secondly, for $A_{n,2}(\mathbf{u}, \mathbf{v})$, we have

$$|A_{n,2}(\mathbf{u}, \mathbf{v})| = \omega_n^{-1} \left| \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} (\alpha_j + N_j) / \sum_{l \in [1:J]} V_l - N_j / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} \right|$$

$$\leq \frac{a \prod_{k=1}^d J_k}{\omega_n \sum_{l \in [1:J]} V_l} + \omega_n^{-1} \left| \frac{n}{\sum_{l \in [1:J]} V_l} - 1 \right| \times \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} N_j / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}. \quad (3.27)$$

The first term converges in probability to zero in view of (3.26) and the condition on J_k , $\prod_{k=1}^d J_k \ll n \omega_n$. For the second term, by (3.26), we can obtain that

$$\omega_n^{-1} \left| \frac{n}{\sum_{l \in [1:J]} V_l} - 1 \right| = O_p \left(\max \left\{ \frac{\prod_{k=1}^d J_k}{n \omega_n}, \frac{1}{\sqrt{n \omega_n^2}} \right\} \right) \rightarrow_p 0,$$

in view of the condition $\prod_{k=1}^d J_k \ll n \omega_n$ again and by noting that $\omega_n \gg n^{-1/2}$. Next, in order to show that $A_{n,2}(\mathbf{u}, \mathbf{v})$ converges to zero in probability uniformly for $(\mathbf{u}, \mathbf{v}) \in [c^{-\gamma}, c]^d \times [c^{-\gamma}, c]^d$ for any $c > 1$ and $\gamma > 1$, it suffices to show that

$$\sup_{\substack{\mathbf{u} \geq c^{-\gamma} \mathbf{1}, \\ \mathbf{v} \geq c^{-\gamma} \mathbf{1}}} \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} N_j / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} = O_{P_0}(1). \quad (3.28)$$

To this end, Lemma 3.5.3 establishes that

$$E_0 \sup_{\substack{\mathbf{u} \geq c^{-\gamma} \mathbf{1}, \\ \mathbf{v} \geq c^{-\gamma} \mathbf{1}}} \left| \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} (N_j - E_0(N_j)) / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}} \right| = O(\omega_n). \quad (3.29)$$

Moreover, we observe that

$$\sup_{\mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}} \frac{\sum_{[j(-\mathbf{u}):j(\mathbf{v})]} E_0(N_j) / n}{\prod_{k=1}^d (j(\mathbf{v})_k - j(-\mathbf{u})_k + 1) J_k^{-1}}$$

$$= \sup_{\mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}} \frac{\int \mathbb{1}_{\cup \{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}}(\mathbf{x}) g_0(\mathbf{x}) d\mathbf{x}}{\int \mathbb{1}_{\cup \{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}}(\mathbf{x}) d\mathbf{x}},$$

which is bounded by $g_0(\mathbf{0})$. Thus, (3.28) follows. We conclude that the posterior probability that

$A_{n,2}$ is smaller than a predetermined positive number uniformly for all $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^d \times \mathbb{R}^d$ goes to one in P_0 -probability.

We write

$$B_{n,1}(\mathbf{u}, \mathbf{v}) = P_n(\mathbf{u}, \mathbf{v})/s_n(\mathbf{u}, \mathbf{v}),$$

where $P_n(\mathbf{u}, \mathbf{v})$ is defined in Lemma 3.5.2 and s_n is defined and investigated in the preceding part of the proof. Thus $B_{n,1} \rightsquigarrow \sqrt{g_0(\mathbf{x}_0)}H_1(\mathbf{u}, \mathbf{v})/\prod_{k=1}^d(u_k + v_k)$ on $\mathbb{L}_\infty([0, c] \times [0, c])$.

We can rewrite $B_{n,2}$ as

$$B_{n,2}(\mathbf{u}, \mathbf{v}) = \frac{n\omega_n \int \mathbb{1}_{\cup\{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}}(\mathbf{x})(g_0(\mathbf{x}) - g_0(\mathbf{x}_0))d\mathbf{x}}{s_n(\mathbf{u}, \mathbf{v})} \quad (3.30)$$

Under Assumption 7, by Lemma 1 of Han and Zhang (2020), we see that the mixed derivatives with the order \mathbf{m} being such that $0 < \sum_k m_k/\eta_k < 1$ in the expansion in Assumption 7 must be zero under the multivariate monotonicity condition. By Assumption 7, as the remainder in the expansion is $o(\omega_n)$ uniformly for all $(\mathbf{u}, \mathbf{v}) \in [0, c] \times [0, c]$, the limit of $B_{n,2}(\mathbf{u}, \mathbf{v})$ is the same as that of

$$\sum_{\mathbf{m} \in M} \frac{\partial^{\mathbf{m}} g_0(\mathbf{x}_0)}{m!} \frac{n\omega_n \int_{\cup\{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}} (\mathbf{x} - \mathbf{x}_0)^{\mathbf{m}} d\mathbf{x}}{s_n(\mathbf{u}, \mathbf{v})}.$$

We note that

$$\begin{aligned} & \int_{\cup\{I_j: j \in [j(-\mathbf{u}):j(\mathbf{v})]\}} (\mathbf{x} - \mathbf{x}_0)^{\mathbf{m}} d\mathbf{x} \\ &= \frac{1}{\prod_{k=1}^d (m_k + 1)} \left[\prod_{k=1}^d \left(\frac{[(x_{0,k} + v_k r_{n,k})J_k] - 1}{J_k} - x_{0,k} \right)^{m_k+1} \right. \\ & \quad \left. - \prod_{k=1}^d \left(\frac{[(x_{0,k} - u_k r_{n,k})J_k] - 1}{J_k} - x_{0,k} \right)^{m_k+1} \right]. \end{aligned}$$

As $r_{n,k} = \omega_n^{1/\eta_k}$ and $\sum_{k=1}^d m_k/\eta_k = 1$ for $\mathbf{m} \in M$, we have $\prod_{k=1}^d r_{n,k}^{m_k+1} = \omega_n^{1+\sum_{k=1}^d \eta_k^{-1}}$. Then it follows

that

$$\begin{aligned} & n\omega_n \prod_{k=1}^d \left(\frac{[(x_{0,k} + v_k r_{n,k})J_k]}{J_k} - x_{0,k} \right)^{m_k+1} \\ &= \prod_{k=1}^d (v_k + O((J_k r_{n,k})^{-1}))^{m_k+1} \rightarrow \prod_{k=1}^d v_k^{m_k+1}, \end{aligned}$$

as $J_k \gg r_{n,k}^{-1}$. By the same argument, we have that

$$n\omega_n \prod_{k=1}^d \left(\frac{[(x_{0,k} - u_k r_{n,k})J_k] - 1}{J_k} - x_{0,k} \right)^{m_k+1} \rightarrow \prod_{k=1}^d (-u_k)^{m_k+1}.$$

In view of (3.30), we have shown that

$$B_{n,2}(\mathbf{u}, \mathbf{v}) \rightarrow \sum_{\mathbf{m} \in M} \frac{\partial^{\mathbf{m}} g_0(\mathbf{x}_0)}{(\mathbf{m} + 1)!} \prod_{k=1}^d \frac{v_k^{m_k+1} - (-u_k)^{m_k+1}}{u_k + v_k},$$

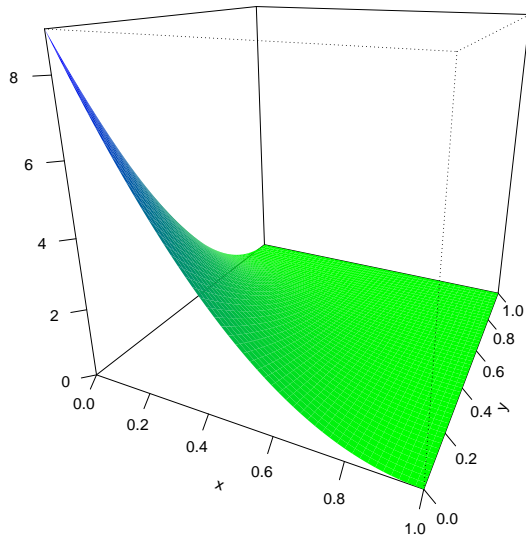
uniformly for $(\mathbf{u}, \mathbf{v}) \in [c^{-\gamma}, c]^d \times [c^{-\gamma}, c]^d$ for any $c > 1$ and $\gamma > 1$.

The second condition of Lemma A.2.1 is shown by Lemma 3.5.5.

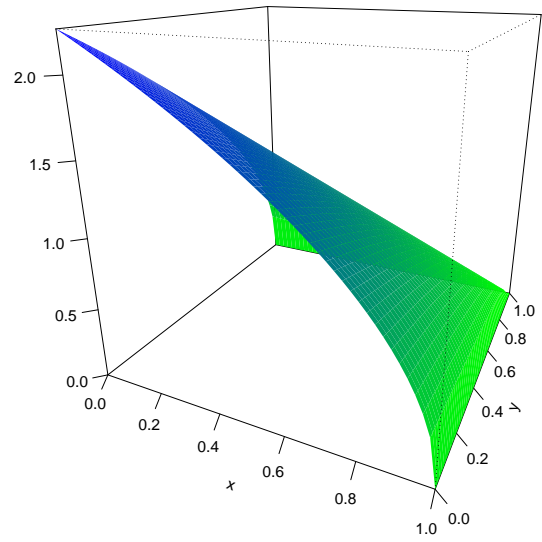
For the third condition, from Proposition 7 in Han and Zhang (2020), we know that

$$\lim_{c \rightarrow \infty} \mathbb{P}(W_c \neq W) = \lim_{c \rightarrow \infty} \int \mathbb{P}(W_c \neq W | H_1) d\mathbb{P}_{H_1} = 0,$$

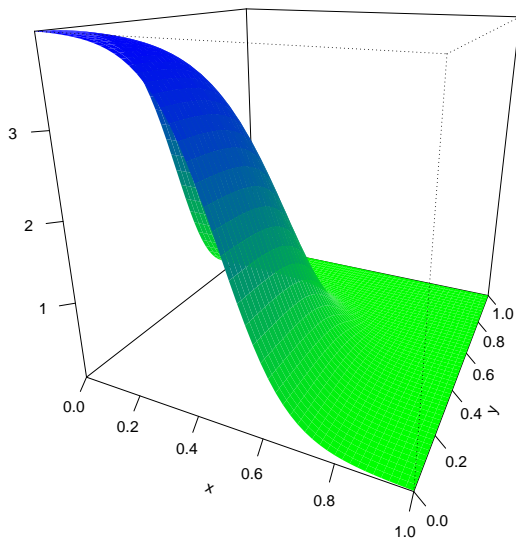
which implies the third condition by Markov's inequality. By verifying all the three conditions in Lemma A.2.1, we conclude the proof of the weak convergence. □



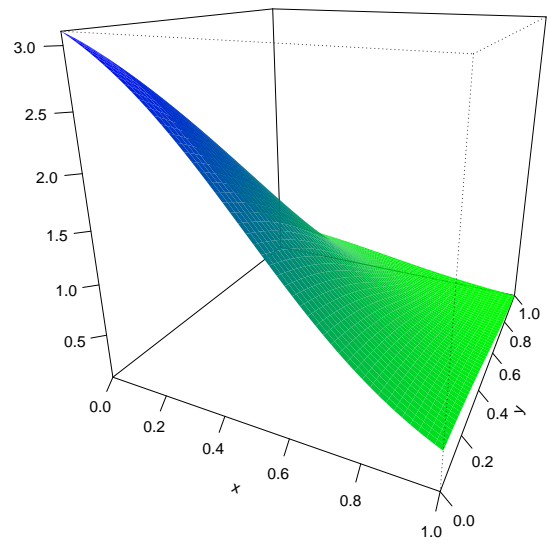
(a) g_1



(b) g_2



(c) g_3



(d) g_4

Figure 3.1: Function graphs of g_0 .

CHAPTER

4

BAYESIAN INFERENCE FOR k -MONOTONE DENSITIES WITH APPLICATIONS TO MULTIPLE TESTING

4.1 Definitions and characterization

Definition 4.1.1 (k -monotonicity). *Let I be subinterval of $(0, \infty)$. A function f on I is said to be 1-monotone on I if f is nonnegative and nonincreasing. For $k \geq 2$, f is said to be k -monotone on I if $(-1)^j f^{(j)}$ is nonnegative, nonincreasing and convex on I , for every $j = 0, \dots, k-2$.*

Let the class of k -monotone functions on I be denoted by \mathcal{F}_I^k . The class of k -monotone probability densities on I will be denoted by $\mathcal{D}_I^k = \{g \in \mathcal{F}_I^k : \int g = 1\}$. We shall be concerned

about k -monotone functions on a bounded interval, which can be taken to be the unit interval $(0, 1)$ without loss of generality. Since the domain is fixed at $(0, 1)$ throughout, we shall drop $(0, 1)$ from the notations $\mathcal{F}_{(0,1)}^k$ and $\mathcal{D}_{(0,1)}^k$, and simply write \mathcal{F}^k and \mathcal{D}^k respectively.

A closely related concept is k -convexity, which is sometimes referred to as k -monotonicity by some authors in the approximation theory literature. There are multiple ways to characterize k -convex functions, and we present an equivalent definition in the following.

Definition 4.1.2 (k -convexity). *A function $f : (0, 1) \rightarrow \mathbb{R}$ is said to be 1-convex on $(0, 1)$ if f is nondecreasing, while for $k \geq 2$, f is said to be k -convex on $(0, 1)$ if $f^{(k-2)}$ exists and is convex on $(0, 1)$. We shall write \mathcal{C}^k for the space of k -convex functions on $(0, 1)$.*

Introduce a probability density function

$$\psi_k(x, \theta) = \frac{k}{\theta} \left(1 - \frac{x}{\theta}\right)_+^{k-1}, \text{ for } x > 0, \theta > 0. \quad (4.1)$$

Note that $\psi_k(\cdot, 1)$ is the probability density function of Beta(1, k). The following result shows that k -monotone functions and densities on $(0, 1)$ admit a useful mixture representation using the kernel ψ_k .

Lemma 4.1.1 (Characterization of k -monotone functions and densities on $(0, 1)$). *A function $f \in \mathcal{F}^k$ if and only if there exist a nondecreasing function $\gamma(t)$ on $(0, 1)$ and $\alpha_j \geq 0$ for $j = 0, 1, \dots, k-2$, such that, for $x \in (0, 1)$,*

$$f(x) = \sum_{j=0}^{k-1} \alpha_j (1-x)^j + \int_0^1 \psi_k(x, t) d\gamma(t). \quad (4.2)$$

A density $g \in \mathcal{D}^k$ if and only if there exists a probability measure Q and $(\beta_j : 0 \leq j \leq k) \in \Delta_{k+1}$ such that, for every $x \in (0, 1)$,

$$g(x) = \sum_{j=0}^{k-1} \beta_j \psi_{j+1}(x, 1) + \beta_k \int_0^1 \psi_k(x, \theta) dQ(\theta). \quad (4.3)$$

The proof of this lemma is based on Taylor expansion and further integration by parts of the integral form remainder term. Similar results can be found in Gao (2008) for a k -monotone distribution function on a compact interval and Williamson (1956) for k -monotone function on the positive half real line. We defer the proof of the lemma to the appendix.

A crucial property of k -monotone functions for deriving posterior contraction rates is that they can be approximated effectively by k -monotone free-knot spline functions in the \mathbb{L}_p -metric, $1 \leq p < \infty$. This property is derived from Theorem 1.1 of Kopotun and Shadrin (2003) on a shape-preserving approximation of k -convex functions.

Let $\mathcal{S}_{N,k}$ denote the space of free knot splines of degree $k - 1$ with N interior knots in $[0, 1]$. To align with the k -convex functions, we introduce a reflection transformation to the argument. Let $\tau(x) = 1 - x$ for $x \in (0, 1)$ and denote $\check{\mathcal{F}}^k = \{f \circ \tau : f \in \mathcal{F}^k\}$. Then shape preserving approximation to \mathcal{F}^k is essentially the same problem of shape-preserving approximation to $\check{\mathcal{F}}^k$. By Definition 4.1.1, for $k \geq 2$, $f \in \check{\mathcal{F}}^k$ if and only if $f^{(j)}$ is nonnegative, nondecreasing, and convex, for every $j = 0, 1, \dots, k - 2$. It is then clear that $\check{\mathcal{F}}^k$ is a subclass of \mathcal{C}^k . Moreover, for $h \in \mathcal{C}^k$ and $k \geq 2$, let $h^{(k-1)}$ denote the right derivative of $h^{(k-2)}$, which is well defined since h is a convex function on $(0, 1)$. We also know that $h^{(k-1)}$ is right continuous. It is not hard to see that $h \in \check{\mathcal{F}}^k$ as well if $h^{(j)}(0+) \geq 0$, for $j = 0, \dots, k - 1$. Indeed, $h^{(k-1)}(0+) \geq 0$ implies that $h^{(k-2)}$ is nondecreasing, and furthermore, it is known that $h^{(k-2)}$ is nonnegative, nondecreasing, and convex. Continuing in the same way, we know that $h^{(j)}$ is nonnegative, nondecreasing, and convex for all $j = 0, \dots, k - 2$, that is, $h \in \check{\mathcal{F}}^k$ by definition. In view of this point, for $f \in \check{\mathcal{F}}^k \subset \mathcal{C}^k$, the shape preserving approximation by a free knot spline function $s \in \mathcal{S}_{N,k} \cap \mathcal{C}^k$ considered in Kopotun and Shadrin (2003) is also a shape preserving approximation in $\check{\mathcal{F}}^k$ (i.e. $s \in \mathcal{S}_{N,k} \cap \check{\mathcal{F}}^k$) provided that $s^{(j)}(0) \geq 0$ for all $j = 0, \dots, k - 1$. By close inspection of the construction of approximating function in Kopotun and Shadrin (2003), this set of conditions is naturally satisfied. We leave the details of the argument in the appendix.

The main result in Kopotun and Shadrin (2003) states that the shape-preserving approximation by free knot splines can be as good as the free knot spline approximation regarding the

number of splines used to construct the approximation function and the approximation error. In fact, Theorem 1.1 of Kopotun and Shadrin (2003) presents a more general result. In what follows, we only use their result with the order of free knot splines fixed at k (i.e. approximation by piecewise polynomials of degree $k - 1$) as this is the only case of interest in the current work.

Proposition 4.1.1 (Theorem 1.1 of Kopotun and Shadrin (2003)). *For any $1 \leq p \leq \infty$ and any $f \in \mathcal{C}^k \cap \mathbb{L}_p(0, 1)$, there exist constants $C_k > 0$ and $C_{k,p} > 0$ such that*

$$d_p(f, \mathcal{S}_{C_k N, k} \cap \mathcal{C}^k) \leq C_{k,p} d_p(f, \mathcal{S}_{N, k}).$$

On the other hand, the approximation error of free-knot splines is well-studied in approximation theory, as can be found in Chapter 12 of DeVore and Lorentz (1993). If the $(k - 1)$ -th derivative of f is bounded, the right-hand side of the last display is bounded by N^{-k} up to some positive constant. Moreover, the shape-preserving approximation to a k -monotone function is not hard to be adapted to the shape-preserving approximation to the k -monotone density. With the help of Lemma 4.1.1, the free knot spline approximation of order k with N interior knots admits a representation as in (4.3), indicating that the mixing distribution Q is supported on a set of at most N points.

To summarize, we obtain the following approximation result, whose proof is deferred to the appendix.

Lemma 4.1.2. *Let $g \in \mathcal{D}^k$ be given by (4.3) such that $|g^{(k-1)}(0+)| < \infty$. Then there exists a discrete probability measure Q_N with N support points in $(0, 1)$ such that with*

$$g_N(x) = \sum_{j=0}^{k-1} \beta_j \psi_{j+1}(x, 1) + \beta_k \int_0^1 \psi_k(x, \theta) dQ_N(\theta) \in \mathcal{D}^k, \quad (4.4)$$

we have that $\|g - g_N\|_\infty \leq C N^{-k}$ for some constant $C > 0$.

4.2 Posterior Contraction Rates

Let $\mathbf{X}_n = (X_1, \dots, X_n)$ be independent and identically distributed (i.i.d.) samples from a k -monotone density g given by the representation (4.3) for a known $k = 1, 2, \dots$. To place a prior on g , it is natural to consider independent priors for the coefficient vector $\beta = (\beta_0, \dots, \beta_k)$ and the mixing distribution Q .

We put a Dirichlet distribution prior on β with parameters $0 < a_j < \infty$, for all $j = 0, 1, \dots, k$. Independently of β , we assign either a Dirichlet process (DP) prior or a Finite Mixtures (FM) prior on Q :

DP: $Q \sim \text{DP}_{aH}$, where $a > 0$ is the precision parameter and H is the center measure supported on $(0, 1)$; see Ghosal and van der Vaart (2017) for definitions;

FM: $Q = \sum_{j=1}^J w_j \delta_{\theta_j}$, with $\theta_1, \dots, \theta_J | J \stackrel{i.i.d.}{\sim} H$, and $(w_1, \dots, w_J) | J \sim \text{Dir}(J; \omega_{1J}, \dots, \omega_{JJ})$, independently. J is given a prior $\Pi(J) = (n^c - 1)n^{-cJ}$ on the set of positive integers.

In the above priors, a , H , $(\omega_{jj} : 1 \leq j \leq J < \infty)$ and c are hyperparameters. We assume that

(C1) H admits a Lebesgue density p_H on $(0, 1)$ such that in a small neighborhood of zero,

$$p_H(\theta) \lesssim \theta^{t_1} \text{ for some } t_1 > 0;$$

(C2) for any interval $(u, v) \subset (0, 1)$ and some $t_2 > 0$, $H((u, v)) \gtrsim (v - u)^{t_2}$.

If $g \in \mathcal{D}^k$ for $k \geq 2$, it is assumed that g is differentiable only up to order $k - 2$. However, $(-1)^{k-2} g^{(k-2)}$ is convex and non-increasing on $(0, 1)$. Hence, we can define $g^{(k-1)}$ uniquely almost everywhere as either the left or right derivative of $g^{(k-2)}$, which are equal except possibly on an at most countable set.

Theorem 4.2.1 (Contraction rate for Dirichlet process mixture prior). *Let the data \mathbf{X}_n be generated from a k -monotone density g_0 on $(0, 1)$ given by*

$$g_0(x) = \sum_{j=0}^{k-1} \beta_{0,j} \psi_{j+1}(x, 1) + \beta_{0,k} \int \psi_k(x, \theta) dQ_0(\theta), \quad (4.5)$$

where k is known. We assume $g_0^{(k-1)}(0+) < \infty$ and $\beta_{0,0} > 0$. Let β be given a Dirichlet prior with positive constant parameters, a_0, \dots, a_k , and independently, put a Dirichlet process prior on Q satisfying Conditions (C1) and (C2). Then the posterior distribution of g contracts at the rate $\epsilon_n = (n/\log n)^{-k/(2k+1)}$ at g_0 with respect to the Hellinger distance, i.e., $E_0[\Pi(d_H(g, g_0) \geq M_n \epsilon_n | \mathbf{X}_n)] \rightarrow 0$ for any $M_n \rightarrow \infty$.

The same posterior contraction rate can be obtained by using a finite mixture prior on the mixing distribution, as presented in the following theorem.

Theorem 4.2.2 (Contraction rate for finite mixture prior). *Let the data \mathbf{X}_n be generated from a k -monotone density g_0 on $(0, 1)$ given by (4.5) with a known k , satisfying $g_0^{(k-1)}(0+) < \infty$ and $\beta_{0,0} > 0$. Let β be given the Dirichlet prior with positive constant parameters a_0, \dots, a_k , and independently, put a finite mixture prior for Q satisfying Conditions (C1) and (C2), with $c > 0$ chosen sufficiently large. Then the posterior distribution of g contracts at g_0 at the rate $\epsilon_n = (n/\log n)^{-k/(2k+1)}$ with respect to the Hellinger distance.*

Remark 6. In Theorem 4.2.2, the same posterior contraction rate can be derived if the prior on J is replaced by a fixed prior that satisfies the condition, $e^{-b_1 j \log j} \leq \Pi(J = j) \leq e^{-b_2 j \log j}$. For instance, a Poisson prior truncated at 0 satisfies the required tail condition.

The posterior contraction rate is substantially improved to a nearly parametric rate using the same prior if the mixing distribution Q_0 is finitely supported on J_0 points, i.e.,

$$g_0(x) = \sum_{j=0}^{k-1} \beta_{0,j} \psi_{j+1}(x, 1) + \beta_{0,k} \sum_{l=1}^{J_0} w_l^0 \psi_k(x, \theta_j^0). \quad (4.6)$$

In the result below, both k and J_0 are allowed to depend on n (and hence the resulting rate involves k and J_0), provided that $\max(\log k, \log J_0) \lesssim \log n$.

Theorem 4.2.3 (Finitely supported mixing). *Let the true density g_0 as given in (4.6) with $\beta_{0,0} > 0$. Let β be given the Dirichlet prior with positive constant parameters a_0, \dots, a_k , and independently, let Q be given a finite mixture prior with $c > 0$ chosen sufficiently large and H satisfying the*

conditions that $H((u, v)) \gtrsim (v - u)^{t_2}$ and $p_H \lesssim \exp\{-t_3/\theta\}$ for any interval $(u, v) \subset (n^{-2}, 1)$, and $\theta \in (0, n^{-2})$, where $t_2, t_3 > 0$ are constants. Then the posterior of g contracts at g_0 at the rate $\epsilon_n = \sqrt{\max(k, J_0)(\log n)/n}$ with respect to the Hellinger metric.

It can be seen from the proof that the fixed prior in Remark 6 also obtains the same rate if $\log J_0 \asymp \log n$.

4.3 Adaptation to k

In the last section, we studied posterior contraction rates assuming that the order of monotonicity k is known. Here, the parameter k serves as a regularity index controlling the complexity of the model, much like a smoothness index. Adapting the rates to different values of k is therefore a highly desirable objective. In the Bayesian framework, a natural approach is to treat k as a model index parameter and put a prior distribution on it. Consequently, the resulting prior becomes a mixture of the priors used for a fixed index value. Under similar situations in smoothness or sparsity settings, the corresponding posterior distribution often adapts to the optimal rate under fairly mild conditions; see Ghosal et al. (2008a); Ghosal and van der Vaart (2017); Castillo et al. (2015), among others. In this section, we show that such an automatic adaptation strategy works in the k -monotone setting as well. This feature is particularly attractive by employing the Bayesian approach, while no parallel result is known in the non-Bayesian literature for the family of k -monotone densities indexed by k . Noting the models \mathcal{D}^k are nested in the following way: $\mathcal{D}^{k+1} \subset \mathcal{D}^k$, we define the true value k_0 as the largest value of k such that $g_0 \in \mathcal{D}^k$. We assume k_0 is finite, which is the case of interest. It is not hard to see that the finiteness of k_0 implies $\beta_{k_0} > 0$ in the characterization (4.3). Otherwise, this k_0 -monotone density would be a polynomial of degree at most $k_0 - 1$, which would correspond to the case $k_0 = \infty$, contradicting the finiteness of k_0 .

Let k be given a prior Π that is one of the two types:

$$(K1) \quad e^{-d_1 k \log k} \leq \Pi(k) \leq e^{-d_2 k \log k} \text{ for some } d_1 \geq d_2 > 0;$$

(K2) $\Pi(k) = (n^r - 1)n^{-rk}$, for some $r > 0$.

Theorem 4.3.1 (Adaptive contraction rate). *Let the monotonicity index k be unknown and endowed with a prior satisfying the conditions (K1) or (K2). Given k , let the prior for β be $\text{Dir}(a_0, \dots, a_k)$ for some a_0, \dots, a_k lying between two fixed positive numbers, and independently, let Q be given either the Dirichlet process prior or the finite mixture prior with a sufficiently large $c > 0$, satisfying Conditions (C1) and (C2). Let the true density be g_0 be given by (4.5) with $k = k_0$ satisfying $|g_0^{(k_0-1)}(0+)| < \infty$ and $\beta_{0,0} > 0$. Then the posterior distribution contracts at g_0 at the rate $\epsilon_n = (n/\log n)^{-k_0/(2k_0+1)}$ with respect to the Hellinger metric.*

If the true k -monotone density has a finite representation as in (4.6) with $k = k_0$, then it is still possible to obtain the nearly parametric posterior contraction rate stated in Theorem 4.2.3 without knowing the true value k_0 of k . As shown in Theorem 4.2.3, this holds when both k_0 and J_0 satisfy that $\max(\log J_0, \log k_0) \lesssim \log n$. It may be noted that, even though $g \in \mathcal{D}^{\bar{k}}$ for any $\bar{k} < k_0$ as well, the corresponding mixture representation with the kernel $\psi_{\bar{k}}$ will not be supported on finitely many points in general.

Theorem 4.3.2 (Adaptive contraction for finite mixture). *Let the monotonicity index k be unknown and endowed with a prior of the type (K2). Given k , let the prior for β be $\text{Dir}(a_0, \dots, a_k)$ for some a_0, \dots, a_k lying between two fixed positive numbers. Independently, let Q be given the finite mixture prior with a sufficiently large $c > 0$. The prior distribution H of a support point θ satisfies, for some $t_2, t_3 > 0$, that $H((u, v)) \gtrsim (v - u)^{t_2}$ for any interval $(u, v) \subset (n^{-2}, 1)$, and the corresponding density $p_H(\theta) \lesssim \exp\{-t_3/\theta\}$ for all $\theta \in (0, n^{-2})$. If g_0 is given by (4.6) for some finite J_0 and $k = k_0$ with $\beta_{0,0} > 0$, then the posterior contracts at g_0 at the rate $\sqrt{\max(k_0, J_0)(\log n)/n}$ with respect to the Hellinger metric.*

It can be seen from the proof that a prior of the type (K1) for k and a fixed prior for J as in Remark 6 may also be used to derive the same rate provided that $\log k_0 \asymp \log n$ and $\log J_0 \asymp \log n$.

4.4 Applications to Multiple Testing

In large-scale hypothesis testing, it is essential to assess the proportion of true null hypotheses when reporting scientific findings. The proportion of null hypotheses, denoted as α , plays a crucial role in the calculation of the positive false discovery rate Storey (2002). Consider a problem of simultaneously testing n hypotheses. For each individual test, the data are summarized using a test statistic, and a p -value is computed based on an exact, approximate, or asymptotic null distribution of the test statistic and the scope of the alternative hypothesis. Furthermore, we assume that the test statistics corresponding to different hypotheses are (nearly) independent, resulting in (nearly) independent p -values. Under a simple null hypothesis, the p -value is calibrated; that is, it has a uniform distribution on $[0, 1]$, provided that the test statistic follows a continuous null distribution. Even when the null hypothesis is composite, certain Bayesian p -values (e.g., the partial posterior predictive p -value of Bayarri and Berger (2000)) asymptotically follow a uniform distribution (cf. Robins et al. (2000)) when the data are sampled using an i.i.d. scheme. The p -values from the alternative hypotheses usually concentrate near the origin and have decreasing density on $[0, 1]$. This feature, along with true null hypotheses outnumbering true alternative hypotheses in practice, is used to estimate the proportion of null hypotheses in Storey's procedure Storey (2002) for controlling the positive false discovery rate (pFDR). It is easy to see that the proportion of the null hypothesis is identifiable if the p -value density under the alternative approaches 0 at 1 (Proposition 4 of Ghosal et al. (2008b)). This assumption is not always true; however, see the discussion in Section 2.2 of Ghosal et al. (2008b). For example, in the two-sided t-test, the density of p -values does not vanish at 1, in which case we can only identify an upper bound for the proportion of null hypotheses. However, if the sample size is reasonably large, the height of the density under the alternative is very small, so the condition holds approximately.

The p -value density under the alternative is explicitly modeled as a monotone decreasing density (k -monotone for $k = 1$) in Langaas et al. (2005). This assumption is extremely mild as it can be seen to hold under the Monotone Likelihood Ratio (MLR) property of the distri-

bution of the test statistic for both one- and two-sided alternatives (Propositions 1 and 2 of Ghosal et al. (2008b)). However, simulation results demonstrate that the Grenander estimator exhibits unstable performance near 1, which significantly affects the quality of the estimator for the positive false discovery rate (pFDR). To enhance the performance, Langaas et al. (2005) recommends using a convex nonincreasing density to fit the density of the p -values. A model-based Bayesian approach to the estimation of the pFDR was adopted in Tang et al. (2007) using certain mixtures of beta densities. The corresponding distribution function under a logarithmic transformation of the argument is completely monotone (Proposition 7 of Ghosal et al. (2008b)), which corresponds to k -monotonicity for all k . Results in Section 3 of Ghosal et al. (2008b) show that the Bayesian procedure under a Dirichlet process prior on the mixing distribution gives consistent posterior for the proportion of null hypotheses and the pFDR. Other model-based and Bayesian approaches to the estimation of pFDR have been proposed based on modeling probit-transformed mixtures of skew-normal densities in Bean et al. (2013) and Ghosal and Roy (2011b) and sufficient conditions for the identification of the proportion of null hypotheses are discussed in Ghosal and Roy (2011a). A review of Bayesian nonparametric methods for multiple testing is available in Ghosal and Roy (2009).

A very appealing condition on the p -value density under the alternative compromising between the generality of the class of monotone densities and the smoothness of the class of completely monotone functions is that the density of the p -value distribution under the alternative belongs to the class of k -monotone density for some k . For instance, the case $k = 2$ corresponding to decreasing convex densities already gives a much more stable estimator of the density (see Langaas et al. (2005)), but it may be harder to ensure under what condition the density of p -values under the alternative would be decreasing and convex. The approach to modeling the density of the p -values under the alternative as a k -monotone density is irresistibly appealing if k can be left unspecified and be adaptively chosen from the data using the technique developed in Section 4.3. The following result quantifies the accuracy of the procedure.

Theorem 4.4.1. *Let U_1, \dots, U_n be independent p -values arising from the simultaneous testing of n hypotheses. We assume that the p -value density g is modeled as k -monotone, where $k \geq 2$. The value of k can be either known or unknown. In the latter case, a prior on g is specified as described in Section 4.2 or 4.3. In both scenarios, α represents the corresponding proportion of null hypotheses. Let g_0 stand for the true density and let the true proportion of null hypotheses be denoted by α_0 . Then under the conditions of Theorems 4.2.1, 4.2.2 or 4.3.1, the posterior distribution of α is consistent at α_0 and contracts at the rate $\epsilon_n = (n/\log n)^{-k/(2(2k+1))}$, that is, for any $M_n \rightarrow \infty$, $\Pi(|\alpha - \alpha_0| > M_n \epsilon_n | U_1, \dots, U_n) \rightarrow 0$ in probability.*

4.5 Simulation

We implement the proposed Bayesian approach for a k -monotone density estimation. Specifically, we employ the Dirichlet process prior for the mixing distribution. To simplify, we only retain the additional uniform component and the mixture component of k -monotone kernels in computation. We consider both scenarios: when the value of k is known and when it is unknown. Simulation results demonstrate the superiority of our method compared to nonparametric maximum likelihood estimation for monotone density, as well as for convex and monotonically nonincreasing density. We present the specifics of our simulation in the following sections.

4.5.1 Estimation accuracy

To perform posterior sampling under the Dirichlet process mixture prior, we utilize the sliced Gibbs sampling algorithm as described in Kalli et al. (2011). We use a uniform base measure on $[n^{-1}, 1]$ for the Dirichlet process prior and set the precision parameter to a fixed value 1. For the simplified model, we set $\beta_j = 0$ for $j = 1, \dots, k - 1$, and give the proportion of the uniform component β_0 a uniform prior on $[0, 1]$. We let k be fixed or assign an appropriate prior for k . In particular, when using the adaptive Bayesian approach, the prior on k is uniformly distributed

over the set $1, \dots, 10$. We select the largest k at 10, which is sufficiently large to approximate common smoothly decreasing densities of interest. In the following, we generate 1000 posterior samples, based on which we make inferences on the unknown density function after dropping the first 2000 burn-in ones in every Bayesian application.

Let the sequence $\theta_{j,J} = j/J$. We consider the following density functions:

- $g_1(x) = \psi_2(x, 1) = 2(1 - x)$,
- $g_2(x) = 0.5f_1(x) + 0.5 = 1.5 - x$,
- $g_3(x) = \sum_{j=1}^3 3^{-1} \psi_2(x, \theta_{j,3})$,
- $g_4(x) = 0.5f_3(x) + 0.5$,
- $g_5(x) = \sum_{j=1}^3 3^{-1} \psi_4(x, \theta_{j,3})$,
- $g_6(x) = \int_0^1 \psi_4(x, \theta) 2\theta d\theta$.

For g_6 , the mixing distribution for θ is Beta(2, 1), and sampling according to g_6 is straightforward.

We take sample sizes of $n = 100, 200$, and 500. For every sample size, we generate independent and identically distributed samples with all six aforementioned models. The proposed Bayesian procedure is applied to each dataset, accounting for both known and unknown values of k . To compare with non-Bayesian methods, we apply the posterior mean density function as the Bayesian estimator. We consider the classical Grenander estimator, as well as the non-parametric maximum likelihood estimator, for convex and nonincreasing densities on the interval $[0, 1]$. To measure the deviation from the true density functions, for each estimate \hat{g} , we compute the mean squared error (MSE) over a grid in $[0, 1]$. This grid is defined as $x_{j,K} = j/K$ for $j = 1, \dots, K$. The MSE is then calculated as follows:

$$\text{MSE}(\hat{g}) = \frac{1}{K} \sum_{j=1}^K (\hat{g}(x_{j,K}) - g_0(x_{j,K}))^2.$$

Here we choose $K = 100$ and f_0 stands for the corresponding f_i , $i = 1, \dots, 6$.

We independently conduct $R = 500$ iterations for each setup and present the average mean squared error (MSE) calculated in Table 4.1. Each row in Table 4.1 corresponds to a specific method applied to the dataset with the corresponding sample size. “Bay” denotes the Dirichlet mixture model with a known k . “Ada” represents the Bayesian methods where k is unknown. “Con” and “Gre” stand for the nonparametric maximum likelihood estimation for the convex and nonincreasing density class and for the nonincreasing density class, respectively.

Table 4.1: Average of MSE.

		g_1	g_2	g_3	g_4	g_5	g_6
$n = 100$	Bay	0.018	0.018	0.027	0.018	0.029	0.028
	Ada	0.024	0.023	0.027	0.026	0.030	0.031
	Con	0.019	0.022	0.041	0.032	0.068	0.076
	Gre	0.058	0.047	0.097	0.068	0.158	0.162
$n = 200$	Bay	0.009	0.011	0.017	0.011	0.021	0.017
	Ada	0.014	0.013	0.016	0.014	0.019	0.017
	Con	0.010	0.011	0.024	0.017	0.040	0.041
	Gre	0.036	0.029	0.058	0.041	0.102	0.102
$n = 500$	Bay	0.003	0.005	0.008	0.006	0.010	0.010
	Ada	0.003	0.006	0.008	0.007	0.014	0.015
	Con	0.004	0.005	0.010	0.008	0.018	0.020
	Gre	0.018	0.015	0.029	0.022	0.052	0.053

In summary, the proposed Bayesian methods for both known and unknown values of k demonstrate superiority over nonparametric likelihood estimations. The adaptive Bayesian approach performs nearly as well as the Bayesian method that employs the optimal choice of k .

4.5.2 Estimation of the proportion of null hypotheses

The simulation setup in this part closely follows that of Langaas et al. (2005). Here, we simulate DNA microarray data that involves multiple hypothesis testing problems. For each of the

m individuals, we collect a dataset of sample size n , denoted by $\mathbf{X}_j = (X_{1,j}, \dots, X_{n,j})$ for $j = 1, \dots, m$, independently drawn from a multi-normal distribution, i.e., $\mathbf{X}_j \stackrel{i.i.d.}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$. We test the hypotheses

$$H_{0,i} : \mu_i = 0, \text{ versus } H_{0,i} : \mu_i \neq 0,$$

based on the t-test statistics. For comparison, we also consider the set of one-sided t-tests,

$$H_{0,i} : \mu_i = 0, \text{ versus } H_{0,i} : \mu_i > 0.$$

The mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are generated as follows. For the null proportion, $\alpha_0 \in \{0.5, 0.8, 0.9, 0.95\}$, we generate a binomial variable n_0 with parameter n and α_0 . Randomly selected n_0 positions among n and set the corresponding $\mu_i = 0$. For the two-sided t-tests, we generate the alternative means by independently sampling from the symmetric bitriangular distribution with parameter $a = \log_2 1.2$ and $b = 2$; see Langaas et al. (2005) for details. For one-sided t-tests, the remaining μ_i are independently generated from the symmetric triangular distribution with the same parameters as in the previous case. To consider the effect of correlation between tests, we consider a specific block diagonal structure for $\boldsymbol{\Sigma}$. We choose a block size $G \in \{50, 100\}$. The within-block correlation ρ takes values in $\{0, 0.25, 0.5, 0.75\}$. Between blocks, the coordinates are independent. Note that $\rho = 0$ means all the tests are pairwise independent.

We choose $n = 2000$ and $m = 10$ throughout. For the Bayesian approach, we continue to use the Dirichlet process mixture prior with parameters defined as in the previous section while considering k as an unknown parameter with a prior distribution. The proposed estimator is the posterior mean of β_0 . For comparison, we also estimate the null proportion by the maximum likelihood estimation for convex and decreasing density class, as proposed in Langaas et al. (2005). This can be easily computed using the `convest` function in the R package `limma`. Each setting is replicated 1000 times, and the densities of these two estimators are plotted in Figure

4.1 – Figure 4.4.

For both two-sided and one-sided tests, the simulation results exhibit a similar pattern. The presence of correlation between tests has a detrimental effect on the performance of both methods. However, our Bayesian procedure demonstrates more stable performance in the presence of within-group correlation and for cases with larger blocks of correlation. Furthermore, our method shows more accurate estimation performance when the proportion of null hypotheses is relatively large. However, when α_0 is moderate, such as 0.5, the convex maximum likelihood estimator appears to be less biased. These findings suggest that our Bayesian approach offers advantages in handling correlated tests and estimating the null proportion accurately, particularly when there is a higher proportion of null hypotheses, which is the very common case. These observations highlight the strengths and limitations of both methods in different scenarios, providing valuable insights into their respective performance characteristics.

4.6 Proofs

Proof of Theorem 4.2.1. We apply the general theory of posterior contraction for i.i.d. observations as in Section 8.2 of Ghosal and van der Vaart (2017). We need to obtain a lower bound for prior concentration in a Kullback-Leibler neighborhood of the true density and bound the size of a sieve in terms of the metric entropy so that the remaining part of the parameter space has an exponentially small prior probability.

We verify the first condition given by (8.4) of Ghosal and van der Vaart (2017) at any g_0 with ϵ_n a constant multiple of $(n/\log n)^{-k/(2k+1)}$:

$$-\log \Pi(K(g_0, g) \leq \epsilon_n^2, V(g_0, g) \leq \epsilon_n^2) \lesssim n \epsilon_n^2. \quad (4.7)$$

By Lemma 4.1.2, there exists $g^*(x) = \sum_{j=0}^{k-1} \beta_j^* \psi_{j+1}(x, 1) + \beta_k^* \sum_{l=1}^{J^*} w_l^* \psi_k(x, \theta_l^*) \in \mathcal{D}^k$ with $(w_l^* : l = 1, \dots, J^*) \in \Delta_{J^*}$ and $(\theta_l^* : l = 1, \dots, J^*) \in (0, 1)^{J^*}$, such that $J^* \lesssim \epsilon_n^{-1/k}$ and $\|g_0 - g^*\|_\infty \lesssim \epsilon_n$.

First, we show that we can maintain the same approximation rate by restricting the choice to $\theta_j^* \notin (0, \epsilon_n^2)$ to ensure that $\|g_0 - g^*\|_2 \lesssim \epsilon_n$. Indeed, if there are $\theta_l^* < \epsilon_n^2$, we write $\bar{w} = \sum_{l: \theta_l^* < \epsilon_n^2} w_l^*$ and define

$$g^\dagger(x) = \sum_{j=0}^{k-1} \beta_j^* \psi_{j+1}(x, 1) + \beta_k^* \sum_{l: \theta_l^* \geq \epsilon_n^2} w_l^* \psi_k(x, \theta_l^*) + \beta_k^* \bar{w} \psi_k(x, \epsilon_n^2).$$

It follows that $g^\dagger \in \mathcal{D}^k$ and $g^\dagger(x) = g^*(x)$ for all $\epsilon_n^2 \leq x < 1$. Since, g_0 is bounded and $\|g^* - g_0\|_\infty \lesssim \epsilon_n$, clearly g^* is bounded. As

$$g^*(0+) - g^\dagger(0+) = \beta_k^* \sum_{l: \theta_l^* < \epsilon_n^2} w_l^* (\psi_k(0, \theta_l^*) - \psi_k(0, \epsilon_n^2)) = \beta_k^* \sum_{l: \theta_l^* < \epsilon_n^2} w_l^* \left(\frac{k}{\theta_l} - \frac{k}{\epsilon_n^2} \right)$$

is nonnegative, g^* is bounded as well. Now $\|g^\dagger - g^*\|_2 \lesssim \|\mathbb{1}_{(0, \epsilon_n^2)}\|_2 = \epsilon_n$, and hence $\|g^\dagger - g_0\|_2 \lesssim \epsilon_n$. This assures that we can assume without loss of generality that $\theta_l^* \geq \epsilon_n^2$ for an \mathbb{L}_2 -approximation of g_0 within the order of ϵ_n using $J^* \lesssim \epsilon_n^{-1/k}$ mixture components.

To bound $K(g_0, g)$ and $V(g_0, g)$, we first bound the Hellinger distance. As

$$d_H(g_0, g) \leq d_H(g_0, g^*) + d_H(g, g^*),$$

and $d_H(g_0, g^*) \leq \|1/g_0\|_\infty^{1/2} \|g - g_0\|_2 \lesssim \epsilon_n$, it suffices to bound $d_H(g, g^*)$. Let $I_j = (\theta_l^* - \epsilon_n^4/2, \theta_l^* + \epsilon_n^4/2)$, $l = 1, \dots, J^*$. We can assume, without loss of generality, that all spacings between $\theta_1^*, \dots, \theta_{J^*}^*$ are bigger than ϵ_n^4 . Indeed, as $\min \theta_l^* \geq \epsilon_n^2$, Lemma B.1.2 eliminates the need for placing multiple support points within ϵ_n^4 -neighborhood to control the \mathbb{L}_1 -distance within a constant multiple of ϵ_n^2 . This implies that I_l , $l = 1, \dots, J^*$, can be assumed to be pairwise disjoint. Let $I_0 = (0, 1) \setminus (\cup_{l=1}^{J^*} I_l)$. Then

$$\begin{aligned} \|g - g^*\|_1 &\leq \sum_{j=0}^{k-1} |\beta_j - \beta_j^*| + \sum_{l=1}^{J^*} \int_{I_l} \|\psi_k(\cdot, \theta) - \psi_k(\cdot, \theta_l^*)\|_1 dQ(\theta) \\ &\quad + \sum_{l=1}^{J^*} |Q(I_l) - w_l^*| + Q(I_0). \end{aligned} \tag{4.8}$$

The fourth term in (4.8) is bounded by the third term because $Q(I_0) = 1 - \sum_{l=1}^{J^*} Q(I_l)$ and $\sum_{l=1}^{J^*} w_l^* = 1$. Since $\|\psi_k(\cdot, \theta) - \psi_k(\cdot, \theta_j^*)\|_1 \leq 2|\theta - \theta_j^*| / \min \theta_l^* \leq 2\epsilon_n^2$ for any $\theta \in I_l$, and $\sum_{l=1}^{J^*} Q(I_l) \leq 1$, the second term is bounded by $\sum_{l=1}^{J^*} 2\epsilon_n^2 Q(I_l) \leq 2\epsilon_n^2$. Therefore $\sum_{j=0}^{k-1} |\beta_j^* - \beta_j| \leq \epsilon_n^2$ and $\sum_{l=1}^{J^*} |Q(I_l) - w_l^*| \leq \epsilon_n^2$ together ensure that $\|g - g^*\|_1 \lesssim \epsilon_n^2$, and hence $d_H(g, g^*) \lesssim \epsilon_n$. Therefore, by Lemma B.1.3, it follows that

$$\max(K(g_0, g), V(g_0, g)) \lesssim \epsilon_n^2,$$

so it suffices to lower-bound $\Pi\{\sum_{j=0}^{k-1} |\beta_j - \beta_{0,j}| \leq \epsilon_n^2, \sum_{l=1}^{J^*} |Q(I_l) - w_l^*| \leq \epsilon_n^2\}$. By Lemma G.13 of Ghosal and van der Vaart (2017), under the assumed conditions on the center measure H , for some constant $C, C' > 0$, we have $\Pi(\sum_{j=1}^{k-1} |\beta_j - \beta_{0,j}| \leq \epsilon_n^2) \gtrsim e^{-Ck \log(1/\epsilon_n)}$ and $\Pi(\sum_{l=1}^{J^*} |Q(I_l) - w_l^*| \leq \epsilon_n^2) \gtrsim e^{-C'J^* \log(1/\epsilon_n)}$. As β and Q are independent, it follows that

$$-\log \Pi\left(\sum_{j=0}^{k-1} |\beta_j - \beta_{0,j}| \leq \epsilon_n^2, \sum_{l=1}^{J^*} |Q(I_l) - w_l^*| \leq \epsilon_n^2\right) \lesssim J^* \log(1/\epsilon_n). \quad (4.9)$$

Equating $J^* \log(1/\epsilon_n)$ with $n\epsilon_n^2$, it is now immediate that (4.7) holds for ϵ_n a constant multiple of $(n/\log n)^{-k/(2k+1)}$.

Define a sieve $\mathcal{D}_n^k = \{g \in \mathcal{D}^k : g(0+) \leq M_n\}$, where $M_n = \exp\{C n^{1/(2k+1)} (\log n)^{2k/(2k+1)} (t_1 + 1)^{-1}\}$ and C is a larger enough positive constant, to be determined later. Instead of verifying the metric entropy condition on \mathcal{D}_n^k , we use the following one from Theorem 5.1 of Rivoirard and Rousseau (2012):

For any $j \in \mathbb{N}$, let $\mathcal{D}_{n,j}^k = \{f \in \mathcal{D}_n^k : j\epsilon_n \leq d_H(f_0, f) \leq (j+1)\epsilon_n\}$. There exists $J_{0,n}$ such that for all $j \geq J_{0,n}$,

$$\log \mathcal{N}(j\epsilon_n/2, \mathcal{D}_{n,j}^k, d_H) \lesssim n\epsilon_n^2 j^2. \quad (4.10)$$

Suppose $M_n \geq 1$, for every $g \in \mathcal{D}_n^k$, $g(0+) \leq M_n$. Let $a_n \in (0, 1)$ such that $\epsilon_n/2 \leq a_n \leq \epsilon_n$. For

every $f \in \mathcal{D}_{n,j}^k$, since $\|g - g_0\|_1 \leq 2d_H(g_0, g) \leq 2(j+1)\epsilon_n$, we have

$$\int_0^{a_n} g(x)dx - \int_0^{a_n} g_0(x)dx \leq 2(j+1)\epsilon_n,$$

which implies

$$\int_0^{a_n} g(x)dx \leq (2j+2 + g_0(0+))\epsilon_n, \text{ and } a_n g(a_n) \leq a_n g_0(0+) + 2(j+1)\epsilon_n,$$

and then

$$g(a_n) \leq g_0(0+) + 4(j+1).$$

Let $\mathcal{D}_{n,j,1}^k = \{g \mathbb{1}_{(0,a_n)} : g \in \mathcal{D}_{n,j}^k\}$ and $\mathcal{D}_{n,j,2}^k = \{g \mathbb{1}_{[a_n,1]} : g \in \mathcal{D}_{n,j}^k\}$. By Lemma B.1.1, we have

$$\begin{aligned} \log \mathcal{N}(j\epsilon_n/4, \mathcal{D}_{n,j,1}^k, d_H) &\lesssim \log(a_n M_n)^{\frac{1}{2k}} [(2j+2 + f_0(0+))\epsilon_n]^{\frac{1}{k}} (j\epsilon_n/4)^{-\frac{1}{k}} \\ &\lesssim n^{1/(2k+1)} (\log n)^{2k/(2k+1)} = n\epsilon_n^2. \end{aligned}$$

By Lemma B.1.1 again, we have

$$\begin{aligned} &\log \mathcal{N}(j\epsilon_n/4, \mathcal{D}_{n,j,2}^k, d_H) \\ &\lesssim \log(f_0(0+) + 4(j+1))^{\frac{1}{2k}} (j\epsilon_n/4)^{-\frac{1}{k}} \\ &\lesssim (\log j)^{\frac{1}{2k}} j^{-\frac{1}{k}} \epsilon_n^{-\frac{1}{k}} \lesssim n\epsilon_n^2 j^2. \end{aligned}$$

Since $\mathcal{D}_{n,j}^k \subseteq \mathcal{D}_{n,j,1}^k + \mathcal{D}_{n,j,2}^k$, then, for j large enough,

$$\log \mathcal{N}(j\epsilon_n/2, \mathcal{D}_{n,j}^k, d_H) \leq \log \mathcal{N}(j\epsilon_n/4, \mathcal{D}_{n,j,1}^k, d_H) + \log \mathcal{N}(j\epsilon_n/4, \mathcal{D}_{n,j,2}^k, d_H) \lesssim n\epsilon_n^2 j^2,$$

which fulfills the condition (4.10).

Next, we control the residual prior probability $\Pi(\mathcal{D}^k \setminus \mathcal{D}_n^k)$. Using the fact that $\psi_k(x, \theta) \leq k/\theta$,

we obtain the estimate

$$\begin{aligned}\Pi(g(0+) > M_n) &\leq \Pi(k \int \theta^{-1} dQ(\theta) > M_n) \\ &= \Pi\left(\int_0^{2k/M_n} \theta^{-1} dQ(\theta) + \int_{2k/M_n}^1 \theta^{-1} dQ(\theta) > M_n/k\right).\end{aligned}$$

Because always $\int_{2k/M_n}^1 \theta^{-1} dQ(\theta) \leq M_n/(2k)$, the residual probability is at most

$$\Pi\left(\int_0^{2k/M_n} \theta^{-1} dQ(\theta) > \frac{M_n}{2k}\right) \leq \frac{2k}{M_n} \mathbb{E} \int_0^{2k/M_n} \theta^{-1} dQ(\theta) \lesssim \int_0^{2k/M_n} \theta^{-1+t_1} d\theta$$

respectively using Markov's inequality and the assumption on the base measure. As the last expression is bounded by a multiple of $M_n^{-t_1}$, it follows that the residual probability is at most $\exp\{-C n \epsilon_n^2\}$, where $C > 0$ can be chosen as large we please for $\epsilon_n = (n/\log n)^{-k/(2k+1)}$ by our choice of M_n . This verifies all required conditions for the applicability of the general theory of posterior contraction rate with $\epsilon_n = (n/\log n)^{-k/(2k+1)}$. \square

Proof of Theorem 4.2.2. The proof is largely similar to that of Theorem 4.2.1 by verifying the three conditions of Theorem 8.9 of Ghosal and van der Vaart (2017) for $\epsilon_n = (n/\log n)^{-k/(2k+1)}$. We highlight the differences in the following.

To estimate the prior concentration in the Kullback-Leibler neighborhood, we need to condition on the event $\{J = J^*\}$ in (4.7), where J^* is such that the uniform approximation using a mixture ψ_k with J^* support points is within $n^{-k/(2k+1)}$. By Lemma 4.1.2, we can assume that $J^* \leq n^{1/(2k+1)}$. To bound the prior probability of $\{\theta_j \in I_j\}$, given $J = J^*$, observe that the condition assumed on H , we have that

$$\Pi(\theta_l \in I_l, 1 \leq l \leq J | J = J^*) \gtrsim \epsilon_n^{4t_2 J^*} \geq \exp\{C_1 J^* \log n\} \geq \exp\{-C_2 n \epsilon_n^2\},$$

where C_1 and C_2 are two positive constants. Along with the estimate $\Pi(J = J^*) = \exp\{-c J^* \log n + \log(n^c - 1)\} \geq \exp\{-C_3 n \epsilon_n^2\}$ for some $C_3 > 0$, the required prior concentration rate is verified.

A sieve is chosen to be $\{g \text{ given by (4.3): } g(0+) \leq M_n\}$, where

$$M_n = \exp\{C n^{1/(2k+1)}(\log n)^{2k/(2k+1)}\}$$

for a large positive constant C . The residual prior probability of the complement of the sieve is bounded by $\sum_{j=1}^{J_n} \Pi(J = j)\Pi(g(0+) \leq M_n | J = j)$. Each term in the sum can be estimated as in the proof of Theorem 4.2.1. It suffices to note that, as in the last theorem, $EQ = H$ because the support points and their weights are independently distributed, these weights sum to one, and the support points are i.i.d. draws from H .

The metric entropy bound for the sieve is obtained by applying Lemma B.1.1, as before. \square

Proof of Theorem 4.2.3. The proof is again obtained by applying the general theory on the posterior contraction rate in Ghosal and van der Vaart (2017) and the verification of the prior concentration condition proceeds in the same way as in the proof of Theorem 4.2.1. However, to derive the stated nearly parametric rate, the estimates of the metric entropy and the residual prior probability will have to be refined.

Suppose that Q_0 is supported on J_0 fixed points $\theta_1^0, \dots, \theta_{J_0}^0$ in $(0, 1)$ with weights $w_1^0, \dots, w_{J_0}^0$. Let c_0 be the minimum of $\{\theta_l^0\}$. Let $I_l = (\theta_l^0 - c_0\epsilon_n^2/2, \theta_l^0 + c_0\epsilon_n^2/2)$, for $l = 1, \dots, J_0$, and $I_0 = (0, 1) \setminus (\cup_{l=1}^{J_0} I_l)$. Clearly, I_l , $1 \leq l \leq J_0$, are pairwise disjoint when n large enough. Then $\|g - g_0\|_1$ is bounded by the expression on the right side of (4.8) with J_0 replacing J^* . Therefore, following the same chain of arguments, the estimate of the prior probability of the Kullback-Leibler neighborhood reduces to

$$\Pi\left(\sum_{j=0}^{k-1} |\beta_j - \beta_{0,j}| \leq \epsilon_n^2\right) \times \Pi(J = J_0) \times \Pi\left(\sum_{l=1}^{J_0} |Q(I_l) - w_l^0| \leq \epsilon_n^2 | J = J_0\right).$$

As before, $-\log \Pi(\sum_{j=0}^{k-1} |\beta_j - \beta_{0,j}| \leq \epsilon_n^2) \lesssim k \log(1/\epsilon_n)$ and

$$-\log \Pi\left(\sum_{l=1}^{J_0} |Q(I_l) - w_l^0| \leq \epsilon_n^2 | J = J_0\right) \lesssim J_0 \log(1/\epsilon_n) \lesssim J_0 \log n,$$

while the prior for J satisfies $-\log \Pi(J = J_0) = c J_0 \log n - \log(n^c - 1) \lesssim J_0 \log n$. Hence the prior concentration condition (4.7) holds for $\epsilon_n = \sqrt{(\max(J_0, k) \log n)/n}$.

Take $\bar{J} = L J_0$ for some $L > 1$ to be determined later. We consider a sieve $\mathcal{L}_n^k = \{g$ given by (4.3): $Q = \sum_{l=1}^J w_l \psi_k(\cdot, \theta_l), (w_l : l \leq J) \in \Delta_J, (\theta_l) \in (n^{-2}, 1)^J, l = 1, \dots, J, J \leq \bar{J}\} = \cup_{J=1}^{\bar{J}} \mathcal{L}_{n, \bar{J}}^k$, say. Then the residual prior is bounded by

$$\Pi(J > \bar{J}) + \Pi(\theta_l < n^{-2}, 1 \leq l \leq J, J \leq \bar{J}). \quad (4.11)$$

The first term in the last display is bounded by

$$\exp\{-c L \max\{k, J_0\} \log n + \log(n^c - 1)\} \leq \exp\{-C \max\{k, J_0\} \log n\},$$

for some $C > 0$. We also observe that

$$\Pi\left(\bigcup_{J \leq \bar{J}} \bigcup_{l \leq J} \{\theta_l < n^{-2}\}\right) = \sum_{j=1}^{\bar{J}} \sum_{l=1}^J \Pi(\theta_l < n^{-2} | J = j) \leq \bar{J}^2 H((0, n^{-2})).$$

Using the inequality

$$\int_0^a e^{-t/x} dx = \frac{a^2}{t} e^{-t/a} - \int_0^a \frac{2}{t} x e^{-t/x} dx \leq \frac{a^2}{t} e^{-t/a}, \quad t > 0, \quad (4.12)$$

the estimate above reduced to a constant multiple of $\bar{J}^2 n^{-4} e^{-t_3 n^2}$. Thus the residual prior probability is bounded by $\exp\{-C \max\{k, J_0\} \log n\}$, where C can be chosen as large as we please by making L large enough.

Next, we estimate the metric entropy of the sieve. For any two arbitrary elements g_1, g_2 of $\mathcal{L}_{n, \bar{J}}^k$, $g_r(x) = \sum_{j=0}^{k-1} \beta_{r,j} \psi_{j+1}(x, 1) + \beta_{r,k} p_r(x)$, where $p_r(x) = \sum_{l=1}^J w_{r,j} \psi_k(x, \theta_l)$, $r = 1, 2$, observe that $\|g_1 - g_2\|_1 \leq \sum_{j=0}^{k-1} |\beta_{1,j} - \beta_{2,j}| + \|p_1 - p_2\|_1$. Using the estimate in Lemma B.1.2, $\|p_1 - p_2\|_1$

can be bounded by

$$\begin{aligned} & \sum_{l=1}^J w_{1,l} \|\psi_k(\cdot, \theta_{1,l}) - \psi_k(\cdot, \theta_{2,l})\|_1 + \sum_{l=1}^J |w_{1,l} - w_{2,l}| \\ & \leq 2n^2 \sum_{l \leq J} |\theta_{1,l} - \theta_{2,l}| + \sum_{l=1}^J |w_{1,l} - w_{2,l}|. \end{aligned} \quad (4.13)$$

Thus if $\sum_{j=0}^{k-1} |\beta_{1,l} - \beta_{2,l}| \leq \epsilon_n^2/2$, $\sum_{l=1}^J |w_{1,l} - w_{2,l}| \leq \epsilon_n^2/4$ and $\max\{|\theta_{1,l} - \theta_{2,l}| : l \leq J\} \leq \epsilon_n^2/(8n^2 J)$, then $d_H(g_1, g_2) \leq \|g_1 - g_2\|_1^{1/2} \leq \epsilon_n$. The $\epsilon_n^2/(4n^2 J)$ covering number of $[0, 1]$ is bounded by $4n^2 J \leq 4n^2 \bar{J}$. The $\epsilon_n^2/2$ -covering number of Δ_k in the ℓ_1 metric and the $\epsilon_n^2/4$ -covering number of Δ_J in the ℓ_1 metric are respectively bounded by $(10/\epsilon_n^2)^{k-1}$ and $(20/\epsilon_n^2)^{J-1} \leq (20/\epsilon_n^2)^{\bar{J}}$ by Proposition C.1 of Ghosal and van der Vaart (2017). Then the ϵ -Hellinger metric entropy of \mathcal{L}_n^k is bounded by

$$\log(\bar{J} \times 4n^2 \bar{J} \times (10/\epsilon_n^2)^{k-1} \times (20/\epsilon_n^2)^{\bar{J}}) \lesssim \bar{J}(\log n + \log(1/\epsilon)).$$

Thus for $\epsilon_n = \sqrt{(\max(J_0, k) \log n)/n}$, the entropy condition (8.5) of Theorem 8.9 of Ghosal and van der Vaart (2017) holds. \square

Proof of Theorem 4.3.1. The first condition follows from the proof of Theorems 4.2.1 and 4.2.2 upon conditioning on $k = k_0$ and using the fact that $-\log \Pi(k = k_0) \lesssim k_0 \log k_0 \leq n\epsilon_n^2$ under (K1) and $-\log \Pi(k = k_0) \lesssim k_0 \log n \lesssim n\epsilon_n^2$ under (K2), where $\epsilon_n = (n/\log n)^{-k_0/(2k_0+1)}$. Then the first condition holds for both the Dirichlet process mixture prior (see the proof of Theorem 4.2.1) and the finite mixture prior (see the proof of Theorem 4.2.2).

To verify the remaining two conditions for posterior contraction rate, we first address the finite mixture prior. For the metric entropy condition, consider the sieve $\mathcal{L}_n = \cup_{k=1}^{k_n} \cup_{j=1}^J \mathcal{L}_{j,k}$, where $\mathcal{L}_{j,k} = \{\sum_{j=0}^{k-1} \beta_j \psi_{j+1}(x, 1) + \beta_k \sum_{l=1}^J w_l \psi_k(\cdot, \theta_l) \in \mathcal{D}^k : (w_l : l \leq J) \in \Delta_J, n^{-2} \leq \theta_l \leq 1, l \leq J\}$. Following the corresponding part in the proof of Theorem 4.2.3, we know that the Hellinger metric entropy of $\mathcal{L}_{J_n, k}$ is bounded by up to a constant multiple of $\max(k, J_n) \log n$. Hence, the

Hellinger entropy of \mathcal{L}_n can be bounded as follows,

$$\log \sum_{k=1}^{k_n} \mathcal{N}(\epsilon_n, \mathcal{L}_{J_n, k}, d_H) \leq \log k_n + \log \mathcal{N}(\epsilon_n, \mathcal{L}_{J_n, k_n}, d_H) \lesssim \max(k_n, J_n) \log n.$$

By choosing J_n the integer part of $L_1(n/\log n)^{1/(2k_0+1)}$ for some $L_1 > 0$, we get $J_n \log n \asymp n\epsilon_n^2$ while maintaining $J_n > J^* \asymp \epsilon_n^{-k_0}$, where J^* is the number of terms used to approximate to derive the estimate in the prior concentration condition. We choose k_n the integer part of $L_2(n/\log n)^{1/(2k_0+1)}$ for some $L_2 > 0$ to fulfil the entropy condition.

Now it remains to bound the residual prior probability of the sieve. Under both (K1) and (K2), the tail estimate of $\Pi(k > k_n) \leq e^{-Ln\epsilon_n^2}$ is obtained with L as large as we please by choosing L_2 sufficiently large. A similar argument applies for the tail $\Pi(J > J_n)$. Now

$$\Pi(\theta_l < n^{-2}, 1 \leq l \leq J, J \leq J_n) \leq \sum_{m=1}^{J_n} \sum_{l=1}^J \Pi(\theta_l \in (0, n^{-2})) \leq J_n^2 n^{-4} e^{-t_3 n^2},$$

where the last inequality is due to (4.12). This expression is also bounded by $e^{-Ln\epsilon_n^2}$ where we can make $L > 0$ as large as we wish. the proof for the finite mixture prior case now follows by an application of the general theory of posterior contraction.

For the Dirichlet process mixture prior, the sieve construction and the residual prior bounding need some modifications. We will elaborate on the differences in the following. Consider the sieve, $\mathcal{E}_n = \cup_{k=1}^{k_n} \mathcal{E}_{k,n}$ where

$$\mathcal{E}_{k,n} = \left\{ \sum_{j=0}^{k-1} \beta_j \psi_{j+1}(x, 1) + \beta_k \sum_{l=1}^{\infty} w_l \psi_k(x, \theta_l) \in \mathcal{D}^k : \right. \\ \left. (w_j : j = 1, 2, \dots) \in \Delta_{\infty}, \sum_{j>J_n} w_j < \epsilon_n^2, \theta_1, \dots, \theta_{J_n} \in (n^{-2}, 1) \right\}.$$

The residual prior probability is bounded in the following,

$$\Pi(\mathcal{E}_n^c) \leq \Pi(k > k_n) + \Pi\left(\sum_{j>J_n} w_j \geq \epsilon_n^2\right) + J_n H((0, n^{-2})).$$

The first and the third terms can be bounded in a similar way as in the previous part. For the second term, by stick-breaking weight representation, $\sum_{l>J_n} w_l = \prod_{l=1}^{J_n} (1 - V_l)$, where $V_l \stackrel{i.i.d.}{\sim} \text{Beta}(1, a)$. Since $-\sum_{l=1}^{J_n} \log(1 - V_l)$ is Gamma distributed with shape parameter J_n and rate parameter A . Then it follows that $\Pi(\sum_{l>J_n} w_l \geq \epsilon_n^2)$ is given by

$$\mathbb{P}\left(-\sum_{l=1}^{J_n} \log(1 - V_l) \leq 2 \log \epsilon_n^{-1}\right) \leq \frac{(2a \log \epsilon_n^{-1})^{J_n}}{(J_n - 1)!} \leq \sqrt{\frac{J_n}{2\pi}} (2eAJ_n^{-1} \log \epsilon_n^{-1})^{J_n}$$

by Stirling's inequality for factorials. Choosing J_n to be the integer part of $L_1(n/\log n)^{1/(2k_0+1)}$ for some L_1 . We can bound the expression by $e^{-Ln\epsilon_n^2}$, where L can be made as large as we like by choosing L_1 large enough.

Following the same argument of the proof of Theorem 4.2.3, we obtain the bound in (4.13) plus $\|\sum_{l \geq J_n} w_l \psi_k(\cdot, \theta_l) - \sum_{l \geq J_n} w'_l \psi_k(\cdot, \theta'_l)\|_1 \leq 2\epsilon_n^2$. Hence, the Hellinger metric entropy of the sieve $\mathcal{E}_{k,n}$ can be bounded by a constant multiple of $J_n \log n$. The proof is concluded by following the same argument used for the finite mixture case. \square

Proof of Theorem 4.4.1. For two density functions g_1, g_2 from model (4.3), we represent them as $g_1(u) = \alpha_1 + (1 - \alpha_1)h_1(u)$ and $g_2(u) = \alpha_2 + (1 - \alpha_2)h_2(u)$, separating out the constant component. We shall bound the $|\alpha_1 - \alpha_2|$ by a constant multiple of the square root of the Hellinger distance between g_1 and g_2 . This will lead to the conclusion in view of Theorems 4.2.1, 4.2.2, and 4.3.1.

For $\alpha_1 > \alpha_2$ and $\alpha_1 \leq g_2(0+)$, the solution s_0 to the equation $g_2(u) = \alpha_1$ exists and is unique due to the strict convexity of g_2 . Then

$$g_2(u) \geq g'_2(s_0)(u - s_0) + \alpha_1 \text{ for every } u \in (0, 1); \quad (4.14)$$

here g'_2 can be considered as either the right or the left derivative, both of which are well-defined for a convex function. As $g_2 \geq \max\{g'_2(s_0)(u - s_0) + \alpha_1, 0\}$ for every $u \in (0, 1)$, and $|g'_2(s_0)| \geq (\alpha_1 - \alpha_2)/(1 - s_0)$, the absolute slope of the line passing through two points on the graph of g_2 ,

(s_0, α_1) and $(1, \alpha_2)$, due to the convexity of g_2 , upon integrating (4.14), it follows that

$$1 \geq \int_0^{s_0} [g_2'(s_0)(u - s_0) + \alpha_1] du \geq \frac{(\alpha_1 - \alpha_2)s_0^2}{2(1 - s_0)} + \alpha_1 s_0 \geq \frac{(\alpha_1 - \alpha_2)s_0}{2(1 - s_0)}.$$

This implies that $s_0 \leq (1 + (\alpha_1 - \alpha_2)/2)^{-1}$, or equivalently, the bound $1 - s_0 \geq (1 + 2/(\alpha_1 - \alpha_2))^{-1} \geq (\alpha_1 - \alpha_2)/3$. Using these estimates

$$\|g_1 - g_2\|_1 \geq \int_{s_0}^1 (g_1(u) - g_2(u)) du \geq \int_{s_0}^1 (\alpha_1 - \frac{\alpha_1 - \alpha_2}{1 - s_0}(s_0 - u) - \alpha_1) du$$

is seen to be bounded below by $(\alpha_1 - \alpha_2)(1 - s_0)/2 \geq (\alpha_1 - \alpha_2)^2/6$. Thus

$$|\alpha_1 - \alpha_2| \leq \sqrt{6 \|g_1 - g_2\|_1} \leq \sqrt{6 d_H(g_1, g_2)}. \quad (4.15)$$

Let $\mathbf{U}_n = \{U_1, \dots, U_n\}$. Hence, for any $M_n \rightarrow \infty$, $\Pi(|\alpha - \alpha_0| > M_n(n/\log n)^{-k/(2(2k+1))} | \mathbf{U}_n) \leq \Pi(d_H(g, g_0) > (M_n^2/6)(n/\log n)^{-k/(2(2k+1))} | \mathbf{U}_n) \rightarrow 0$ in probability under the true distribution. \square

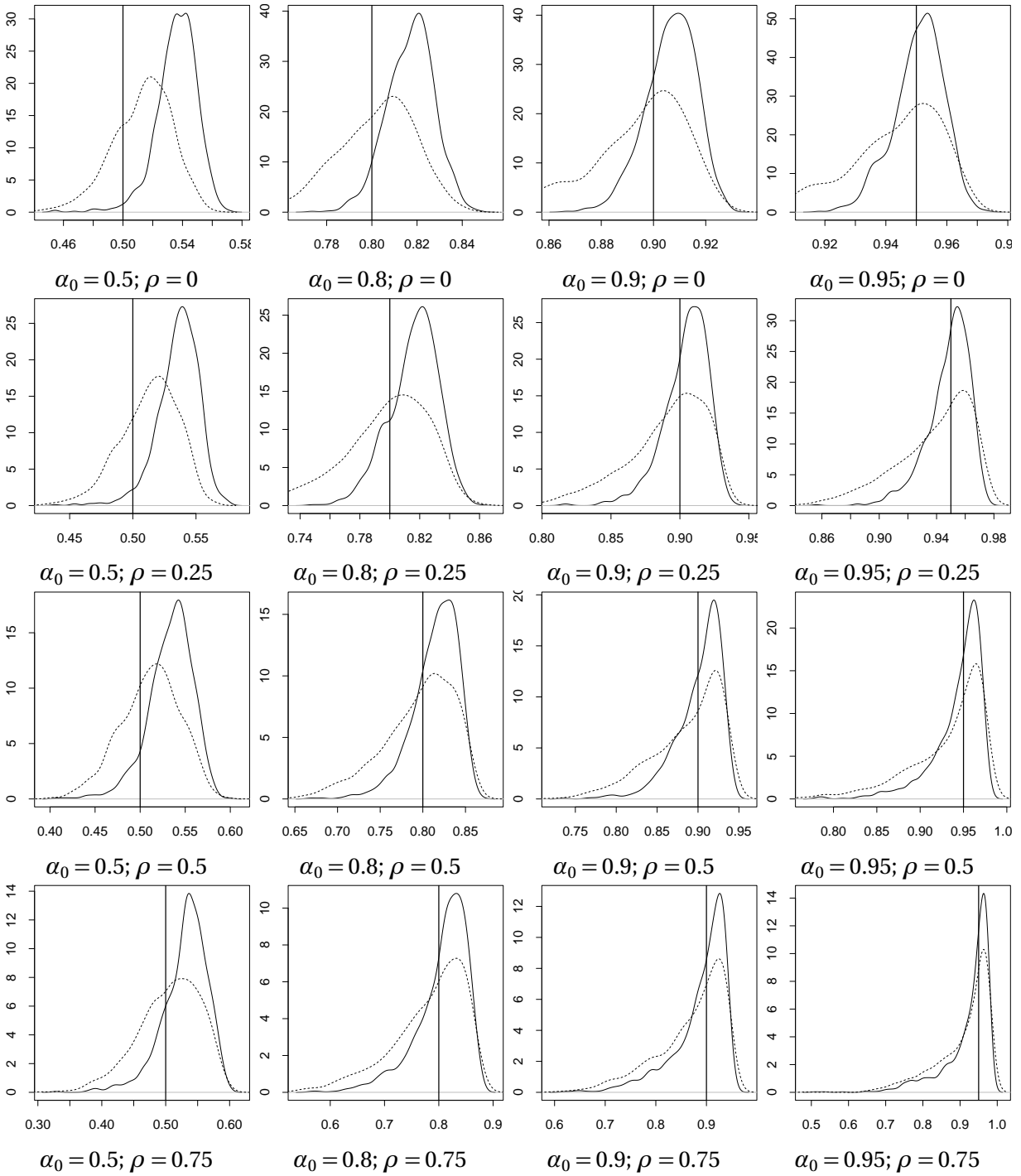


Figure 4.1: Density plots of the estimated α_0 for p -values from two-sided t -tests ($G = 50$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density.

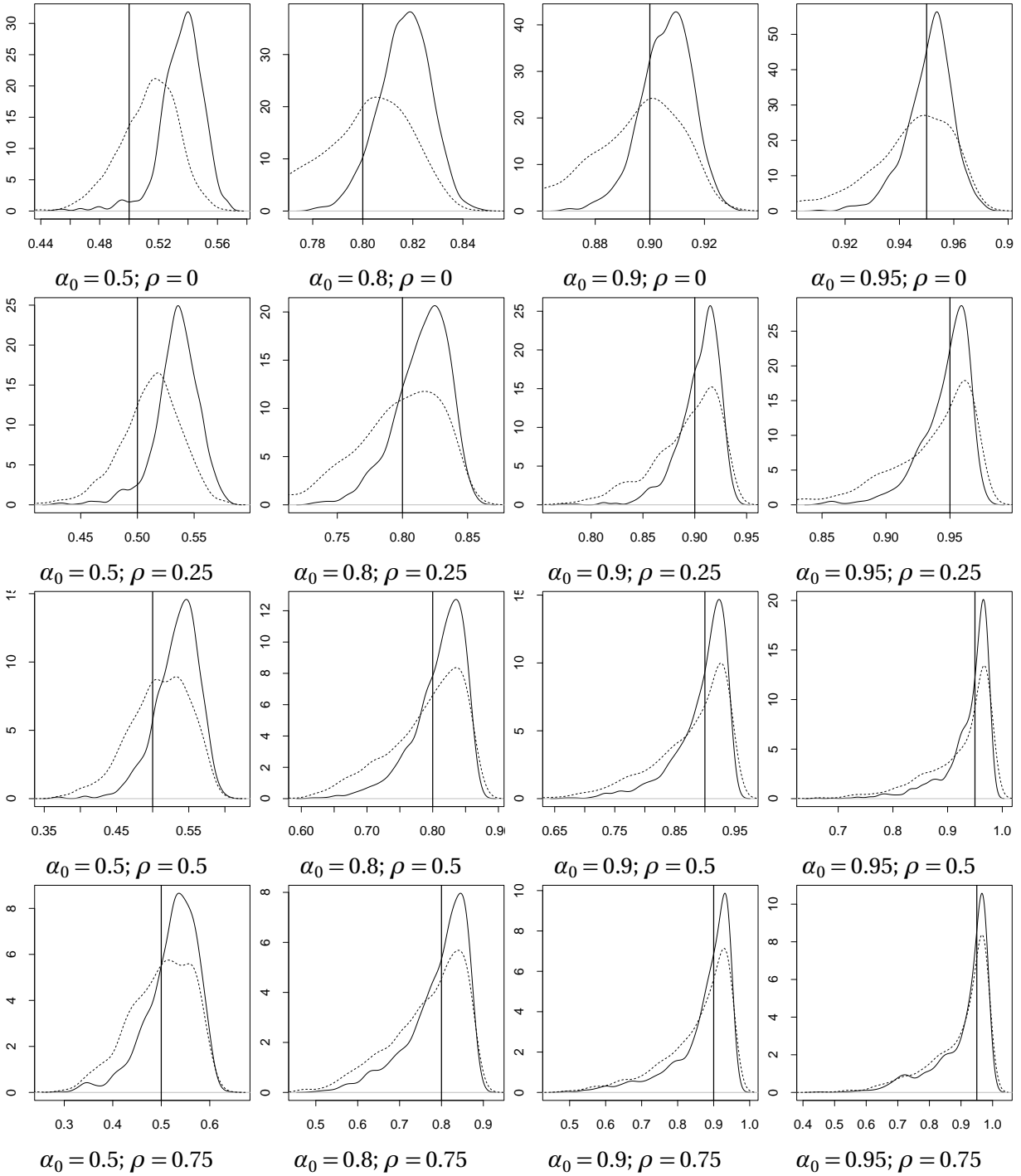


Figure 4.2: Density plots of the estimated α_0 for p -values from two-sided t-tests ($G = 100$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density.

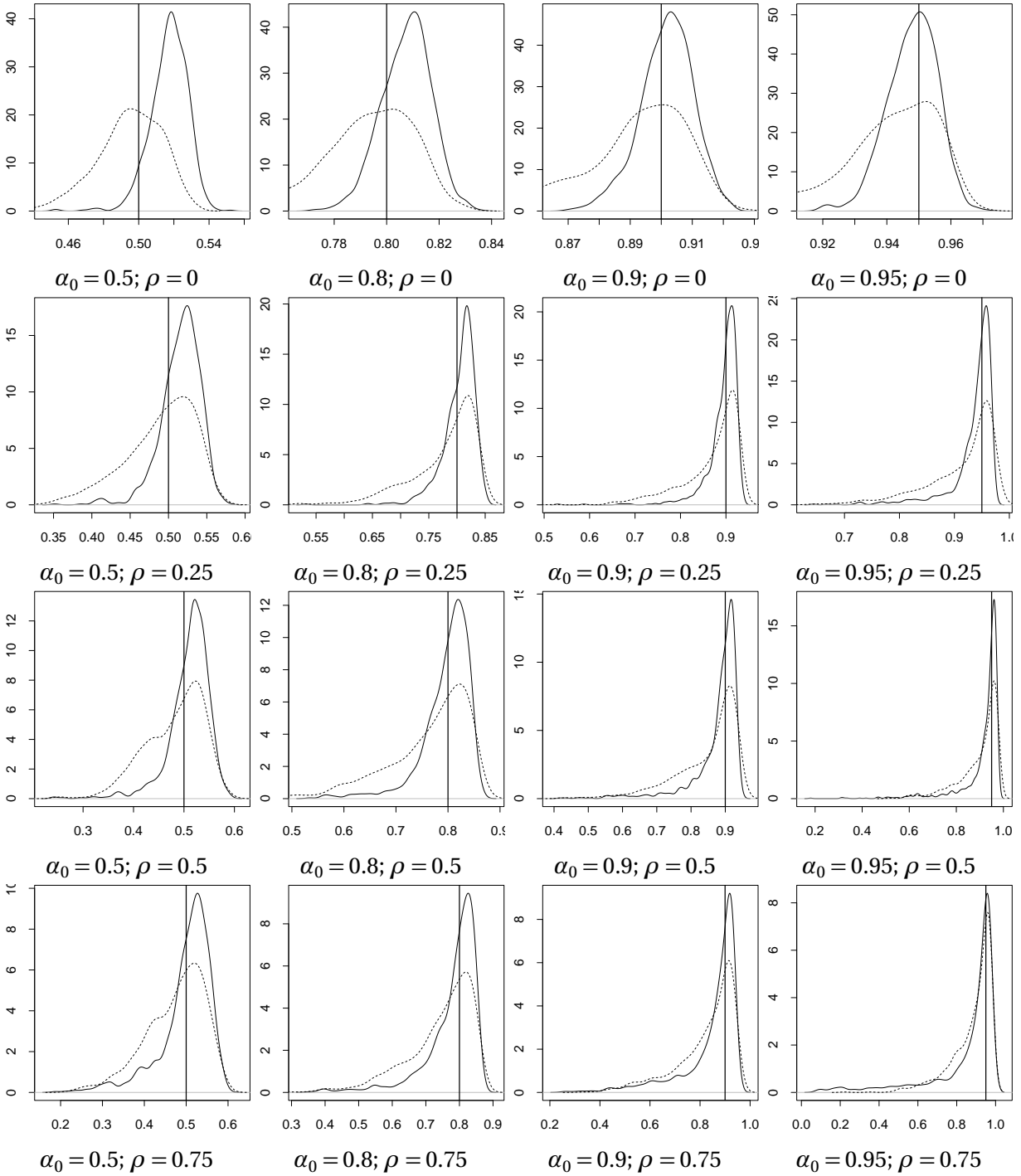


Figure 4.3: Density plots of the estimated α_0 for p -values from one-sided t-tests ($G = 50$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density.

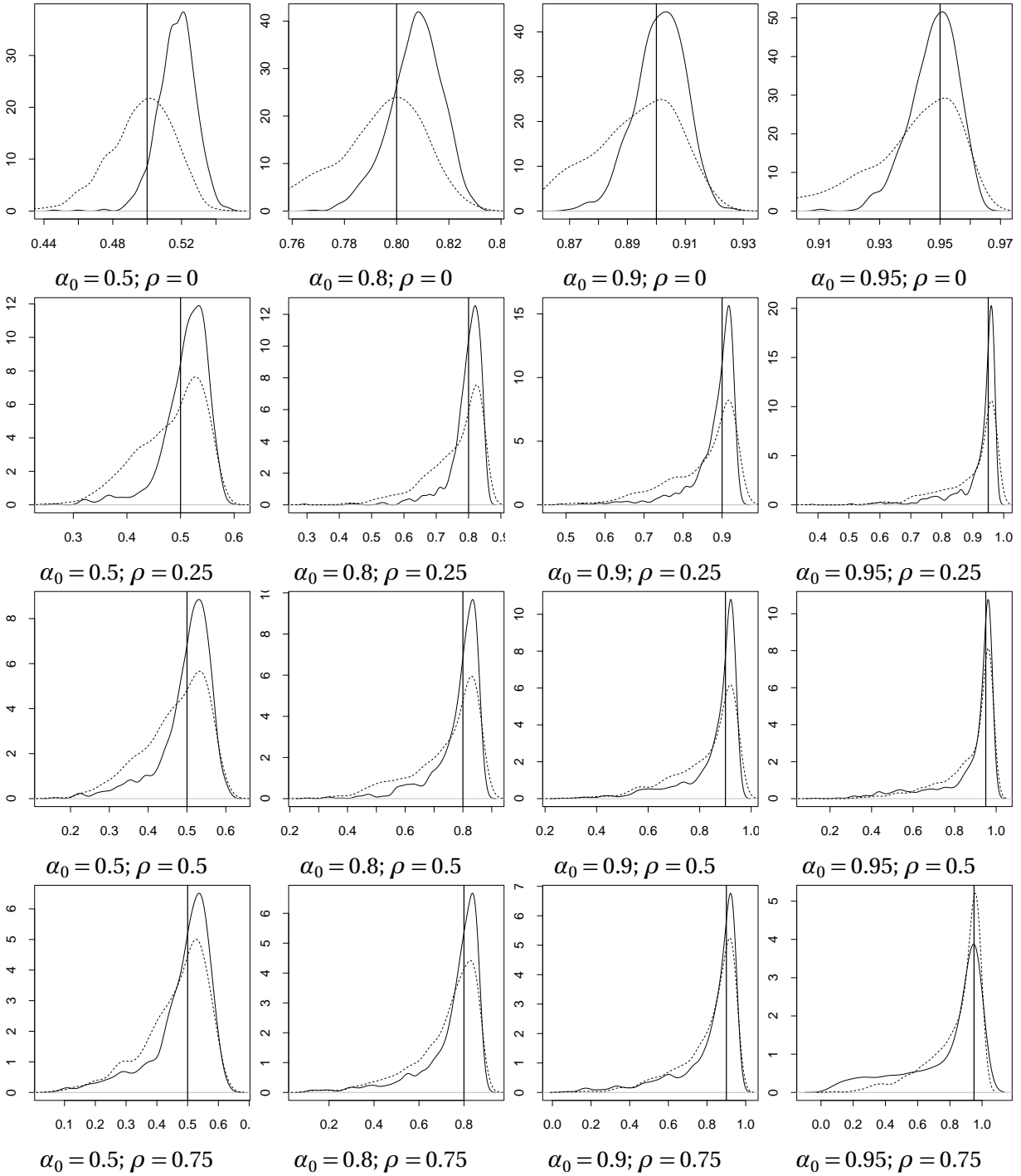


Figure 4.4: Density plots of the estimated α_0 for p -values from one-sided t -tests ($G = 100$). The vertical line indicates the true α_0 , which is also marked below each figure, along with the within-group correlation coefficient. The solid lines are the densities of the posterior mean of α_0 , and the dashed lines are the densities of estimated α_0 by fitting a convex decreasing density.

REFERENCES

- Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26:641–647.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions. The Theory and Application of Isotonic Regression*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, London-New York-Sydney.
- Bayarri, M. J. and Berger, J. O. (2000). p values for composite null models. *J. Amer. Statist. Assoc.*, 95(452):1127–1142, 1157–1170. With comments and a rejoinder by the authors.
- Beal, M., Ghahramani, Z., and Rasmussen, C. (2001). The infinite hidden markov model. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.
- Bean, G. J., Dimarco, E. A., Mercer, L. D., Thayer, L. K., Roy, A., and Ghosal, S. (2013). Finite skew-mixture models for estimation of positive false discovery rates. *Stat. Methodol.*, 10:46–57.
- Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.*, 46(2):745–780.
- Bernardo, J., Berger, J., Dawid, A., Smith, A., et al. (1998). Regression and classification using gaussian process priors. *Bayesian statistics*, 6:475.
- Bhaumik, P. and Ghosal, S. (2015). Bayesian two-step estimation in differential equation models. *Electron. J. Stat.*, 9(2):3124–3154.
- Bhaumik, P. and Ghosal, S. (2017a). Bayesian inference for higher-order ordinary differential equation models. *J. Multivariate Anal.*, 157:103–114.
- Bhaumik, P. and Ghosal, S. (2017b). Efficient Bayesian estimation and uncertainty quantification in ordinary differential equation models. *Bernoulli*, 23(4B):3537–3570.
- Bhaumik, P., Shi, W., and Ghosal, S. (2022). Two-step Bayesian methods for generalized regression driven by partial differential equations. *Bernoulli*, 28(3):1625–1647.
- Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26:607–616.
- Brunk, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, pages 177–197. Cambridge Univ. Press, London.
- Brunner, L. J. and Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.*, 17(4):1550–1566.

- Cai, B. and Dunson, D. B. (2007). Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *J. Amer. Statist. Assoc.*, 102(480):1158–1171.
- Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018.
- Chakraborty, M. and Ghosal, S. (2021a). Bayesian inference on monotone regression quantile: coverage and rate acceleration. *Preprint*.
- Chakraborty, M. and Ghosal, S. (2021b). Convergence rates for Bayesian estimation and testing in monotone regression. *Electron. J. Stat.*, 15(1):3478–3503.
- Chakraborty, M. and Ghosal, S. (2021c). Coverage of credible intervals in nonparametric monotone regression. *Ann. Statist.*, 49(2):1011–1028.
- Chakraborty, M. and Ghosal, S. (2022). Rates and coverage for monotone densities using projection-posterior. *Bernoulli*, 28(2):1093–1119.
- Chatterjee, S., Guntuboyina, A., and Sen, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, 43(4):1774–1800.
- Chatterjee, S., Guntuboyina, A., and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli*, 24(2):1072–1100.
- Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923.
- de Leeuw, J., Hornik, K., and Mair, P. (2009). Isotone optimization in R: Pool-adjacent-violators algorithm (pava) and active set methods. *J. Statist. Software*, 32(5):1–24.
- Deng, H., Han, Q., and Zhang, C.-H. (2021). Confidence intervals for multiple isotonic regression and other monotone models. *Ann. Statist.*, 49(4):2021–2052.
- Deng, H. and Zhang, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs. *Ann. Statist.*, 48(6):3672–3698.
- DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.
- Diaconis, P. W. and Freedman, D. (1998). Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli*, 4(4):411–444.
- Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.*, 29(1):124–152.
- Durot, C. (2007). On the L_p -error of monotonicity constrained estimators. *Ann. Statist.*, 35(3):1080–1104.

- Durot, C., Kulikov, V. N., and Lopuhaä, H. P. (2012). The limit distribution of the L_∞ -error of Grenander-type estimators. *Ann. Statist.*, 40(3):1578–1608.
- Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90(430):577–588.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230.
- Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, pages 287–302. Academic Press, New York.
- Fokianos, K., Leucht, A., and Neumann, M. H. (2020). On integrated L^1 convergence rate of an isotonic regression estimator for multivariate observations. *IEEE Trans. Inform. Theory*, 66(10):6389–6402.
- Gao, F. (2008). Entropy Estimate for k -Monotone Functions via Small Ball Probability of Integrated Brownian Motions. *Electronic Comm. Probab.*, 13(none):121 – 130.
- Gao, F. and Wellner, J. A. (2009). On the rate of convergence of the maximum likelihood estimator of a k -monotone density. *Sci. China Ser. A*, 52(7):1525–1538.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000a). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.
- Ghosal, S., Lember, J., and van der Vaart, A. (2008a). Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89.
- Ghosal, S. and Roy, A. (2009). Bayesian nonparametric approach to multiple testing. In *Perspectives in mathematical sciences. I*, volume 7 of *Stat. Sci. Interdiscip. Res.*, pages 139–164. World Sci. Publ., Hackensack, NJ.
- Ghosal, S. and Roy, A. (2011a). Identifiability of the proportion of null hypotheses in skew-mixture models for the p -value distribution. *Electron. J. Stat.*, 5:329–341.
- Ghosal, S. and Roy, A. (2011b). Predicting false discovery proportion under dependence. *J. Amer. Statist. Assoc.*, 106(495):1208–1218.
- Ghosal, S., Roy, A., and Tang, Y. (2008b). Posterior consistency of Dirichlet mixtures of beta densities in estimating positive false discovery rates. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 of *Inst. Math. Stat. (IMS) Collect.*, pages 105–115. Inst. Math. Statist., Beachwood, OH.
- Ghosal, S., Sen, A., and van der Vaart, A. W. (2000b). Testing monotonicity of regression. *Ann. Statist.*, 28(4):1054–1082.
- Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

- Gijbels, I., Hall, P., Jones, M. C., and Koch, I. (2000). Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, 87(3):663–673.
- Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).
- Groeneboom, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, pages 539–555. Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA.
- Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation Under Shape Constraints*, volume 38 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York. Estimators, algorithms and asymptotics.
- Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel.
- Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, 28(1):20–39.
- Han, Q. (2021). Set structured global empirical risk minimizers are rate optimal in general dimensions. *Ann. Statist.*, 49(5):2642–2671.
- Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.*, 47(5):2440–2471.
- Han, Q. and Zhang, C.-H. (2020). Limit distribution theory for block estimators in multiple isotonic regression. *Ann. Statist.*, 48(6):3251–3282.
- He, X., Ng, P., and Portnoy, S. (1998). Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(3):537–550.
- Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 18(3):1259–1294.
- Huang, J. and Wellner, J. A. (1995). Estimation of a monotone density or monotone hazard under random censoring. *Scand. J. Statist.*, 22(1):3–33.
- Huang, Y. and Zhang, C.-H. (1994). Estimating a monotone density from censored observations. *Ann. Statist.*, 22(3):1256–1274.
- Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Stat. Comput.*, 21(1):93–105.
- Kopotun, K. and Shadrin, A. (2003). On k -monotone approximation by free knot splines. *SIAM J. Math. Anal.*, 34(4):901–924.
- Kulikov, V. N. and Lopuhaä, H. P. (2005). Asymptotic normality of the L_k -error of the Grenander estimator. *Ann. Statist.*, 33(5):2228–2255.

- Langaas, M., Lindqvist, B. H., and Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, 67(4):555–572.
- Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53.
- Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.
- Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B*, 40(2):113–146. With discussion.
- Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101(2):303–317.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.*, 12(1):351–357.
- Luss, R., Rosset, S., and Shahar, M. (2012). Efficient regularized isotonic regression with application to gene-gene interaction search. *Ann. Appl. Stat.*, 6(1):253–283.
- MacEachern, S. N. and Müller, P. (1998). Estimating mixture of dirichlet process models. *J. Comp. Graph. Statist.*, 7(2):223–238.
- Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406.
- Petrone, S. (1999). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.*, 27(1):105–126.
- Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A*, 31:23–36.
- Rivoirard, V. and Rousseau, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.*, 7(2):311–333.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester.
- Robins, J. M., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.*, 95(452):1143–1167, 1171–1172. With comments and a rejoinder by the authors.
- Saarela, O. and Arjas, E. (2011). A method for Bayesian monotonic multiple regression. *Scand. J. Stat.*, 38(3):499–513.
- Salomond, J.-B. (2014a). Adaptive Bayes test for monotonicity. In *The contribution of young researchers to Bayesian statistics*, volume 63 of *Springer Proc. Math. Stat.*, pages 29–33. Springer, Cham.

- Salomond, J.-B. (2014b). Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Stat.*, 8(1):1380–1404.
- Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(1):159–175.
- Smythe, R. T. (1974). Sums of independent random variables on partially ordered sets. *Ann. Probability*, 2:906–917.
- Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, 64(3):479–498.
- Stout, Q. F. (2013). Isotonic regression via partitioning. *Algorithmica*, 66(1):93–112.
- Tang, Y., Ghosal, S., and Roy, A. (2007). Nonparametric Bayesian estimation of positive false discovery rates. *Biometrics*, 63(4):1126–1134, 1312.
- Teh, Y., Jordan, M., Beal, M., and Blei, D. (2004). Sharing clusters among related groups: Hierarchical dirichlet processes. In Saul, L., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press.
- Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York. with Applications to Statistics.
- Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40(3):364–372.
- Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.*, 36(1-3):45–54.
- Wang, X. (2008). Bayesian free-knot monotone cubic spline regression. *J. Comput. Graph. Statist.*, 17(2):373–387.
- Williamson, R. E. (1956). Multiply monotone functions and their Laplace transforms. *Duke Math. J.*, 23:189–207.
- Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.*, 9(2):443–448.
- Yang, Y. and Tokdar, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.*, 43(2):652–674.

Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.*, 44(3):1069–1102.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555.

APPENDICES

APPENDIX

A

AUXILIARY RESULTS FOR MULTIVARIATE MONOTONE FUNCTION ESTIMATION

A.1 Supporting results for Section 2.2 and 2.3

We first define the Hölder smooth function class on $[0, 1]^d$.

Definition A.1.1. *The Hölder space $\mathcal{H}(\alpha, L)$, for $\alpha \in (0, 1]$ and $L \in \mathbb{R}_{\geq 0}$, consists of all functions $f : [0, 1]^d \rightarrow \mathbb{R}$ such that*

$$\sup_{\substack{x \neq y \\ x, y \in [0, 1]^d}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^\alpha} \leq L.$$

For a more general definition, we refer the readers to Definition C.4 of Ghosal and van der

Vaart (2017).

Lemma A.1.1. *If X_1, \dots, X_n are a random sample from a density g on $[0, 1]$, $J \rightarrow \infty$, and $n/J^d \gg \log J$. If g is bounded, then for some constants $C > 0$,*

$$P_0(\max\{N_j : j \in [1 : J]\} \leq C n/J^d) \rightarrow 1.$$

If g is bounded away from zero, then for some constant $C' > 0$, we have

$$P_0(\min\{N_j : j \in [1 : J]\} \geq C' n/J^d) \rightarrow 1.$$

Proof. For every j , $N_j \sim \text{Bin}(n, G(I_j))$. If g is bounded from above by a , then $G(I_j)$ is bounded by a/J^d . Following the same argument of the proof of Lemma A.2 of Chakraborty and Ghosal Chakraborty and Ghosal (2021c), we obtain that, by large deviation probability,

$$P_0(N_j > C n/J^d) \leq 2 \exp\{-C' n/J^d\}.$$

By the condition $n/\log J \gg J^d$, we have

$$P_0(\max N_j > C n/J^d) \leq 2 \exp\{-C'' n/J^d\} \rightarrow 0.$$

The second claim follows from a similar argument. □

Lemma A.1.2. *Let G^* be a probability measure on $[0, 1]^d$ such that $\max\{G^*(I_j) : j \in [1 : J]\} \lesssim J^{-d}$. For a given $f : [0, 1]^d \rightarrow \mathbb{R}$ and J , let $f_j : [0, 1]^d \rightarrow \mathbb{R}$ be defined by $f_j(\mathbf{x}) = \sum_{j \in [1 : J]} \theta_j \mathbb{1}\{\mathbf{x} \in I_j\}$, $\mathbf{x} \in [0, 1]^d$, where θ_j is any value between $f((j-1)/J)$ and $f(j/J)$. Then $\|f - f_j\|_{p, G^*} \lesssim J^{-1/p}$. Moreover, for some appropriate choices of θ_j , $j \in [1 : J]$, we can ensure that $f \in \mathcal{M}$.*

Proof. For θ_j any value between $f((j-1)/J)$ and $f(j/J)$,

$$\begin{aligned} \|f - f_J\|_{1,G^*} &= \sum_j \int_{I_j} |f - \theta_j| dG^* \\ &\leq \sum_j (f(j/J) - f((j-1)/J)) G^*(I_j) \\ &\lesssim J^{-d} \sum_j (f(j/J) - f((j-1)/J)). \end{aligned}$$

To get the upper bound of the summation in the last inequality, we first decompose the index set $[1 : J]$ in the following way. For every $j \in [1 : J]$, Let A_j be the largest possible subset of $[1 : J]$ in the form $\{\dots, j-2 \cdot 1, j-1, j, j+1, j+2 \cdot 1, \dots\}$, which is a chain with respect to the coordinatewise partial order on the index set. Then we count the number of different A_j . Note that A_j can be identified by its minimal element. The minimal element of A_j should satisfy that at least one of its coordinates is 1, otherwise, we can subtract this element by 1 while the smaller element is still in $[1 : J]$, thus should be in A_j , contradicting the fact of the minimal element. The number of different minimal elements is no larger than dJ^{d-1} , by choosing a coordinate equal to 1 among all d coordinates and setting the rest ones free in $\{1, \dots, J\}$. The construction of A_j gives

$$\sum_{l \in A_j} (f(l/J) - f((l-1)/J)) \leq f(\mathbf{1}) - f(\mathbf{0}).$$

Then we have

$$\|f - f_J\|_{1,G^*} \lesssim J^{-d} (dJ^{d-1} (f(\mathbf{1}) - f(\mathbf{0}))) \lesssim J^{-1}.$$

The monotonicity constraint will be maintained by choosing, for $j \in [1, J]$,

$$\theta_j = \frac{\int_{I_j} f(\mathbf{x}) d\mathbf{x}}{G(I_j)},$$

or $\theta_j = f((j-1)/J)$, for instance.

For $p > 1$, note that

$$\|f - f_J\|_{p, G^*}^p \leq (f(\mathbf{1}) - f(\mathbf{0}))^{p-1} \|f - f_J\|_{1, G^*}.$$

Then the conclusion follows. □

Remark 7. For $p > 1$, the \mathbb{L}_p approximation rate in Lemma A.1.2 may not be improved. To see this, consider $f = \sum_{j=1}^d \mathbb{1}\{j : x_j > c_j\}$, where c is a fixed vector with irrational coordinates in $[0, 1]$. Note that c is never on the boundary of any hypercube used for partitioning. Clearly, f is a multivariate monotone function with a discontinuity at any x that shares a coordinate with c . Let j^* be the index such that $c \in I_{j^*}$. and generally for a given J , for $k = 1, \dots, d$,

$$\min\{c_{j_k^*} - (j_k^* - 1)/J, j_k^*/J - c_{j_k^*}\} \gtrsim 1/J.$$

For any hypercube I_j used in the partition such that $j_k = j_k^*$ for some $k = 1, \dots, d$, there is a jump of size at least 1 within I_j . Hence, no matter how θ is chosen, $\int_{I_j} |f - f_J|^p \gtrsim J^{-d}$ for all such hypercubes. The number of hypercubes with this property is of the order J^{d-1} , and hence it follows that $\int |f - f_J|^p \gtrsim J^{-1}$. This shows that the approximation order cannot be improved using only equispaced knot points to form the hypercubes for the piecewise constant approximation.

Remark 8. In view of Lemma A.1.1, if $J^d(\log n)/n \rightarrow 0$, then the empirical distribution satisfies the condition $\max\{G_n(I_j) : j \in [1 : J]\} \lesssim J^{-d}$ in probability, and hence $\|f - f_J\|_{1, G_n} \lesssim J^{-d}$, and the implicit constant of proportionality in \lesssim does not depend on f .

Lemma A.1.3. *Suppose J is deterministic and satisfies $J \rightarrow \infty$ and $J^d/n \rightarrow 0$. For \mathbf{X} either deterministic or random, under Assumptions 1-3, we have*

- (i) $\hat{\sigma}_n^2$ converges in probability to σ_0^2 at the rate of $\max\{n^{-1/2}, J^d/n, J^{-1}\}$.
- (ii) If we endow σ^2 with an Inverse-Gamma prior $IG(\beta_1, \beta_2)$ for some $\beta_1 > 0, \beta_2 > 0$, σ^2 contracts around σ_0^2 as the same rate $\max\{n^{-1/2}, J^d/n, J^{-1}\}$.

Proof. Let $\theta_{0,j} = N_j^{-1} \sum_{i: X_i \in I_j} f_0(\mathbf{X}_i)$. By (2.5),

$$\begin{aligned}
\hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n} \sum_{i=1}^n (f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]})^2 + \frac{1}{n} \sum_{j \in [1:J]} N_j (\theta_{0,j} - \zeta_j)^2 \\
&\quad + \frac{2}{n} \sum_{i=1}^n \varepsilon_i (f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]}) + \frac{2}{n} \sum_{j \in [1:J]} N_j \bar{\varepsilon}|_{I_j} (\theta_{0,j} - \zeta_j) \\
&\quad + \frac{2}{n} \sum_{i=1}^n (f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]}) (\theta_{0, [\mathbf{X}_i]} - \zeta_{[\mathbf{X}_i]}) \\
&\quad - \frac{1}{n} \sum_{j \in [1:J]} \frac{N_j^2 (\theta_{0,j} - \zeta_j)^2 + N_j^2 (\bar{\varepsilon}|_{I_j})^2 + 2N_j^2 \bar{\varepsilon}|_{I_j} (\theta_{0,j} - \zeta_j)}{N_j + \lambda_j^{-2}} \\
&= \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n} \sum_{i=1}^n (f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]})^2 + \frac{2}{n} \sum_{i=1}^n \varepsilon_i (f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]}) \\
&\quad + \frac{2}{n} \sum_{i=1}^n (f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]}) (\theta_{0, [\mathbf{X}_i]} - \zeta_{[\mathbf{X}_i]}) \\
&\quad + \frac{1}{n} \sum_{j \in [1:J]} \frac{\lambda_j^{-2} N_j (\theta_{0,j} - \zeta_j)^2}{N_j + \lambda_j^{-2}} + \frac{2}{n} \sum_{j \in [1:J]} \frac{\lambda_j^{-2} N_j \bar{\varepsilon}|_{I_j} (\theta_{0,j} - \zeta_j)}{N_j + \lambda_j^{-2}} \\
&\quad + \frac{1}{n} \sum_{j \in [1:J]} \frac{N_j^2 (\bar{\varepsilon}|_{I_j})^2}{N_j + \lambda_j^{-2}}.
\end{aligned}$$

Note that λ_j^{-2} , ζ_j and f_0 are all bounded. Then we can bound $|\hat{\sigma}_n^2 - \sigma_0^2|$ up to a constant by

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 - \sigma_0^2 \right| + \frac{1}{n} \sum_{i=1}^n |f_0(\mathbf{X}_i) - \theta_{0, [\mathbf{X}_i]}| + \frac{1}{n} \sum_{j \in [1:J]} (\theta_{0,j} - \zeta_j)^2 \\
&\quad + \frac{1}{n} \left| \sum_{j \in [1:J]} \frac{N_j \bar{\varepsilon}|_{I_j} (\theta_{0,j} - \zeta_j)}{N_j + \lambda_j^{-2}} \right| + \frac{1}{n} \sum_{j \in [1:J]} N_j (\bar{\varepsilon}|_{I_j})^2.
\end{aligned} \tag{A.1}$$

The first term of (A.1) is $O_{\mathbb{P}_0}(n^{-1/2})$. By the monotonicity of f_0 , the second term is bounded by

$$n^{-1} \sum_{j \in [1:J]} N_j (f_0(j/J) - f_0((j-1)/J)).$$

By Remark 8, following the same argument of the proof of Lemma A.1.2, we have the second term is $O_{\mathbb{P}_0}(J^{-1})$ for random \mathbf{X} and $O(J^{-1})$ for deterministic \mathbf{X} under Assumption 1. The third term is bounded by a constant multiple of J^d/n since the hyperparameters ζ_j and $\theta_{0,j}$ are

bounded. Noting that $E[(\bar{\epsilon}|_{I_j})^2|\mathbf{X}] = \sigma_0^2/N_j$, by Markov inequality, we know that the last term is $O_{P_0}(J^d/n)$. For the fourth term, by Cauchy–Schwarz inequality, we have

$$\left| \sum_{j \in [1:J]} \frac{N_j \bar{\epsilon}|_{I_j}}{N_j + \lambda_j^{-2}} (\overline{f_0(\mathbf{X})}|_{I_j} - \zeta_j) \right| \lesssim J^{d/2} \sqrt{\sum_{j \in [1:J]} (\bar{\epsilon}|_{I_j})^2} = O_{P_0}(J^d).$$

Combine all of the results and the first claim follows.

Given the first claim, we can prove the second one by following the same proof of Proposition 4.1 (b) of Yoo and Ghosal Yoo and Ghosal (2016). \square

A.2 Supporting results for Section 2.4

Lemma A.2.1. *Let \mathbb{D}_n , $n \geq 1$, be a set of random observations with distribution P_0^n . Let W_n^* , $W_{n,c}^*$, $c > 0$, $n = 1, 2, \dots$, be random variables and let $\mathcal{L}_n^* = \mathcal{L}(W_n^*|\mathbb{D}_n)$, $\mathcal{L}_{n,c}^* = \mathcal{L}(W_{n,c}^*|\mathbb{D}_n)$ stand for their conditional distributions given \mathbb{D}_n respectively, viewed as random measures on \mathbb{R} . Let W , W_c , $c > 0$, be random variables and H be a random process. Define $\mathcal{L}_c = \mathcal{L}(W_c|H)$ and $\mathcal{L} = \mathcal{L}(W|H)$. Assume that*

- (i) *for every $c > 0$, $\mathcal{L}_{n,c}^* \rightsquigarrow \mathcal{L}_c$;*
- (ii) *$\lim_{c \rightarrow \infty} \limsup_{n \rightarrow \infty} P(W_{n,c}^* \neq W_n^*|\mathbb{D}_n) = 0$ in P_0^n -probability;*
- (iii) *$P(W_c \neq W) \rightarrow 0$ as $c \rightarrow \infty$.*

Then $\mathcal{L}_n^ \rightsquigarrow \mathcal{L}$.*

Proof. Let \mathfrak{M} denote the collection of random probability measure on $(\mathbb{R}, \mathfrak{B})$, where \mathfrak{B} is the Borel σ -algebra. Fix a uniformly continuous function $f : \mathfrak{M} \rightarrow [0, 1]$. For a chosen $\epsilon > 0$, get $0 < \eta < \epsilon$, k and g_1, \dots, g_k uniformly continuous functions from \mathbb{R} to $[0, 1]$ depending on f only, such that $\sum_{j=1}^k \left| \int g_j dQ - \int g_j dQ' \right| < 2k\eta$ implies $|f(Q) - f(Q')| < \epsilon$.

For any n and c , we have

$$|Ef(\mathcal{L}_n^*) - Ef(\mathcal{L})| \leq E|f(\mathcal{L}_n^*) - f(\mathcal{L}_{n,c}^*)| + E|f(\mathcal{L}_{n,c}^*) - Ef(\mathcal{L}_c)| + E|f(\mathcal{L}_c) - f(\mathcal{L})|,$$

so it suffices to bound each term for all sufficiently large n and a suitable $c > 0$.

Using (iii), get $c' > 0$ such that $P(W_c \neq W|H) < \epsilon$ for all $c > c'$ on a set E with $P(E^c) < \epsilon$.

From (ii), get $c^* \geq c'$ and $N^* \geq 1$ such that $P(W_{n,c^*} \neq W_n|\mathbb{D}_n) < \eta$ on a set E_n with $P(E_n^c) < \eta$ for all $n \geq N^*$.

From (i), get $N \geq N^*$ such that $|Ef(\mathcal{L}_{n,c^*}^*) - Ef(\mathcal{L}_{c^*})| < \epsilon$ for all $n \geq N$.

Since $|f| \leq 1$,

$$E|f(\mathcal{L}_n^*) - f(\mathcal{L}_{n,c^*}^*)| \leq \epsilon + P(|f(\mathcal{L}_n^*) - f(\mathcal{L}_{n,c^*}^*)| > \epsilon),$$

so it suffices to control

$$P\left(\sum_{j=1}^k \left| \int g_j d\mathcal{L}_{n,c^*} - \int g_j d\mathcal{L}_n \right| \geq k\eta\right).$$

The j th term

$$E[|g_j(W_{n,c^*}^*) - g_j(W_n^*)||\mathbb{D}_n] \leq \eta + P(|W_{n,c^*}^* - W_n^*| > \eta|\mathbb{D}_n) < 2\eta, \quad j = 1, \dots, k,$$

on the event E_n for all $n \geq N$. Hence $P(|f(\mathcal{L}_n^*) - f(\mathcal{L}_{n,c^*}^*)| > \epsilon) < \epsilon$ on E_n for all $n \geq N$.

By the same argument, $E|f(\mathcal{L}_c^*) - f(\mathcal{L})| \leq \epsilon + P(|f(\mathcal{L}_c) - f(\mathcal{L})| > \epsilon)$, and

$$E[|g(W_{c^*}) - g(W)|] \leq \eta + P(|W_{c^*} - W| > \eta) < 2\eta,$$

assuring that $P(|f(\mathcal{L}_c) - f(\mathcal{L})| > \epsilon) < \epsilon$ on E .

Piecing these together, using the value $c = c^*$, for all $n \geq N$, we obtain that

$$|Ef(\mathcal{L}_n^*) - Ef(\mathcal{L})| \leq 3\epsilon + P(E_n^c) + P(E^c) < 5\epsilon.$$

Since $\epsilon > 0$ is arbitrary, this completes the proof. \square

Lemma A.2.2. For $(u, v) \in [0, c1] \times [0, c1]$ such that $u_k \leq x_{0,k}$, $v_k \leq 1 - x_{0,k}$ for all $s+1 \leq k \leq d$, under the conditions of Theorem 4.1, we have that $\omega_n^2 \sum_{j \in [j(-u); j(v)]} N_j \rightarrow_{P_0} \prod_{k=1}^s (u_k + v_k) D_s(u, v)$.

Proof. Let $\mathbf{x}_s = (x_{0,1}, \dots, x_{0,s}, x_{s+1}, \dots, x_d)$. For $s < d$, let

$$D_s^J(\mathbf{u}, \mathbf{v}) = \int_{\left[\left(\lfloor (x_{0,k} - u_k) J_k \rfloor - 1 \right) / J_k < x_k \leq \left(\lfloor (x_{0,k} + v_k) J_k \rfloor \right) / J_k \right]_{s+1 \leq k \leq d}} \mathbf{g}(\mathbf{x}_s) dx_{s+1} \cdots dx_d,$$

and $D_d^J(\mathbf{u}, \mathbf{v}) = D_d(\mathbf{u}, \mathbf{v}) = \mathbf{g}(\mathbf{x}_0)$. As $0 < a_1 \leq \mathbf{g}(\mathbf{x}) \leq a_2 \leq \infty$ and $J_k \rightarrow \infty$ for every $k = s+1, \dots, d$, we have $|D_s^J(\mathbf{u}, \mathbf{v}) - D_s(\mathbf{u}, \mathbf{v})| \lesssim \max\{J_k^{-1} : s+1 \leq k \leq d\} \rightarrow 0$ as $n \rightarrow \infty$.

By the continuity of $\mathbf{g}(\mathbf{x})$ at \mathbf{x}_0 , and using the facts that $u_k, v_k \leq c$, $r_{n,k} \rightarrow 0$ for $1 \leq k \leq s$, and $J_k \gg r_{n,k}^{-1}$, it follows that

$$\begin{aligned} \mathbb{E}[\omega_n^2 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j] &= n \omega_n^2 \int_{I_{j(-\mathbf{u}):j(\mathbf{v})}} \mathbf{g}(\mathbf{x}) d\mathbf{x} \\ &= n \omega_n^2 \prod_{k=1}^s (u_k r_{n,k} + v_k r_{n,k} + O(J_k^{-1})) (D_s^J(\mathbf{u}, \mathbf{v}) + o(1)) \\ &= n \omega_n^{2 + \sum_{k=1}^s \beta_k^{-1}} \prod_{k=1}^s (u_k + v_k + O((r_{n,k} J_k)^{-1})) (D_s^J(\mathbf{u}, \mathbf{v}) + o(1)) \\ &\rightarrow \prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v}). \end{aligned}$$

Further, as $\sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j$ is binomially distributed,

$$\text{Var}(\omega_n^2 \sum_{j \in [j(-\mathbf{u}):j(\mathbf{v})]} N_j) \leq \omega_n^2 \left(\prod_{k=1}^s (u_k + v_k) D_s(\mathbf{u}, \mathbf{v}) + o(1) \right) \rightarrow 0.$$

Thus conclusion now follows from Chebyshev's inequality. \square

Lemma A.2.3 (Theorem 2.11.9 of van der Vaart and Wellner (1996)). *For each n , let Z_{n1}, \dots, Z_{nm_n} be independent stochastic processes indexed by a totally bounded semi-metric space (\mathcal{F}, ρ) . Let $\|Z_{ni}\|_{\mathcal{F}} = \sup\{|Z_{ni}(f)| : f \in \mathcal{F}\}$, $i = 1, \dots, m_n$, $n \in \mathbb{N}$. Suppose that*

- (i) $\sum_{i=1}^{m_n} \mathbb{E} \|Z_{ni}\|_{\mathcal{F}}^2 \mathbb{1}_{\{\|Z_{ni}\|_{\mathcal{F}} > \eta\}} \rightarrow 0$ for every $\eta > 0$.
- (ii) $\sup\{\sum_{i=1}^{m_n} \mathbb{E}(Z_{ni}(f) - Z_{ni}(g))^2 : \rho(f, g) < \delta_n\} \rightarrow 0$ for every $\delta_n \rightarrow 0$.

(iii) $\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_n)} d\epsilon \rightarrow 0$ for every $\delta_n \rightarrow 0$, where

$$\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_n) = \min\{N : \cup_{j=1}^N \mathcal{F}_{\epsilon_j}^n \supset \mathcal{F}, E \sup_{f, g \in \mathcal{F}_{\epsilon_j}^n} |Z_{ni}(f) - Z_{ni}(g)|^2 \leq \epsilon^2\}.$$

Then the sequence $\sum_{i=1}^{m_n} (Z_{ni} - EZ_{ni})$ is asymptotically tight in $\mathbb{L}_\infty(\mathcal{F})$.

Lemma A.2.4. *Let*

$$E_n = \left\{ a_1 n / \left(2 \prod_{k=1}^d J_k \right) \leq N_j \leq 2 a_2 n / \left(\prod_{k=1}^d J_k \right) \right.$$

for all j }, where a_1 and a_2 are respectively lower and upper bounds of the density g . If

$$n / \left(\prod_{k=1}^d J_k \right) \gg \log(n),$$

then $P_0(E_n) \rightarrow 1$.

Proof. This can be shown with the same lines of argument of the proof of Lemma A.2 of Chakraborty and Ghosal (2021c) by replacing J there with $\prod_{k=1}^d J_k$ and noting that $n / \prod_{k=1}^d J_k \gg \log(n) \gtrsim \log(\prod_{k=1}^d J_k)$. \square

Lemma A.2.5. *Under the conditions of Theorem 4.1, $\hat{\sigma}_n^2$ converges to σ_0^2 in probability at the rate $\max\{n^{-1/2}, n^{-1} \prod_{k=1}^d J_k, \max(J_k^{-1} : 1 \leq k \leq d)\}$.*

Proof. $|\hat{\sigma}_n^2 - \sigma_0^2|$ is bounded by up to a constant multiple of

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 - \sigma_0^2 \right| + \frac{1}{n} \sum_{i=1}^n (f_0(\mathbf{X}_i) - \overline{f_0(\mathbf{X})})_{I_j}^2 \\ & + \frac{1}{n} \sum_{j \in [1:J]} \frac{\lambda_j^{-2} N_j}{N_j + \lambda_j^{-2}} (\overline{f_0(\mathbf{X})})_{I_j} - \zeta_j)^2 \\ & + \left| \frac{1}{n} \sum_{j \in [1:J]} \frac{N_j \bar{\epsilon}_{I_j}}{N_j + \lambda_j^{-2}} (\overline{f_0(\mathbf{X})})_{I_j} - \zeta_j \right| + \frac{1}{n} \sum_{j \in [1:J]} N_j (\bar{\epsilon}_{I_j})^2. \end{aligned} \tag{A.2}$$

The first term of (A.2) is $O_{P_0}(n^{-1/2})$. By the monotonicity of f_0 , the second term is bounded by

$$n^{-1} \sum_{j \in [1:J]} N_j (f_0(j/J) - f_0((j-1)/J))^2.$$

Because $|f_0(\mathbf{x})| \leq \max\{f_0(\mathbf{0}), f_0(\mathbf{1})\}$ for every $\mathbf{x} \in [0, 1]^d$, on the event E_n defined in Lemma A.2.4 with probability tending to 1, the last display is further bounded by a multiple of

$$\prod_{k=1}^d J_k^{-1} \cdot \sum_{j \in [1:J]} f_0(j/J) - f_0((j-1)/J).$$

Note that $[1:J]$ can be partitioned into no more than $\sum_{k=1}^d \prod_{p \neq k} J_p$ subsets $\{A_q\}_q$, where each A_q is the largest possible set such that for every pair of $j_1, j_2 \in A_q$, there exists an integer a such that $j_1 = j_2 + a1$. Thus the expression is bounded by

$$\prod_{k=1}^d J_k^{-1} \sum_{k=1}^d \prod_{p \neq k} J_p |f_0(\mathbf{1}) - f_0(\mathbf{0})| \lesssim \max\{J_k^{-1} : 1 \leq k \leq d\},$$

since $J_k \rightarrow \infty$ for every k . Hence, the second term in (A.2) is $O_{P_0}(\max_k J_k^{-1})$. On the event E_n , the third term is bounded by a constant multiple of $\prod_{k=1}^d J_k/n$ since the hyperparameters, ζ_j and λ_j , and the regression function f_0 are bounded. Noting that $\text{Var}(\bar{\epsilon}|_{I_j} | \mathbf{X}) = \sigma_0^2/N_j$ and using Lemma A.2.4, the fourth term is $O_{P_0}(\prod_{k=1}^d J_k/n)$. The expectation of the last term is $O(\prod_{k=1}^d J_k/n)$. It follows that the last term is $O_{P_0}(\prod_{k=1}^d J_k/n)$. \square

Lemma A.2.6. *Let $\mathcal{J} = \{\mathcal{J}_1 \times \dots \times \mathcal{J}_m\} \subseteq \mathbb{N}^m$ and $\{S_j, j \in \mathcal{J}\}$ be a collection of random variables. Assume that for every $1 \leq k \leq m$ and every $(j_1, \dots, j_{k-1}, j_{k+1}, \dots, j_m)$, $\{S_{(j_1, \dots, j_{k-1}, j, j_{k+1}, \dots, j_m)}, \mathcal{F}_j^{(k)}, j \in \mathcal{J}_k\}$ is a martingale. Then for $p > 1$, we have that*

$$\begin{aligned} \mathbb{P}(\max_{j \in \mathcal{J}} |S_j| > \epsilon) &\leq \frac{(p/(p-1))^{p(m-1)}}{\epsilon^p} \mathbb{E}|S_{(n_1, \dots, n_m)}|^p; \\ \mathbb{E}(\max_{j \in \mathcal{J}} |S_j|^p) &\leq (p/(p-1))^{pm} \mathbb{E}|S_{(n_1, \dots, n_m)}|^p. \end{aligned}$$

Proof. The proof is adapted from Lemma 1 of Smythe (1974). Without loss of generality, we

assume that $\mathcal{J}_k = \{1, 2, \dots, n_k\}$ for every $k \leq m$. Define

$$\begin{aligned}\tau_1 &= \inf\{t_1 : \max\{|S_{(t_1, j_2, \dots, j_m)}| : j_k \in \mathcal{J}_k \text{ for } k \geq 2\} > \epsilon\}, \\ \tau_2 &= \inf\{t_2 : \max\{|S_{(\tau_1, t_2, \dots, j_m)}| : j_k \in \mathcal{J}_k \text{ for } k \geq 3\} > \epsilon\}, \\ &\quad \vdots \\ \tau_m &= \inf\{t_m : \max\{|S_{(\tau_1, \dots, \tau_{m-1}, j_m)}| : j_m \in \mathcal{J}_m\} > \epsilon\},\end{aligned}$$

where $\inf \emptyset = \infty$ by convention. Then $\{\max\{|S_j| : j \in \mathcal{J}\} > \epsilon\} = \cup_{j \in \mathcal{J}} \{\tau = j\}$, where $\tau = (\tau_1, \dots, \tau_m)^\top$.

As $\{|S_{(j, j_2, \dots, j_m)}|^p, \mathcal{F}_j^{(1)}, j \in \mathcal{J}_1\}$ is a nonnegative submartingale, $E(|S_j|^p \mathbb{1}_{\{\tau=j\}})$ can be bounded by

$$E(E(|S_{(n_1, j_2, \dots, j_m)}|^p | \mathcal{F}_{j_1}^{(1)}) \mathbb{1}_{\{\tau=j\}}) = E(|S_{(n_1, j_2, \dots, j_m)}|^p \mathbb{1}_{\{\tau=j\}}),$$

for every $j = (j_1, \dots, j_m) \in \mathcal{J}$. Hence it follows that

$$P(\max_{j \in \mathcal{J}} |S_j| > \epsilon) \leq \epsilon^{-p} \sum_{j \in \mathcal{J}} E(|S_j|^p \mathbb{1}_{\{\tau=j\}}) \leq \epsilon^{-p} \sum_{j \in \mathcal{J}} E(|S_{(n_1, \tau_2, \dots, \tau_m)}|^p \mathbb{1}_{\{\tau=j\}}) \leq \epsilon^{-p} E(|S_{(n_1, \tau_2, \dots, \tau_m)}|^p).$$

It is clear that

$$\{\max\{|S_{(n_1, j, j_3, \dots, j_m)}|^p : j_k \in \mathcal{J}_k, k \geq 3\}, \mathcal{F}_j^{(2)}, j \in \mathcal{J}_2\}$$

is a nonnegative submartingale since $\{|S_{(n_1, j, j_3, \dots, j_m)}|^p, \mathcal{F}_j^{(2)}, j \in \mathcal{J}_2\}$ is a nonnegative submartingale. Hence by Doob's inequality,

$$\begin{aligned}& E(|S_{(n_1, \tau_2, \dots, \tau_m)}|^p) \\ & \leq E(\max_{j \in \mathcal{J}_2} \{\max\{|S_{(n_1, j, j_3, \dots, j_m)}|^p : j_k \in \mathcal{J}_k, k \geq 3\}\}) \\ & \leq \left(\frac{p}{p-1}\right)^p E(\max\{|S_{(n_1, n_2, j_3, \dots, j_m)}|^p : j_k \in \mathcal{J}_k, k \geq 3\}).\end{aligned}$$

Using Doob's inequality repeatedly in the subsequent coordinates, we have

$$\mathbb{E}(|S_{(n_1, \tau_2, \dots, \tau_m)}|^p) \leq \left(\frac{p}{p-1}\right)^{p(m-1)} \mathbb{E}(|S_{(n_1, \dots, n_m)}|^p),$$

which gives the first inequality. The second one follows from the similar argument above. \square

Lemma A.2.7 (Lemma C.7 of Han and Zhang (2020)). *Let $g : [0, c]^d \times [0, c]^d \rightarrow \mathbb{R}$, $g(\mathbf{u}, \mathbf{v}) = \prod_{k=1}^d (u_k + v_k)$. Then*

$$|g(\mathbf{u}, \mathbf{v}) - g(\mathbf{u}', \mathbf{v}')| \leq (2c)^d \sqrt{d} (\|\mathbf{u} - \mathbf{u}'\| + \|\mathbf{v} - \mathbf{v}'\|).$$

Lemma A.2.8 (Lemma C.5 of Han and Zhang (2020)). *Let \mathbb{G}_n be the empirical measure with respect to G . Under Assumption 1 and Assumption 2, for some $c_0 \geq 1$ and $l \in L^*$, we can find a sequence $u_n \downarrow 0$ such that with probability at least $1 - O(n^{-2})$,*

$$\sup_{\substack{c_0^{-1} \mathbf{1} \leq \mathbf{u} \leq c_0 \mathbf{1} \\ \mathbf{v} \geq \mathbf{0}}} \left(\omega_n \prod_k (u_k^{l_k} \vee v_k^{l_k}) \right)^{-1} \cdot \left| \frac{\mathbb{G}_n(\mathbf{X} - \mathbf{x}_0)^l \mathbb{1}_{[x_0 - u \circ r_n, x_0 + v \circ r_n]}}{\mathbb{G}_n \mathbb{1}_{[x_0 - u \circ r_n, x_0 + v \circ r_n]}} - \frac{G_n(\mathbf{X} - \mathbf{x}_0)^l \mathbb{1}_{[x_0 - u \circ r_n, x_0 + v \circ r_n]}}{G_n \mathbb{1}_{[x_0 - u \circ r_n, x_0 + v \circ r_n]}} \right|,$$

is bounded from above by u_n .

Lemma A.2.9 (Lemma C.6 of Han and Zhang (2020), Supplement C). *Let (a, b, γ) be such that $a > 1$, $0 < b < \gamma < b + (a + 1)$. Let $R_{a,b,\gamma}(c)$ be defined as in the proof of Lemma 6.7 and $H_i(\mathbf{u}, \mathbf{v})$ as in Theorem 4.1, for $i = 1, 2$. Then there exists some positive constant C depending on d, a and σ_0 such that for any $c > 1$ and $i = 1, 2$,*

$$\mathbb{E} \left[\sup_{(\mathbf{u}, \mathbf{v}) \in R_{a,b,\gamma}} |H_i(\mathbf{u}, \mathbf{v}) - H_i(\mathbf{u}, \mathbf{v} \mathbb{1}_{[s+1:d-1]})| \right] \leq C \sqrt{c^{as-\gamma-a} \mathbb{1}_{\{s=d\}} \log c}.$$

Lemma A.2.10 (Lemma C.8 of Han and Zhang (2020)). *Let $H_i(\mathbf{u}, \mathbf{v})$ be as in Theorem 4.1, for*

$i = 1, 2$. Then for any $\epsilon > 0$, there exists $\rho_\epsilon > 0$ such that

$$\mathbb{P}\left(\min_{\substack{0 \leq v_k \leq 1, s+1 \leq k \leq d \\ v_d = 0}} \max_{0 \leq u_k \leq 1, 1 \leq k \leq d} H_i(\mathbf{u}, \mathbf{v}) \geq \rho_\epsilon\right) \geq 1 - \epsilon.$$

A.3 Supporting results for Chapter 3

Lemma A.3.1. For any $c > 1$, let $h : [0, c]^d \times [0, c]^d \mapsto \mathbb{R}$ be given by $h(\mathbf{a}, \mathbf{b}) = \prod_{k=1}^d (a_k + b_k)$. Then $|h(\mathbf{a}, \mathbf{b}) - h(\mathbf{a}', \mathbf{b}')| \lesssim \|\mathbf{a} - \mathbf{a}'\| + \|\mathbf{b} - \mathbf{b}'\|$, where the implicit constant multiple depends only on c and d .

Proof. As h is a polynomial function of $2d$ arguments and the order with respect to each argument is one while holding the rest fixed, then, by the first-order Taylor expansion, we have that

$$|h(\mathbf{a}, \mathbf{b}) - h(\mathbf{a}', \mathbf{b}')| \leq \sum_{k=1}^d \frac{\partial h}{\partial a_k}(\mathbf{a}', \mathbf{b}') |a_k - a'_k| + \sum_{k=1}^d \frac{\partial h}{\partial b_k}(\mathbf{a}', \mathbf{b}') |b_k - b'_k|.$$

Note that $\partial h / \partial a_k(\mathbf{a}', \mathbf{b}') = \partial h / \partial b_k(\mathbf{a}', \mathbf{b}') = \prod_{l \neq k} (a'_l + b'_l) \leq (2c)^{d-1}$. Thus $|h(\mathbf{a}, \mathbf{b}) - h(\mathbf{a}', \mathbf{b}')|$ is bounded by a multiple of $\|\mathbf{a} - \mathbf{a}'\|_1 + \|\mathbf{b} - \mathbf{b}'\|_1$. By the Cauchy–Schwarz inequality, the sum of the \mathbb{L}_1 -norms is bounded further by $\sqrt{d}(\|\mathbf{u} - \mathbf{u}'\| + \|\mathbf{v} - \mathbf{v}'\|)$, which gives the desired inequality. \square

Lemma A.3.2 (Lemma C.6 of Han and Zhang (2020), Supplement C). Let $a \in (1, \infty)$, $b \in (0, \gamma)$, and $\gamma \in (b, a + b - 1)$. Let $R_{a,b,\gamma}(c) = \{(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d : u_k \in [0, c], 1 \leq k \leq d; u_d \leq c^{-\gamma}; 0 \leq v_k \leq c^a, 1 \leq k \leq d; 0 \leq v_d \leq c^{-b}\}$. Let $H(\mathbf{u}, \mathbf{v})$ be a centered Gaussian process with covariance

$$\text{Cov}(H(\mathbf{u}, \mathbf{v}), H(\mathbf{u}', \mathbf{v}')) = \prod_{k=1}^d (u_k \wedge u'_k + v_k \wedge v'_k), \quad (\text{A.3})$$

for all $(\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}') \in \mathbb{R}_{\geq 0}^d \times \mathbb{R}_{\geq 0}^d$. Then there exists a constant $C_{a,d} > 0$ such that for any $c > 1$,

$$\mathbb{E}\left(\sup_{(\mathbf{u}, \mathbf{v}) \in R_{a,b,\gamma}(c)} |H(\mathbf{u}, \mathbf{v}) - H(\mathbf{0}, \mathbf{v})|\right) \leq C_{a,d} \sqrt{c^{a(d-1)-\gamma} \log c}.$$

Lemma A.3.3. For the Gaussian process $H(\mathbf{u}, \mathbf{v})$ with covariance kernel (A.3),

$$\max\{H(\mathbf{0}, \mathbf{v}) : 0 \leq v_k \leq 1, 1 \leq k \leq d\} \geq 0 \text{ a.s.}$$

Proof. For $d = 1$, $H(0, v)$ is a Brownian motion on $v \geq 0$. The first statement is simply induced by the reflection principle. When $s = d$, by noting that $H(\mathbf{0}, (1, \dots, 1, v_d))$ is a Brownian motion with respect to $v_d > 0$, it holds that

$$0 \leq \mathbb{P}\left(\max_{\substack{0 \leq v_k \leq 1 \\ 1 \leq k \leq d}} H(\mathbf{0}, \mathbf{v}) \leq 0\right) \leq \mathbb{P}\left(\max_{0 \leq v_d \leq 1} H(\mathbf{0}, (1, \dots, 1, v_d)) \leq 0\right) = 0$$

by the fact that

$$\mathbb{P}\left(\max_{0 \leq v_d \leq 1} H(\mathbf{0}, (1, \dots, 1, v_d)) \leq \rho\right) = 1 - 2\mathbb{P}(H(\mathbf{0}, \mathbf{1}) \geq \rho),$$

and the continuity of normal distribution. □

Lemma A.3.4. If $U \sim \text{Gamma}(\Delta, 1)$, then for every $m \in \mathbb{Z}_{>0}$, $\mathbb{E}(U - \mathbb{E}(U))^{2m} \lesssim \Delta^m$ as $\Delta \rightarrow \infty$.

Proof. Suppose $n_\Delta \leq \Delta < n_\Delta + 1$ for some integer n_Δ . We will assume that $\Delta > n_\Delta$ in this proof. If Δ is an integer, the argument can proceed in a similar and simpler way. Let Z_i denote independent random variables with standard exponential distribution for $i = 1, \dots, n_\Delta$. Let $Z_{n_\Delta+1} \sim \text{Gamma}(\Delta - n_\Delta, 1)$. Then U is equal to $\sum_{i=1}^{n_\Delta+1} Z_i$ in distribution. By Marcinkiewicz–Zygmund inequality, there exists a constant $C_m > 0$ depending only on m such that,

$$\mathbb{E}(U - \mathbb{E}(U))^{2m} = \mathbb{E}\left[\sum_{i=1}^{n_\Delta+1} (Z_i - \mathbb{E}(Z_i))\right]^{2m} \leq C_m \mathbb{E}\left[\sum_{i=1}^{n_\Delta+1} (Z_i - \mathbb{E}(Z_i))^2\right]^m. \quad (\text{A.4})$$

By Jensen's inequality,

$$\left[\frac{1}{n_\Delta + 1} \sum_{i=1}^{n_\Delta+1} (Z_i - \mathbb{E}(Z_i))^2\right]^m \leq \frac{1}{n_\Delta + 1} \sum_{i=1}^{n_\Delta+1} (Z_i - \mathbb{E}(Z_i))^{2m}.$$

The the right-hand side of (A.4) is bounded by $(n_\Delta + 1)^{m-1} \sum_{i=1}^{n_\Delta+1} \mathbb{E}(Z_i - \mathbb{E}(Z_i))^{2m}$. For fixed m , the lemma follows when $\Delta \rightarrow \infty$. □

APPENDIX

B

AUXILIARY RESULTS FOR k -MONOTONE DENSITY ESTIMATION

B.1 Supporting results for Chapter 4

The following lemma is adapted from Theorem 3 of Gao and Wellner (2009), which gives an upper bound of the Hellinger metric entropy of k -monotone functions.

Lemma B.1.1. *Let \mathcal{F} be the set of nonnegative k -monotone functions on an interval $[p, p + A]$ such that $f(p) \leq B$ and $\int f \leq M$ for any $f \in \mathcal{F}$, then*

$$\log \mathcal{N}(2\epsilon, \mathcal{F}, d_H) \leq \log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, d_H) \lesssim |\log AB|^{1/(2k)} M^{1/k} \epsilon^{-1/k}.$$

The following lemma gives a property of the kernel function $\psi_k(\cdot; \theta)$ that will be used in our analysis.

Lemma B.1.2. *For $\psi_k(x, \theta)$ as defined in (4.1), we have*

$$\|\psi_k(\cdot, \theta) - \psi_k(\cdot, \theta')\|_1 \leq 2(1 - \min\{\theta, \theta'\} / \max\{\theta, \theta'\}). \quad (\text{B.1})$$

Proof. Without loss of generality, assume that $0 < \theta < \theta'$. Let $\delta_k(x) = \psi_k(x, \theta) - \psi_k(x, \theta')$. It is easy to see that (B.1) holds for $k = 1$. In fact, the equality in (B.1) holds for $k = 1$ by direct calculation using the fact that $\psi_k(\cdot, \theta)$ and $\psi_k(\cdot, \theta')$ are two densities of uniform distributions on $(0, \theta)$ and $(0, \theta')$ respectively. It is clear that $\delta_k(x) \equiv 0$ on $[\theta', 1)$. If $k \geq 2$, we first claim that there exists a unique solution x_0 to the equation $\delta_k(x) = 0$ for $x \in (0, \theta')$. Since $\delta_k(x) < 0$ for all $x \in [\theta, \theta')$, we restrict x in $(0, \theta)$. Noting that $\delta_k(0) = k(\theta^{-1} - \theta'^{-1}) > 0$ and $\delta_k(\theta) = -k\theta'^{-1}(1 - \theta/\theta')^{k-1} < 0$, by the continuity of δ_k , there exists at least one x such that $\delta_k(x) = 0$. Additionally, by (4.1), for $x \in (0, \theta)$, the equation $\delta_k(x) = 0$ is equivalent to $\{(\theta' - x)/(\theta - x)\}^{k-1} = (\theta'/\theta)^k$. Since the function on the left-hand side of the equation is strictly increasing in $x \in (0, \theta)$, there can be only one solution. By continuity again, $\delta_k(x) > 0$ for $x \in (0, x_0)$ and $\delta_k(x) < 0$ for $x \in (x_0, \theta')$, and hence the \mathbb{L}_1 -distance in (B.1) as

$$2 \int_0^{x_0} \left[\frac{k}{\theta} \left(1 - \frac{x}{\theta}\right)^{k-1} - \frac{k}{\theta'} \left(1 - \frac{x}{\theta'}\right)^{k-1} \right] dx = 2 \left[\left(1 - \frac{x_0}{\theta'}\right)^k - \left(1 - \frac{x_0}{\theta}\right)^k \right].$$

Rewriting the equation $\delta_k(x) = 0$ as $(1 - x/\theta)^k = \frac{\theta - x}{\theta' - x} (1 - x/\theta')^k$, the expression for the \mathbb{L}_1 -distance reduces to $2(1 - x_0/\theta')^{k-1}(1 - \theta/\theta')$. Since $0 < x_0 < \theta < \theta'$, the bound $2(1 - \theta/\theta')$ is immediate. \square

Lemma B.1.3 (Lemma B.2 of Ghosal and van der Vaart (2017)). *For every pair of probability*

densities p and q ,

$$K(p, q) \lesssim d_H^2(p, q)(1 + \log \|p/q\|_\infty) \leq 2d_H^2(p, q)\|p/q\|_\infty,$$

$$V(p, q) \lesssim d_H^2(p, q)(1 + \log \|p/q\|_\infty)^2 \leq 2d_H^2(p, q)\|p/q\|_\infty.$$

Proof of Lemma 4.1.1. For sufficiency, note that f given in (4.2) is continuously differentiable up to the order $k-2$. The derivatives are given by

$$\begin{aligned} (-1)^j f^{(j)}(x) &= \sum_{l=j}^{k-1-j} \alpha_l l(l-1)\cdots(l-j+1)(1-x)^{l-j} \\ &\quad + \int_0^1 \frac{k(k-1)\cdots(k-j)}{t^{j+1}} \left(1 - \frac{x}{t}\right)_+^{k-1-j} d\gamma(t), \end{aligned} \quad (\text{B.2})$$

for $j = 0, 1, \dots, k-2$. It is obvious that the expression in (B.2) are nonnegative and nonincreasing as $\alpha_j \geq 0$. The derivative functions (B.2) are also convex as the summation and integral of convex functions are also convex.

To prove the necessity of the characterization, expand f in a Taylor series at $a \in (0, 1)$ using the fact that $f^{(k-2)}$ is absolutely continuous on $[a, x]$ or $[x, a]$:

$$f(x) = \sum_{j=0}^{k-2} \frac{(x-a)^j}{j!} f^{(j)}(a) + \int_a^x \frac{(x-t)^{k-2}}{(k-2)!} f^{(k-1)}(t) dt,$$

where $f^{(k-1)}$ can be either the right or the left derivative function of the convex or concave function $f^{(k-2)}$, as they are different only on up to countably many points. Note that $f^{(k-1)}$ is a monotone piecewise constant function with bounded variation on $[a, x]$ or $[x, a]$. Applying integration by parts on the remainder once and letting a tend to 1, we obtain

$$\begin{aligned} f(x) &= \sum_{j=0}^{k-1} \frac{(x-1)^j}{j!} f^{(j)}(1-) - \int_x^{1-} \frac{(x-t)^{k-1}}{(k-1)!} df^{(k-1)}(t) \\ &= \sum_{j=0}^{k-1} \frac{(x-1)^j}{j!} f^{(j)}(1-) + \int_{0+}^{1-} \frac{k}{t} \left(1 - \frac{x}{t}\right)_+^{k-1} d\gamma(t), \end{aligned}$$

where $\gamma(t) = \int_{0+}^t (-1)^k u^k d f^{(k-1)}(u)/k!$ for any $t > 0$. Note that γ is nondecreasing as $(-1)^k f^{(k-1)}$ is nondecreasing. Then the characterization for k -monotone functions follows.

The characterization for k -monotone densities follows from proper normalization to a probability density function, which leads to a constraint, $(\beta_j : 0 \leq j \leq k) \in \Delta_{k+1}$. \square

B.2 Discrete approximation

We shall show the following shape-preserving approximation result using free knot splines. Indeed, we consider the shape-preserving free knot spline approximation of the functions in $\check{\mathcal{F}}^k$, which can be transformed into the approximant of \mathcal{F}^k since if $f \in \check{\mathcal{F}}^k \cap \mathbb{L}_p$ and $s \in \mathcal{S}_{N,k}$, then $f \circ \tau \in \mathcal{F}^k \cap \mathbb{L}_p$, $s \circ \tau \in \mathcal{S}_{N,k}$, and $\|f - s\|_p = \|f \circ \tau - s \circ \tau\|_p$ and vice versa.

However, Lemma B.2.2 is not a consequence of Proposition 4.1.1. Since $\check{\mathcal{F}}^k$ is a proper subset of \mathcal{C}^k , for $f \in \check{\mathcal{F}}^k \cap \mathbb{L}_p$, typically $d_p(f, \mathcal{S}_{N,k} \cap \check{\mathcal{F}}^k) \geq d_p(f, \mathcal{S}_{N,k} \cap \mathcal{C}^k)$. Thus we can not directly derive Lemma B.2.2 from Proposition 4.1.1. Fortunately, we can follow the argument of the proof of Proposition 4.1.1 by modifying some supporting lemmas therein, and the k -convex free knot spline approximant, constructed in Kopotun and Shadrin (2003), of a k -convex function is a free knot spline in $\check{\mathcal{F}}^k$ provided the function to be approximated is not only k -convex but in $\check{\mathcal{F}}^k$ as well.

We introduce some notations used in Kopotun and Shadrin (2003) and also used in the following lemmas. For $(a, b) \subset (0, 1)$, set

$$\mathcal{C}_*^k(a, b) = \{f \in \mathcal{C}^k : \max\{|f^{(j)}(a+)|, |f^{(j)}(b-)| : j = 0, 1, \dots, k-1\} < \infty\}$$

$$\check{\mathcal{F}}_*^k(a, b) = \{f \in \check{\mathcal{F}}^k : \max\{|f^{(j)}(b-)| : j = 0, 1, \dots, k-1\} < \infty\}.$$

Note that, if $f \in \check{\mathcal{F}}^k$, $f^{(j)}$ is nonnegative and nondecreasing for every $j = 0, 1, \dots, k-1$. Then $\check{\mathcal{F}}_*^k(a, b) \subset \mathcal{C}_*^k(a, b)$ as $f^{(j)}(a)$ are all bounded up to the order $k-1$. For $f \in \mathcal{C}_*^k(a, b)$, let

$\mathcal{C}^k[f](a, b)$ stand for the set

$$\left\{ g \in \mathcal{C}^k : \begin{array}{l} g^{(j)}(a+) = f^{(j)}(a+), 0 \leq j \leq k-2, g^{(k-1)}(a+) \geq f^{(k-1)}(a+); \\ g^{(j)}(b-) = f^{(j)}(b-), 0 \leq j \leq k-2, g^{(k-1)}(b-) \leq f^{(k-1)}(b-). \end{array} \right\}.$$

In view of the following lemma, we can assume, without loss of generality that f has bounded derivatives up to the order $k-1$.

Lemma B.2.1. *Let $f \in \check{\mathcal{F}}^k \cap \mathbb{L}_p(0, 1)$ for $1 \leq p \leq \infty$. Then for any $\epsilon > 0$, there exists $f_\epsilon \in \check{\mathcal{F}}_*^k(0, 1)$ such that $\|f - f_\epsilon\|_p < \epsilon$.*

Proof of Lemma B.2.1. This proof follows from some modifications from (Kopotun and Shadrin 2003, Lemma 4.4). First, we construct f_ϵ in the following. For $u \in (0, 1)$, denote the Taylor polynomial of f up to the degree $k-1$ at u as $T_u(x) = \sum_{l=0}^{k-1} f^{(l)}(u+) l! (x-u)^l / l!$. For some $\delta \in (0, 1)$, define $f_\epsilon(x)$ to be $f(x)$ if $x \in [0, 1-\delta]$ and $T_{1-\delta}(x)$ if $x \in [1-\delta, 1]$. By the proof of (Kopotun and Shadrin 2003, Lemma 4.4), we know that $\|f - f_\epsilon\|_p \rightarrow 0$ as $\delta \rightarrow 0$. To conclude the proof, it suffices to show that $f_\epsilon \in \check{\mathcal{F}}^k$. By definition, $f^{(j)}((1-\delta)+)$ are all nonnegative for $j = 0, 1, \dots, k-1$. Then on $[1-\delta, 1]$, the derivative values $T_{1-\delta}^{(j)}$ are nonnegative and nondecreasing for $j = 0, 1, \dots, k-1$. As $T_{1-\delta}$ is the Taylor polynomial of degree $k-1$, obviously, $f_\epsilon^{(j)}$ are nonnegative and nondecreasing on $[0, 1]$ up to the order $k-1$. Then we can say that $f_\epsilon^{(j)}$ is nonnegative, nondecreasing, and convex on $[0, 1]$ for every $j = 0, 1, \dots, k-1$, that is, $f_\epsilon \in \check{\mathcal{F}}^k$. \square

Lemma B.2.2. *For any $f \in \check{\mathcal{F}}^k \cap \mathbb{L}_p(0, 1)$, there exists some $s \in \mathcal{S}_{C_k N, k} \cap \check{\mathcal{F}}^k$ such that $\|f - s\|_p \leq C_{k,p} d_p(f, \mathcal{S}_{N,k})$ and $s^{(j)}(0+) = f^{(j)}(0+)$ for $j = 0, 1, \dots, k-2$.*

Proof of Lemma B.2.2. In view of Lemma B.2.1, we can assume that $f \in \check{\mathcal{F}}_*^k(0, 1) \subset \mathcal{C}_*^k(0, 1)$. Following the proof of (Kopotun and Shadrin 2003, Theorem 1), we can construct a spline $s \in \mathcal{S}_{C_k N, k}$ such that $s \in \mathcal{C}^k[f](0, 1)$ and $\|f - s\|_p \leq C_{k,p} d_p(f, \mathcal{S}_{N,k})$.

Next, we will show that $s \in \check{\mathcal{F}}^k$. As $s \in \mathcal{C}^k[f](0, 1)$, $s^{(k-1)}$ is of piecewise constant and nondecreasing due to the convexity of $s^{(k-2)}$ and, moreover, $s^{(k-1)}(0+) \geq f^{(k-1)}(0+) \geq 0$. Then $s^{(k-1)}$ is nonnegative. As $s^{(k-2)}(0+) = f^{(k-2)}(0+) \geq 0$, $s^{(k-2)}$ is nonnegative, nondecreasing and

convex. Noting that $s^{(j)}(0) = f^{(j)}(0) \geq 0$ for $j = 0, 1, \dots, k-3$, by induction, it is easy to see that $s^{(j)}$ is nonnegative, nondecreasing, and convex for every $j = 0, 1, \dots, k-3$ since $s^{(j+1)}$ is nonnegative. Hence, we conclude that $s \in \mathcal{F}^k$. \square

Proof of Lemma 4.1.2. Observe that $g \in \mathbb{L}_\infty(0, 1)$ as $|g^{(k-1)}(0+)| < \infty$. Note that \mathcal{D}^k is a subclass of \mathcal{F}^k . By Lemma B.2.2, for any $g \in D^k$, there exists a $\tilde{g} \in \mathcal{S}_{N,k} \cap \mathcal{F}^k$ such that $\|g - \tilde{g}\|_p \leq C d_p(g, \mathcal{S}_{N,k})$ and $\tilde{g}(1-) = g(1-) = 0$ for $j = 0, 1, \dots, k-2$. By Lemma 4.1.1, $\tilde{g}(x) = \alpha_{k-1}(1-x)^{(k-1)}/(k-1)! + \int_0^1 k t^{-1}(1-x/t)_+^{k-1} d\gamma(t)$ for some nonnegative α_{k-1} and some nondecreasing function γ on $(0, 1)$. The first polynomial term can be incorporated into the integral by defining $\gamma(1) = \gamma(1-) + \alpha_{k-1}/k!$. Since \tilde{g} is a piecewise polynomial of degree $k-1$, we conclude that γ is piecewise constant function with jumps at the knots of the spline. Let $g_N = \tilde{g} / \int \tilde{g}$ satisfying the structure requirement. Now g_N maintains the desired approximation rate:

$$\|g - g_N\|_\infty \leq \|g - \tilde{g}\|_\infty + \left\| \tilde{g} - \frac{\tilde{g}}{\int \tilde{g}} \right\|_\infty \leq \frac{1 + \|g\|_\infty}{1 - \|g - \tilde{g}\|_\infty} \|g - \tilde{g}\|_\infty. \quad (\text{B.3})$$

By (DeVore and Lorentz 1993, Theorem 12.4.5), $\|g - \tilde{g}\|_\infty \leq C_{k,g} N^{-k}$, provided $|g^{(k-1)}(0+)| < \infty$. Thus the right-hand side of (B.3) can be further bounded by $C'_{k,g} N^{-k}$. \square