Electronic Journal of Statistics
Vol. 17 (2023) 798–822
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2115

Posterior contraction and testing for multivariate isotonic regression*

Kang Wang

Department of Statistics,
North Carolina State University,
Raleigh, NC 27695, U.S.A.
e-mail: kwang22@ncsu.edu

and

Subhashis Ghosal

Department of Statistics,
North Carolina State University,
Raleigh, NC 27695, U.S.A.
e-mail: sghosal@ncsu.edu

Abstract: We consider the nonparametric regression problem with multiple predictors and an additive error, where the regression function is assumed to be coordinatewise nondecreasing. We propose a Bayesian approach to make inferences on the multivariate monotone regression function, obtain the posterior contraction rate, and construct a universally consistent Bayesian testing procedure for multivariate monotonicity. To facilitate posterior analysis, we temporarily set aside the shape restrictions, and endow a prior on blockwise constant regression functions with independently normally distributed heights. The unknown variance of the error term is either estimated by the marginal maximum likelihood estimate, or is equipped with an inverse-gamma prior. Then the unrestricted block-heights are a posteriori also independently normally distributed given the error variance, by conjugacy. To comply with the shape restrictions, we project samples from the unrestricted posterior onto the class of multivariate monotone functions, inducing the "projection-posterior distribution", to be used for making an inference. 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 $n^{-1/(2+d)}$. Then 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, that is, the level of the Bayesian test goes to zero, and the power at any fixed alternative goes to one. Moreover, we show that for a smooth alternative function, power goes to one as long as its distance to the class of multivariate monotone functions is at least of the order of the estimation error for a smooth function. To the best of our knowledge, no other test for multivariate monotonicity is available in the Bayesian or the frequentist literature.

MSC2020 subject classifications: 60G10, 62G20, 62F15.
Keywords and phrases: Multivariate isotonic regression, contraction rate, Bayesian tests for multivariate monotonicity.

Received June 2022.

*Research is partially supported by NSF grant number DMS-1916419.

1. Introduction

Shape restricted inference is an important nonparametric statistical technique with a long history. Functions with qualitative shape restrictions, such as monotonic functional relationships between variables, are quite common in natural sciences, sociology, economics and many other areas. Shape restrictions on the function space can also serve as a relaxation to restricted parametric models, such as log-concave density estimation. The shape constraints themselves yield function estimators with good statistical properties without resorting to the subjective selection of the smoothness level, such as in kernel or spline smoothing. Starting from early works on statistical inference under order restrictions, problems with monotonicity constraints on parameters of interest, regression functions, and probability densities were extensively studied. For the univariate monotone function estimation problem, the least squares estimator for an isotonic regression function and the maximum likelihood estimator for a decreasing density function have interesting geometrical representations respectively as the slope of the greatest convex minorant and the least concave majorant of a cumulative sum diagram. The limit distribution at an interior point on which the function has a positive derivative is well known as the rescaled Chernoff's distribution; see Grenander [18], Prakasa Rao [30], Brunk [4], Groeneboom [19, 20], Barlow et al. [2], and Robertson et al. [31]. Asymptotic global behaviors of the least squares estimators under monotone constraints are well developed with respect to various metrics; see Groeneboom [19], Kukilov and Lopuhaä [25], Durot [12], and Durot et al. [13]. Zhang [44] and Bellec [3] studied the non-asymptotic risk bounds of the least squares estimators. Testing for monotonicity was studied in the univariate case by Akakpo et al. [1], Hall and Heckman [21] and Ghosal et al. [16]. Applications of shape-restricted inference in various areas, like causal inference, genetics, and material science, are still of growing interest; more details can be found in Westling et al. [42], Luss and Rosset [27] and Vittorietti et al. [40].

Compared to the well-studied case of univariate monotone shape-restricted inference, convergence results for multivariate monotonicity were lacking until recent years. Among different possible multivariate monotonicity restrictions, coordinatewise monotonicity is popularly considered. This naturally arises in some modeling contexts studied in Robertson et al. [31], Saarlera and Arjas [32] and Fokianos et al. [15]. In the frequentist literature, the least squares estimator based on independent and identically distributed (i.i.d.) under the multivariate coordinatewise monotonicity constraint has received the most attention. For both a fixed grid design or a random design, the minimax rate is given by $n^{-\min\{2/(d+2),1/d\}}$ with respect to the squared empirical $\mathbb{L}_2$-metric when the true regression function is coordinatewise nondecreasing and is of bounded variation (Chatterjee et al. [8], Han et al. [23]). Han [22] showed that some special global empirical risk minimizers, such as the least squares estimator in multivariate isotonic regression, are rate optimal even when the entropy integral concerned therein diverges rapidly. For $d \geq 2$ and $p \geq 1$, the minimax risk under the general empirical $\mathbb{L}_p$-loss $n^{-1}\sum_{i=1}^{n} \mathrm{E}|\hat{f}(x_i) - f(x_i)|^p$ of the estimator $\hat{f}$

of a function f for deterministic predictors on a grid, and the integrated $\mathbb{L}_p$-risk $\int \mathrm{E}|\hat{f}(x) - f(x)|^p dG(x)$ for a random predictor $X \sim G$, are bounded below by a multiple of $n^{-\min\{1/d, p/(d+2)\}}$ under some conditions on the signal-to-noise ratio and the error term; see Deng and Zhang [11]. In addition to the least squares estimator or the empirical risk minimization estimators, other estimators, such as a block-estimator modifying the min-max formula for the isotonic least squares solution (Robertson et al. [31]), have been proposed and studied; see Fokianos et al. [15], Deng and Zhang [11] and Han and Zhang [24]. The computation of isotonic regression with respect to a general partial ordering minimizing the $\mathbb{L}_q$-metric also attracted attention. One solution is to put this question under the framework of convex optimization with linear constraints; see, for example, Dykstra and Robertson [14], de Leeuw [10] and Meyer [28]. A sequential partitioning algorithm is designed for isotonic regression under the weighted $\mathbb{L}_1$-metric that computes in $O(n \log n)$ time for the coordinatewise isotonic regression with 2-dimensional grid designs and in $O(n^2 \log n)$ time for the $d \geq 3$ case; see Stout [37] for details. Stout [38] gave another algorithm with better computation time under the $\mathbb{L}_1$-metric for the unweighted data. In terms of the $\mathbb{L}_2$-metric, which leads to the usual isotonic least squares estimator, the algorithm in Spouge et al. [36] can compute in $O(n^2)$ time for a two-dimensional grid data.

Bayesian approaches to isotonic regression are also available in the literature. Most of these approaches deal with a univariate isotonic regression function. Neelon and Dunson [29] modeled the regression function as a piecewise linear function and incorporated the monotonicity constraints in the priors of the sequential slopes. Shivley et al. [35] considered Bayesian regression splines under the monotonicity constraint, which is incorporated into the spline coefficients through a mixture of a constrained normal distribution and a probability distribution on the boundary of the constrained parameter set. Lin and Dunson [26] considered a Gaussian process prior, and projected posterior samples on the space of monotone functions to obtain an induced posterior distribution, which is subsequently used to make inferences. Chakraborty and Ghosal [5, 6, 7] used the same idea with a piecewise constant prior and obtained results on posterior contraction and frequentist coverage of Bayesian credible intervals. For the multivariate monotone regression, Saarela and Arjas [32] used marked point processes to construct piecewise constant sample paths for the function. They considered a homogeneous Poisson process prior on the random point positions and endowed the associated marks, i.e., the function value at the point, with the uniform distribution prior supported on the allowed interval restricted by the shape constraints. Chipman et al. [9] applied a constrained sum-of-trees to model monotone regression functions. To obtain posterior samples, Markov chain Monte Carlo (MCMC) methods are used for each method mentioned above. The Bayesian testing procedure for monotonicity in the univariate case has also been proposed by a few authors. Salomond [33] and Chakraborty and Ghosal [5, 7] developed tests based on the posterior distribution of a discrepancy of the function from the unrestricted posterior with its monotone projection.

Scott et al. [34] used smoothing splines and regression splines to model the regression function, and incorporated the monotonicity constraints into the prior for the coefficients. To test for monotonicity, they considered the Bayes factor and converted the monotonicity hypothesis to a condition on the minimum of the derivative functions. To the best of our knowledge, no test for multivariate monotonicity, Bayesian or frequentist, is yet available in the literature.

In this paper, we consider a Bayesian approach to multivariate monotone regression, using the projection technique. We show that the resulting induced posterior supported on block-wise constant multivariate monotone function contracts at the optimal rate with respect to an $\mathbb{L}_1$-metric. The basis of the result is a new $\mathbb{L}_1$-approximation property for multivariate monotone functions by piecewise constant functions. We then construct a test for multivariate monotonicity based on the posterior probability of a slight enlargement of the set of multivariate monotone functions. We show that the resulting Bayesian test is universally consistent in that the size of the test goes to zero, and the power goes to one at any fixed alternative, as the sample size increases to infinity. We further show that, even for alternatives approaching the null region, the power can go to one, provided that the alternative maintains a distance of at least a sufficiently large multiple of the posterior contraction rate determined by its smoothness. These results generalize the testing results of Chakraborty and Ghosal [5] to multidimensional predictors.

The rest of this paper is organized as follows. In Section 2, we describe the prior distribution and the projection-posterior approach. Posterior contraction rates and the properties of the Bayesian test for monotonicity are presented in Section 3. Simulation studies to judge the qualities of the proposed estimation and testing procedure in finite sample sizes are conducted in Section 4. Proofs are deferred to Section 5. Certain auxiliary results and their proofs are presented in the appendix.

2. Setup, prior and posterior

We shall use the following notations and symbols throughout the paper. The notation $\mathbb{R}$ stands for the real line, $\mathbb{Z}$ for the set of integers, and $\mathbb{Z}_>$ for the set of positive integers. Vectors and matrices will be denoted by bold letters, and the default form of a vector is assumed to be in the column form. For $\boldsymbol{a} \in \mathbb{R}^d$, let a_k denote the k-th coordinate, $k = 1, \ldots, d$. The symbols $\mathbf{1}$ and $\mathbf{0}$ will respectively denote the d-dimensional all-one and all-zero vectors. For a real x, $\lfloor x \rfloor$ (respectively, $\lceil x \rceil$) will stand for the greatest integer less (respectively, the smallest integer greater) than or equal to x. The indicator function of a set A is denoted by $\mathbb{1}_A(\cdot)$. 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 pth 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 $\rho(f, \mathcal{F}) = \inf\{\rho(f, g) : g \in \mathcal{F}\}$. The symbol $\lesssim$ will stand for an inequality up to a constant multiple, and $\asymp$ will stand for equality in order. For two positive real sequences, a_n and b_n, we

also say $a_n \gg b_n$ if $b_n = o(a_n)$. Let $\mathrm{N}(\nu, \sigma^2)$ stand for the normal distribution with mean ν and variance σ^2.

Consider the natural partial ordering $\preceq$ on $\mathbb{R}^d$ given by: $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$ if $x_{1,k} \leq x_{2,k}$ for every $1 \leq k \leq d$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^d \times \mathbb{R}^d$ and we also use $\boldsymbol{x}_2 \succeq \boldsymbol{x}_1$ if $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$. For $(\boldsymbol{j}_1, \boldsymbol{j}_2) \in \mathbb{Z}^d \times \mathbb{Z}^d$ and $\boldsymbol{j}_1 \preceq \boldsymbol{j}_2$, let $[\boldsymbol{j}_1 : \boldsymbol{j}_2] = \{\boldsymbol{j} \in \mathbb{Z}^d : \boldsymbol{j}_1 \preceq \boldsymbol{j} \preceq \boldsymbol{j}_2\}$.

Definition 1. *A function $f : I \to \mathbb{R}$, where $I \subset \mathbb{R}^d$, is called multivariate monotone if $f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_2)$ whenever $\boldsymbol{x}_1 \preceq \boldsymbol{x}_2$.*

The space of all multivariate monotonic functions on $[0,1]^d$ will be denoted by $\mathcal{M}$.

We consider the nonparametric multivariate regression model

$$Y = f(\boldsymbol{X}) + \varepsilon, \tag{2.1}$$

where $\boldsymbol{X}$ is the d-dimensional predictor and ε is an error term with zero mean and finite variance, independent of $\boldsymbol{X}$. We shall assume, essentially without loss of generality, that the domain of $\boldsymbol{X}$ is $[0,1]^d$. Instead of a traditional smoothness assumption on the regression function f, we assume that f is multivariate monotonic.

We observe the data $\mathbb{D}_n$ consisting of n samples $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ independently from the model. The predictor variable $\boldsymbol{X}$ 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. The main objective of this paper is to study the contraction rate of the posterior distribution and construct a Bayesian test for multivariate monotonicity with some desirable large sample frequentist properties. To facilitate Bayesian inference, we construct a likelihood based on the working model assumption that $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$, although the actual distribution may be non-normal. For a given f, let $p_{f,\sigma}(y|x) = (\sqrt{2\pi}\sigma)^{-1} \exp[-(y - f(x))^2/(2\sigma^2)]$ stand for the conditional density of Y given $X = x$.

Let $G_n = n^{-1} \sum_{i=1}^n \delta_{\boldsymbol{X}_i}$ denote the empirical distribution of $\boldsymbol{X}$. For a deterministic predictor variable $\boldsymbol{X}$, this is a sequence of deterministic distributions, while for a random 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 $(\boldsymbol{X}, Y)$. The expectation with respect to P_0 will be denoted by E_0.

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 are 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 [26] and Chakraborty and Ghosal [5, 7]. A generalization of this approach using a broader "immersion map" was used by Wang and Ghosal [41] to study the coverage of a Bayesian credible interval of a multivariate monotone regression function at a given point.

To obtain posterior contraction rate in terms of a global metric like an $\mathbb{L}_1$-distance, we follow the projection-posterior approach, as in the univariate case of Chakraborty and Ghosal [5]. Given $J \in \mathbb{Z}_{>0}$, let $I_{\mathbf{1}} = [0, J^{-1}]^d$ and $I_{\boldsymbol{j}} = \prod_{k=1}^d ((j_k - 1)/J, j_k/J]$ for $\boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}] \backslash \{\mathbf{1}\}$. Let $\mathcal{F}_J = \{f : f = \sum_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} \theta_{\boldsymbol{j}} \mathbb{1}_{I_{\boldsymbol{j}}}, \theta_{\boldsymbol{j}} \in \mathbb{R}\}$, the set of piecewise constant functions. If f were unrestricted, a conjugate prior for the model (2.1) is given by letting

$$\theta_{\boldsymbol{j}} \overset{\text{ind}}{\sim} \mathrm{N}(\zeta_{\boldsymbol{j}}, \sigma^2 \lambda_{\boldsymbol{j}}^2), \qquad \boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}], \tag{2.2}$$

where $\zeta_{\boldsymbol{j}}, \lambda_{\boldsymbol{j}}$ are hyperparameters, and then either by choosing J deterministically (increasing with n) or by putting a prior on J. The prior and the resulting posterior are both supported within $\mathcal{F}_J$, and the posterior is given by

$$\theta_{\boldsymbol{j}} | (\mathbb{D}_n, \sigma^2, J) \overset{\text{ind}}{\sim} \mathrm{N}((N_{\boldsymbol{j}} \bar{Y}|_{I_{\boldsymbol{j}}} + \zeta_{\boldsymbol{j}} \lambda_{\boldsymbol{j}}^{-2})/(N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}), \sigma^2/(N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2})), \tag{2.3}$$

where $N_{\boldsymbol{j}} = \sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in I_{\boldsymbol{j}}\}$, the number of observed points falling in the $\boldsymbol{j}$th block, and $\bar{Y}|_{I_{\boldsymbol{j}}} = \sum_{i=1}^n Y_i \mathbb{1}\{\boldsymbol{X}_i \in I_{\boldsymbol{j}}\}/N_{\boldsymbol{j}}$, $\boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]$. The resulting posterior for f will be referred to as the "unrestricted posterior", which is not supported within $\mathcal{M}$. The projection map then produces an induced distribution suitable for an inference, to be referred to as the "projection-posterior" distribution.

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 [17]). 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.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 [5] 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 assumption throughout.

Assumption 1 (Design). The predictor variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are deterministic and that $\max\{N_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{1} : \boldsymbol{J}]\} \lesssim n/J^d$, 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 $\zeta_{\boldsymbol{j}}$ and $\lambda_{\boldsymbol{j}}$ in the prior on the coefficients $\theta_{\boldsymbol{j}}$ satisfy $\max_{\boldsymbol{j}} |\zeta_{\boldsymbol{j}}| < \infty$ and $0 < \min_{\boldsymbol{j}} \lambda_{\boldsymbol{j}}^2 \le \max_{\boldsymbol{j}} \lambda_{\boldsymbol{j}}^2 < \infty$.

If the number J of steps in each direction is not chosen deterministically, then it is given a prior supported on $\mathbb{Z}_{>0}$ satisfying the tail condition

$$\exp\{-b_2 J^d \log J\} \le \pi(J) \le \exp\{-b_1 J^d \log J\}, \tag{2.4}$$

where b_1 and b_2 are positive hyperparameters.

To deal with the parameter σ^2, we can plug in the marginal maximum likelihood estimator (MLE) of σ^2. Under the Gaussian working model, the marginal MLE is given by

$$\hat{\sigma}_n^2 = \frac{1}{n}\left[\sum_{i=1}^n \big(Y_i - \sum_{\boldsymbol{j}:\boldsymbol{X}_i \in I_{\boldsymbol{j}}} \zeta_{\boldsymbol{j}}\big)^2 - \sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} \frac{N_{\boldsymbol{j}}^2 (\bar{Y}|_{I_j} - \zeta_{\boldsymbol{j}})^2}{N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}}\right]. \tag{2.5}$$

An alternative is to adopt a fully Bayesian approach, endowing σ^2 with an Inverse-Gamma prior $\mathrm{IG}(\beta_1, \beta_2)$, for some $\beta_1 > 0, \beta_2 > 0$. By conjugacy, the marginal posterior distribution is

$$\sigma^2 | \mathbb{D}_n \sim \mathrm{IG}(\beta_1 + n/2, \beta_2 + n\hat{\sigma}_n^2/2). \tag{2.6}$$

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

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

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

$$\mathcal{C} = \{\boldsymbol{\theta} = (\theta_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{1} : \boldsymbol{J}]) : \theta_{\boldsymbol{j}_1} \le \theta_{\boldsymbol{j}_2}, \text{ if } \boldsymbol{j}_1 \preceq \boldsymbol{j}_2\}. \tag{2.8}$$

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 $p \ge 1$, 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 2.1. *For any f in $\mathcal{F}_J$ and $p \ge 1$, 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,

$$\min_{\boldsymbol{\theta}^* \in \mathcal{C}} \sum_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} |\theta_{\boldsymbol{j}} - \theta^*_{\boldsymbol{j}}|^p G^*(I_{\boldsymbol{j}}). \tag{2.9}$$

The solution of isotonic optimization problem in (2.9) is available in some R packages like 'isotone', see de Leeuw [10]. 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 [37], 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 $G^*(I_{\boldsymbol{j}}) > 0$ for all $\boldsymbol{j}$, 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

$$\min_{\theta^* \in \mathcal{C}} \sum_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} N_{\boldsymbol{j}} |\theta^*_{\boldsymbol{j}} - \theta_{\boldsymbol{j}}|^p, \tag{2.10}$$

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

3. Main results

Let a sample from the projection-posterior defined by the minimization of an $\mathbb{L}_1$-distance, be denoted by $f^* = \sum_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} \theta^*_{\boldsymbol{j}} \mathbb{1}_{\boldsymbol{j}}$. 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 3.1. *Let J be deterministic, Assumptions 2–3 hold and let G^* be a distribution on $[0,1]^d$ possibly depending on n and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ satisfying the conditions that*

$$\mathrm{E}\big[\max_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} G^*(I_{\boldsymbol{j}})\big] \lesssim J^{-d}, \qquad \mathrm{E}\big[\sum_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} G^*(I_{\boldsymbol{j}})(N_{\boldsymbol{j}}+1)^{-1}\big] \lesssim J^d/n. \tag{3.1}$$

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

$$\mathrm{E}_0\Pi(\|f^* - f_0\|_{1,G^*} > M_n\epsilon_n|\mathbb{D}_n) \to 0 \text{ for any } M_n \to \infty. \tag{3.2}$$

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 $J^d(\log n)/n \to 0$. Then the assertion (3.2) 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 [5]. 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 A.1. Hence the $\mathbb{L}_p$-contraction rate of the corresponding procedure will be suboptimal.

The distribution of a random predictor $\boldsymbol{X}$ 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 3.2. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. with distribution G admitting a density function g bounded and bounded away from 0. Let J be deterministic, $J \to \infty$ and $J^d(\log n)/n \to 0$. Then under Assumptions 2 and 3, for $\epsilon_n = \max\{\sqrt{J^d/n}, J^{-1}\}$ and any $M_n \to \infty$, $\mathrm{E}_0\Pi(\|f^* - f_0\|_{1,\lambda} > M_n\epsilon_n|\mathbb{D}_n) \to 0$ where f^* is the $\mathbb{L}_1(G_n)$-projection of f sampled from the unrestricted posterior.*

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, $\Pi(\rho(f, \mathcal{M}) \le \delta_n|\mathbb{D}_n)$, where ρ is a suitable metric, usually an $\mathbb{L}_1$-distance. This idea was also pursued in Salomond [33] and Chakraborty and Ghosal [5] for Bayesian tests for monotonicity in the univariate case, respectively using the $\mathbb{L}_\infty$- 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 C.4 of Ghosal and van der Vaart [17]) even if the alternative approaches the null, provided that happens sufficiently slowly.

Theorem 3.3. *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 $M_n \to \infty$ be predetermined and $J \asymp n^{1/(2+d)}$. Then for the test $\phi_n = \mathbb{1}\{\Pi(\rho(f, \mathcal{M}_J) \le M_n n^{-1/(d+2)} | \mathbb{D}_n) < \gamma\}$, we have*

(i) $\mathrm{E}_0 \phi_n \to 0$ *for any fixed* $f_0 \in \mathcal{M}$;
(ii) $\mathrm{E}_0(1-\phi_n) \to 0$ *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) $\sup\{\mathrm{E}_0(1-\phi_n) : f_0 \in \mathcal{H}(\alpha, L), \rho(f_0, \mathcal{M}) > \tau_n(\alpha)\} \to 0$, *where*

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

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 $\alpha \le 1$ by putting a prior on J and using a larger enlargement in terms of the weaker Hellinger distance on the density

$$p_{f,\sigma}(\boldsymbol{x}, y) = (\sigma\sqrt{2\pi})^{-1} \exp[-(y - f(\boldsymbol{x}))^2/(2\sigma^2)] \tag{3.3}$$

of $(\boldsymbol{X}, Y)$ (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 3.4. *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 $(\boldsymbol{X}, Y)$ induced on the regression function, that is,*

$$\rho^2(f_1, f_2) = 2\big\{1 - (2\pi\sigma^2)^{-1/2} \int \exp[-(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2/(8\sigma^2)] dG(\boldsymbol{x})\big\}. \tag{3.4}$$

Let J be given a prior satisfying (2.4). Consider the test $\phi_n = \mathbb{1}\{\Pi(\rho(f, \mathcal{M}_J) \le M_0\sqrt{(J^d \log n)/n} | \mathbb{D}_n) < \gamma\}$, 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 $f_0 \in \mathcal{M}$, $\mathrm{E}_0 \phi_n \to 0$;*
(ii) *for any fixed f_0 integrable on $[0,1]^d$, and $f_0 \notin \overline{\mathcal{M}}$, $\mathrm{E}_0(1-\phi_n) \to 0$, 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* $C > 0$,

$$\sup\{\mathrm{E}_0(1-\phi_n) : f_0 \in \mathcal{H}(\alpha, L), \rho(f_0, \mathcal{M}) > C(n/\log n)^{-\alpha/(1+2\alpha)}\} \to 0.$$

Remark 3.1. 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 [17] for deterministic predictors.

Remark 3.2. As the distribution G of the random predictor is typically unknown, the tests used in Theorems 3.3 and 3.4 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 (3.4) may be respectively replaced by the Lebesgue $\mathbb{L}_1$-metric and by ρ defined by

$$\rho^2(f_1, f_2) = 2\{1 - (2\pi\sigma^2)^{-1/2} \int \exp[-(f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}))^2/(8\sigma^2)]d\boldsymbol{x}\}. \tag{3.5}$$

Then the conclusions of the theorems hold. For Theorem 3.3, this follows by following the same arguments by using Part (iii) of Theorem 3.1 instead of Part (ii). For Theorem 3.4, we use the equivalence of the metrics (3.4) and (3.5) 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 [17]. 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.

4. Numerical results

4.1. 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 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 $Y = f_0(\boldsymbol{X}) + \varepsilon$ with $\boldsymbol{X}$ uniformly distributed over $[0,1]^2$ and independent errors $\varepsilon \sim \mathrm{N}(0, 0.1^2)$. 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 $\boldsymbol{X}$ 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 1.

TABLE 1

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)

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.

4.2. 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 $\log(dn^{1/4})$ 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 method. We confine to the normal linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$, $i = 1, \ldots, n$. 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, $f_1, \ldots, f_6$ 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 choose $J = 3$ and consider the partition $I_{\boldsymbol{j}}$ for $\boldsymbol{j} = (1,1), \ldots, (3,3)$. On each $I_{\boldsymbol{j}}$, we fit a linear model and test whether any t-value of the slope coefficients $\beta_{1,\boldsymbol{j}}$ and $\beta_{2,\boldsymbol{j}}$ is smaller than $t_{N_{\boldsymbol{j}}-3,1-\eta/18}$ by the Bonferroni adjustment.

We generate 200 datasets for each sample size $n = 100, 200$, and 500. The predictors $\boldsymbol{X}$ 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 and 3.

We can see from Tables 2 and 3 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 looks conservative. However, in the case of f_6, for which the slope coefficients are zero and on the boundary of the null hypothesis, the Bonferroni adjustment seems not that conservative,

TABLE 2
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 3
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

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 the 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 of f_7. The proposed methods enjoy power enhancement when the signal-to-noise ratio increases. We can see this by comparing cases of 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.

5. Proofs

Proof of Proposition 2.1. For a given h, let $\bar{h} = \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \lambda(I_{\boldsymbol{j}})^{-1} \int_{I_{\boldsymbol{j}}} h\mathrm{d}\lambda \cdot \mathbb{1}_{I_{\boldsymbol{j}}}$. Clearly, $\bar{h} \in \mathcal{M}$ if $h \in \mathcal{M}$. Since f is constant on $I_{\boldsymbol{j}}$, for every $\boldsymbol{x} \in I_{\boldsymbol{j}}$,

$$\left| \frac{\int_{I_{\boldsymbol{j}}} h\mathrm{d}\lambda}{\lambda(I_{\boldsymbol{j}})} - f(\boldsymbol{x}) \right|^p = \frac{\left| \int_{I_{\boldsymbol{j}}} (h-f)\,\mathrm{d}\lambda \right|^p}{\lambda(I_{\boldsymbol{j}})^p} \le \frac{\int_{I_{\boldsymbol{j}}} |h-f|^p \mathrm{d}\lambda}{\lambda(I_{\boldsymbol{j}})}, \tag{5.1}$$

by Jensen's inequality. Taking integrals on both sides of (5.1) over $I_{\boldsymbol{j}}$, it follows that $\int_{I_{\boldsymbol{j}}} |\bar{h} - f|^p \, \mathrm{d}\lambda \le \int_{I_{\boldsymbol{j}}} |h - f|^p \, \mathrm{d}\lambda$. 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 $\mathbb{L}_p$-losses. □

Proof of Theorem 3.1. 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 $f_{0,J} = \sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} f_0(\boldsymbol{j}/\boldsymbol{J}) \mathbb{1}_{I_{\boldsymbol{j}}}$. Then $f_{0,J} \in \mathcal{M}_J$. As f^* is the $\mathbb{L}_1(G^*)$-projection of f onto $\mathcal{M}_J$ and $f_0 \in \mathcal{M}$,

$$\begin{aligned} \|f^* - f_0\|_{1,G^*} &\le \|f^* - f\|_{1,G^*} + \|f - f_{0,J}\|_{1,G^*} + \|f_{0,J} - f_0\|_{1,G^*} \\ &\le 2\|f - f_{0,J}\|_{1,G^*} + \|f_{0,J} - f_0\|_{1,G^*}. \end{aligned} \tag{5.2}$$

By Lemma A.2, $\|f_{0,J} - f_0\|_{1,G^*} \lesssim J^{-1}$ as $G^*(I_{\boldsymbol{j}}) \lesssim J^{-1}$ is assumed. Hence it suffices to bound $\|f - f_{0,J}\|_{1,G^*}$.

Without loss of generality, we assume that $N_{\boldsymbol{j}} > 0$ for all $\boldsymbol{j}$. Let $\bar{f}_{0,J} = \sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} \theta_{0,\boldsymbol{j}} \mathbb{1}_{I_{\boldsymbol{j}}}$, where $\theta_{0,\boldsymbol{j}} = N_{\boldsymbol{j}}^{-1} \sum_{i:\boldsymbol{X}_i \in I_{\boldsymbol{j}}} f_0(\boldsymbol{X}_i)$. Then Lemma A.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 $\mathrm{E}_0 \Pi(\|f - \bar{f}_{0,J}\|_{1,G^*} > M_n \sqrt{J^d/n} | \mathbb{D}_n) \to 0$.

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

$$\begin{aligned} &\Pi(\|f - \bar{f}_{0,J}\|_{1,G^*} > M_n \sqrt{J^d/n} | \mathbb{D}_n, \sigma) \\ &= \Pi\Big(\sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} G^*(I_{\boldsymbol{j}}) \, |\theta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}}| > M_n \sqrt{J^d/n} | \mathbb{D}_n, \sigma \Big) \\ &\le \Pi\Big(\sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} G^*(I_{\boldsymbol{j}}) \, |\theta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}}|^2 > M_n^2 J^d/n | \mathbb{D}_n, \sigma \Big) \\ &\le M_n^{-2} J^{-d} \sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} n G^*(I_{\boldsymbol{j}}) \mathrm{E}[(\theta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}})^2 | \mathbb{D}_n, \sigma]. \end{aligned} \tag{5.3}$$

We decompose

$$\mathrm{E}[(\theta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}})^2 | \mathbb{D}_n, \sigma] = \mathrm{Var}(\theta_{\boldsymbol{j}} | \mathbb{D}_n, \sigma) + (\mathrm{E}(\theta_{\boldsymbol{j}} | \mathbb{D}_n, \sigma) - \theta_{0,\boldsymbol{j}})^2. \tag{5.4}$$

We observe that

$$\sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} n G^*(I_{\boldsymbol{j}}) \mathrm{Var}(\theta_{\boldsymbol{j}} | \mathbb{D}_n, \sigma) \le \frac{\sigma^2}{\min\{1, \min_{\boldsymbol{j}} \{\lambda_{\boldsymbol{j}}^{-2}\}\}} \sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} \frac{n G^*(I_{\boldsymbol{j}})}{N_{\boldsymbol{j}} + 1}. \tag{5.5}$$

From (2.3), we know

$$\begin{aligned} &\sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} n G^*(I_{\boldsymbol{j}}) (\mathrm{E}(\theta_{\boldsymbol{j}} | \mathbb{D}_n, \sigma) - \theta_{0,\boldsymbol{j}})^2 \\ &= \sum_{\boldsymbol{j} \in [\boldsymbol{1}:\boldsymbol{J}]} n G^*(I_{\boldsymbol{j}}) \left(\frac{N_{\boldsymbol{j}} \bar{\varepsilon}|_{I_{\boldsymbol{j}}} + \lambda_{\boldsymbol{j}}^{-2} \zeta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}} \lambda_{\boldsymbol{j}}^{-2}}{N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}} \right)^2 \end{aligned}$$

$$\lesssim \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{nG^*(I_{\boldsymbol{j}})N_{\boldsymbol{j}}^2(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2}{(N_{\boldsymbol{j}}+1)^2} + \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{nG^*(I_{\boldsymbol{j}})}{(N_{\boldsymbol{j}}+1)^2}$$
$$\lesssim \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{nG^*(I_{\boldsymbol{j}})}{N_{\boldsymbol{j}}+1} \tag{5.6}$$

by noting that $\mathrm{E}[(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2|\boldsymbol{X},\sigma] = \sigma^2/N_{\boldsymbol{j}}$. Hence, the expectations of the expressions in (5.5) and (5.6) are bounded by a constant multiple of J^d in view of (3.1). Combining these with (5.3) and (5.4), it follows that

$$\mathrm{E}\Pi(\|f - \bar{f}_{0,J}\|_{2,G^*} > M_n\sqrt{J^d/n}|\mathbb{D}_n,\sigma) \lesssim M_n^{-2}, \tag{5.7}$$

and hence the first part of the theorem is established.

If $\max\{N_{\boldsymbol{j}} : \boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]\} \lesssim n/J^d$, then Lemma A.3 ensures that the estimator and the posterior for σ are consistent. For $G^* = G_n$, the condition (3.1) holds because $nG_n(I_{\boldsymbol{j}})(N_{\boldsymbol{j}}+1)^{-1} \le 1$. If $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are i.i.d. with a bounded density g, then $\max\{N_{\boldsymbol{j}} : \boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]\} \lesssim n/J^d$ by Lemma A.1, provided that $J^d(\log n)/n \to 0$. If $G^* = G_n$ for either random or deterministic predictors $\boldsymbol{X}_i$, (5.5) is bounded by J^d up to some positive constant. If $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are i.i.d. G, then owing to $N_{\boldsymbol{j}} \sim \mathrm{Bin}(n, G(I_{\boldsymbol{j}}))$, we have that

$$\mathrm{E}_0[(N_{\boldsymbol{j}}+1)^{-1}] = \begin{cases} \frac{1-(1-G(I_{\boldsymbol{j}}))^{n+1}}{(n+1)G(I_{\boldsymbol{j}})}, & \text{if } G(I_{\boldsymbol{j}}) > 0; \\ 1, & \text{if } G(I_{\boldsymbol{j}}) = 0, \end{cases} \tag{5.8}$$

so that $nG(I_{\boldsymbol{j}})\mathrm{E}[(N_{\boldsymbol{j}}+1)^{-1}] \le 1$, implying that (3.1) holds for $G^* = G$. This completes the proof of the second part of the theorem. □

Proof of Corollary 3.2. For f^* the $\mathbb{L}_1(G_n)$-projection, by the triangle inequality,

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

where, as in the proof of the last theorem, $\bar{f}_{0,J} = \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}\theta_{0,\boldsymbol{j}}\mathbb{1}_{I_{\boldsymbol{j}}}$, with $\theta_{0,\boldsymbol{j}} = N_{\boldsymbol{j}}^{-1}\sum_{i:\boldsymbol{X}_i\in I_{\boldsymbol{j}}} f_0(\boldsymbol{X}_i)$. From Lemma A.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 3.1,

$$\mathrm{E}_0\Pi(\|f^* - \bar{f}_{0,J}\|_{1,G_n} > M_n\epsilon_n|\mathbb{D}_n) \to 0, \tag{5.9}$$

since we also have $\|\bar{f}_{0,J} - f_0\|_{1,G_n} \lesssim J^{-1}$ by Lemmas A.1 and A.2. Thus it suffices to show that

$$\mathrm{E}_0\Pi(\big|\|f^* - \bar{f}_{0,J}\|_{1,G} - \|f^* - \bar{f}_{0,J}\|_{1,G_n}\big| > M_n\epsilon_n|\mathbb{D}_n) \to 0. \tag{5.10}$$

Clearly, we have

$$\big|\|f^* - \bar{f}_{0,J}\|_{1,G} - \|f^* - \bar{f}_{0,J}\|_{1,G_n}\big|$$

$$\leq \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} G_n(I_{\boldsymbol{j}})|\theta^*_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}}| \cdot \max_{\boldsymbol{j}} |G(I_{\boldsymbol{j}})/G_n(I_{\boldsymbol{j}}) - 1|. \tag{5.11}$$

Under the additional condition on the lower bound for g, Lemma A.1 implies that the last factor is $O_{P_0}(1)$. Thus (5.11) is bounded by a constant multiple of $\|f^* - \bar{f}_{0,J}\|_{1,G_n}$ on an event with P_0-probability tending to 1. Then this claim follows from Theorem 3.1. As g is bounded and bounded away from 0, $\|f^* - f_0\|_{1,G} \asymp \|f^* - f_0\|_{1,\lambda}$, then the corollary follows. □

Proof of Theorem 3.3. (i) Since $\rho(f, \mathcal{M}_J) \leq \|f - f_{0,J}\|_{1,G}$, the conclusion follows from Theorem 3.1.

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

$$\rho(f, \mathcal{M}_J) \geq \|f_0 - f^*\|_{1,G} - \|f - f_0\|_{1,G} \geq \rho(f_0, \mathcal{M}_J) - \|f - f_0\|_{1,G}. \tag{5.12}$$

Thus by the triangle inequality,

$$\begin{aligned} &\Pi(\rho(f, \mathcal{M}) \leq M_n n^{-1/(d+2)} | \mathbb{D}_n) \\ &\leq \Pi(\|f - f_0\|_{1,G} \geq \rho(f_0, \mathcal{M}_J) - M_n n^{-1/(d+2)} | \mathbb{D}_n). \end{aligned} \tag{5.13}$$

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 $\theta_{0,\boldsymbol{j}} = \int_{I_{\boldsymbol{j}}} f_0 \mathrm{d}G/G(I_{\boldsymbol{j}})$ and then $f_{0,J} = \sum_{\boldsymbol{j}} \theta_{0,\boldsymbol{j}} \mathbb{1}_{I_{\boldsymbol{j}}}$. By the martingale convergence theorem, $\|f_0 - f_{0,J}\|_{1,G} \to 0$. Proceeding as in the proof of Theorem 3.1, we conclude that

$$\mathrm{E}_0 \Pi(\|f - f_{0,J}\|_{1,G} > M_n \sqrt{J^d/n} | \mathbb{D}_n) \to 0, \tag{5.14}$$

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 (5.14), 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 (5.13) is, for large n, bounded by $\Pi(\|f - f_0\|_{1,G} \geq C n^{-\alpha/(2+d)}/2) \to_{P_0} 0$, since $n^{-\alpha/(2+d)} \gg n^{-1/(2+d)}$. If $\alpha = 1$, the corresponding bound for the event of interest reduces to $\Pi(\|f - f_0\|_{1,G} \geq (C-1) M_n n^{-1/(d+2)}/2 | \mathbb{D}_n) \to_{P_0} 0$. □

Proof of Theorem 3.4. With $p_{f,\sigma}$ defined by (3.3), 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

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

Thus the Kullback-Leibler ball $\{f : K(p_{f_0,\sigma}; p_{f,\sigma}) \leq \epsilon^2, V(p_{f_0,\sigma}; p_{f,\sigma^2}) \leq \epsilon^2\}$ 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 $\|f_0 - f_{0,J}\|^2_{2,G} \leq (f_0(\mathbf{1}) - f_0(\mathbf{0}))\|f_0 - f_{0,J}\|_{1,G}$, 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 $\|f_0 - f_{0,J}\|_{1,G} \le c\epsilon^2$ for some sufficiently small $c > 0$. If a value $\bar{J}$, possibly depending on n, achieves this, then using (2.4), we can lower bound the required $\mathbb{L}_2(G)$-prior concentration by

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

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

$$\Pi(J \ge J_n)/\Pi(\|f - f_0\|_{2,G} \le C\epsilon) = o(e^{-2n\epsilon^2}), \tag{5.15}$$

and hence by Theorem 8.20 of Ghosal and van der Vaart [17], $\Pi(J > J_n|\mathbb{D}_n) \to_{P_0} 0$. If $\epsilon = \epsilon_n \to 0$ is chosen so that $n\epsilon_n^2 \to \infty$ 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 $\Pi(J \ge J_n)/\Pi(\|f - f_0\|_{2,G} \le C\epsilon_n) = o(e^{-2n\epsilon_n^2})$. Hence by Theorem 8.20 of Ghosal and van der Vaart [17] again, $\Pi(J > J_n|\mathbb{D}_n) \to_{P_0} 0$.

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

$$\mathrm{E}_0\Pi(\|f - f_{0,J}\|_{2,G} \ge M_0\sqrt{J^d(\log n)/n}, J \le J_n|\mathbb{D}_n) < \delta, \tag{5.16}$$

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

$$\sum_{J=1}^{J_n} \Pi(J|\mathbb{D}_n)\Pi\Big(\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} (\theta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}})^2 G(I_{\boldsymbol{j}}) \ge M_0^2 J^d(\log n)/n \Big| \mathbb{D}_n\Big). \tag{5.17}$$

By Markov's inequality and Assumption 3,

$$\begin{aligned}&\max_{J\le J_n} \Pi\Big(\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}_n]} (\theta_{\boldsymbol{j}} - \theta_{0,\boldsymbol{j}})^2 G(I_{\boldsymbol{j}}) \ge M_0^2 J^d(\log n)/n \Big| \mathbb{D}_n\Big)\\ &\quad\le \max_{J\le J_n} \frac{n}{M_0^2 J^d \log n} \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}_n]} G(I_{\boldsymbol{j}})\big[\mathrm{Var}(\theta_{\boldsymbol{j}}|\mathbb{D}_n) + (\mathrm{E}(\theta_{\boldsymbol{j}}|\mathbb{D}_n) - \theta_{0,\boldsymbol{j}})^2\big]\end{aligned}$$

which is bounded in probability by a constant multiple of

$$\max_{J\le J_n} \frac{n}{M_0^2 J^d \log n} \sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} G(I_{\boldsymbol{j}})[(N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2})^{-1} + (\bar{Y}|_{I_{\boldsymbol{j}}} - \theta_{0,\boldsymbol{j}})^2] \tag{5.18}$$

It is clear that $G(I_{\boldsymbol{j}}) \asymp J^{-d}$. By Lemma A.1,

$$\mathrm{P}_0\Big(\bigcap_{J=1}^{J_n}\{C_1 n/J^d \le \min_{\boldsymbol{j}} N_{\boldsymbol{j}} \le \max_{\boldsymbol{j}} N_{\boldsymbol{j}} \le C_2 n/J^d\}\Big) \to 1,$$

provided $n/J_n \gg \log J_n$, for two constant C_1 and $C_2 > 0$. Then $N_{\boldsymbol{j}} \asymp n/J^d$ uniformly for all $\boldsymbol{j} \le \boldsymbol{J}$ and J. By the union bound of sub-Gaussian variables (see van der Vaart and Wellner [39], 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 $|N_{\boldsymbol{j}}^{-1} \sum_{i:\boldsymbol{X}_i \in I_j} f_0(\boldsymbol{X}_i) - \theta_{0,\boldsymbol{j}}|^2 \lesssim J^d(\log n)/n$ uniformly for all $\boldsymbol{j}$ and J with high probability. Thus we establish the claim in (5.16).

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 $\Pi(J > J_n|\mathbb{D}_n) \to_{\mathrm{P}_0} 0$. Since $\rho(f, \mathcal{M}_J) \lesssim \rho(f, \mathcal{M}) \le \rho(f, f_0)$, the claim follows from (5.16).

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 (5.15), and consequently, $\Pi(J > J_n|\mathbb{D}_n) \to_{\mathrm{P}_0} 0$. Let $\mathcal{F}_n^* = \bigcup_{J=1}^{J_n}\{\sum_{\boldsymbol{j}\in[\boldsymbol{1},\boldsymbol{J}_n]} \theta_{\boldsymbol{j}} \mathbb{1}_{I_j} : |\theta_{\boldsymbol{j}}| \le n\}$. 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 $J_n^d \log n \le n\epsilon^2$. 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 [17]). Therefore, as $\rho(f_0, \mathcal{M}_J)$ is bounded by a positive fixed constant from below, by (5.12), it follows that $\Pi(\rho(f, \mathcal{M}_J) \le M_0\sqrt{(J^d \log n)/n}|\mathbb{D}_n) \to_{\mathrm{P}_0} 0$.

To prove Part (iii), we observe by Lemma A.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 [17] that the $\mathbb{L}_1(G)$-posterior contraction rate is $(n/\log n)^{-\alpha/(2\alpha+d)}$. Thus, as $\rho(f_0, \mathcal{M}_J) \ge C(n/\log n)^{-\alpha/(2\alpha+d)}$ for a sufficiently large constant $C > 0$, from (5.12) and the probabilistic bound $(n/\log n)^{1/(2\alpha+d)}$ for J, the conclusion follows. □

A. Auxiliary results

Lemma A.1. *If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are a random sample from a density g on $[0,1]$, $J \to \infty$, and $n/J^d \gg \log J$. If g is bounded, then for some constants $C > 0$,*

$$\mathrm{P}_0(\max\{N_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{1} : \boldsymbol{J}]\} \le Cn/J^d) \to 1.$$

If g is bounded away from zero, then for some constant $C' > 0$, we have

$$\mathrm{P}_0(\min\{N_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{1} : \boldsymbol{J}]\} \ge C'n/J^d) \to 1.$$

Proof. For every $\boldsymbol{j}$, $N_{\boldsymbol{j}} \sim \mathrm{Bin}(n, G(I_{\boldsymbol{j}}))$. If g is bounded from above by a, then $G(I_{\boldsymbol{j}})$ is bounded by a/J^d. Following the same argument of the proof of Lemma A.2 of Chakraborty and Ghosal [6], we obtain that, by large deviation probability, $\mathrm{P}_0(N_{\boldsymbol{j}} > Cn/J^d) \le 2\exp\{-C'n/J^d\}$. By the condition $n/\log J \gg J^d$, we have $\mathrm{P}_0(\max N_{\boldsymbol{j}} > Cn/J^d) \le 2\exp\{-C''n/J^d\} \to 0$. The second claim follows from a similar argument. □

Lemma A.2. *Let G^* be a probability measure on $[0,1]^d$ such that $\max\{G^*(I_{\boldsymbol{j}}) : \boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]\} \lesssim J^{-d}$. For a given $f : [0,1]^d \to \mathbb{R}$ and J, let $f_J : [0,1]^d \to \mathbb{R}$ be defined by $f_J(\boldsymbol{x}) = \sum_{\boldsymbol{j} \in [\mathbf{1}:\boldsymbol{J}]} \theta_{\boldsymbol{j}} \mathbb{1}\{\boldsymbol{x} \in I_{\boldsymbol{j}}\}$, $\boldsymbol{x} \in [0,1]^d$, where $\theta_{\boldsymbol{j}}$ is any value between $f((\boldsymbol{j}-\mathbf{1})/J)$ and $f(\boldsymbol{j}/J)$. Then $\|f - f_J\|_{p,G^*} \lesssim J^{-1/p}$. Moreover, for some appropriate choices of $\theta_{\boldsymbol{j}}$, $\boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]$, we can ensure that $f \in \mathcal{M}$.*

Proof. For $\theta_{\boldsymbol{j}}$ any value between $f((\boldsymbol{j}-\mathbf{1})/J)$ and $f(\boldsymbol{j}/J)$,

$$\begin{aligned}\|f - f_J\|_{1,G^*} &= \sum_{\boldsymbol{j}} \int_{I_{\boldsymbol{j}}} |f - \theta_{\boldsymbol{j}}| \mathrm{d}G^* \\ &\le \sum_{\boldsymbol{j}} (f(\boldsymbol{j}/J) - f((\boldsymbol{j}-\mathbf{1})/J)) G^*(I_{\boldsymbol{j}}) \\ &\lesssim J^{-d} \sum_{\boldsymbol{j}} (f(\boldsymbol{j}/J) - f((\boldsymbol{j}-\mathbf{1})/J)).\end{aligned}$$

To get the upper bound of the summation in the last inequality, we first decompose the index set $[\mathbf{1} : \boldsymbol{J}]$ in the following way. For every $\boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]$, Let $A_{\boldsymbol{j}}$ be the largest possible subset of $[\mathbf{1} : \boldsymbol{J}]$ in the form $\{\ldots, \boldsymbol{j} - 2\cdot\mathbf{1}, \boldsymbol{j} - \mathbf{1}, \boldsymbol{j}, \boldsymbol{j} + \mathbf{1}, \boldsymbol{j} + 2\cdot\mathbf{1}, \ldots\}$, which is a chain with respect to the coordinatewise partial order on the index set. Then we count the number of different $A_{\boldsymbol{j}}$. Note that $A_{\boldsymbol{j}}$ can be identified by its minimal element. The minimal element of $A_{\boldsymbol{j}}$ should satisfy that at least one of its coordinates is 1, otherwise, we can subtract this element by $\mathbf{1}$ while the smaller element is still in $[\mathbf{1} : \boldsymbol{J}]$, thus should be in $A_{\boldsymbol{j}}$, 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 $\{1, \ldots, J\}$. The construction of $A_{\boldsymbol{j}}$ gives $\sum_{\boldsymbol{l} \in A_{\boldsymbol{j}}} (f(\boldsymbol{l}/J) - f((\boldsymbol{l}-\mathbf{1})/J)) \le f(\mathbf{1}) - f(\mathbf{0})$. 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 $\boldsymbol{j} \in [\mathbf{1}, \boldsymbol{J}]$, $\theta_{\boldsymbol{j}} = \int_{I_{\boldsymbol{j}}} f(\boldsymbol{x}) d\boldsymbol{x}/G(I_{\boldsymbol{j}})$, or $\theta_{\boldsymbol{j}} = f((\boldsymbol{j}-\mathbf{1})/J)$, for instance.

For $p > 1$, note that $\|f - f_J\|_{p,G^*}^p \le (f(\mathbf{1}) - f(\mathbf{0}))^{p-1} \|f - f_J\|_{1,G^*}$. Then the conclusion follows. □

Remark A.1. For $p > 1$, the $\mathbb{L}_p$-approximation rate in Lemma A.2 may not be improved. To see this, consider $f = \sum_{j=1}^d \mathbb{1}\{j : x_j > c_j\}$, where $\boldsymbol{c}$ is a fixed vector with irrational coordinates in $[0,1]$. Note that $\boldsymbol{c}$ is never on the boundary of any hypercube used for partitioning. Clearly, f is a multivariate monotone function with a discontinuity at any $\boldsymbol{x}$ that shares a coordinate with $\boldsymbol{c}$. Let $\boldsymbol{j}^*$ be the index such that $\boldsymbol{c} \in I_{\boldsymbol{j}^*}$. and generally for a given J, for $k = 1, \ldots, d$,

$\min\{c_{j_k^*} - (j_k^* - 1)/J, j_k^*/J - c_{j_k^*}\} \gtrsim 1/J$. For any hypercube $I_{\boldsymbol{j}}$ used in the partition such that $j_k = j_k^*$ for some $k = 1, \ldots, d$, there is a jump of size at least 1 within $I_{\boldsymbol{j}}$. Hence, no matter how $\boldsymbol{\theta}$ is chosen, $\int_{I_{\boldsymbol{j}}} |f - f_J|^p \gtrsim J^{-d}$ 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 A.2. In view of Lemma A.1, if $J^d(\log n)/n \to 0$, then the empirical distribution satisfies the condition $\max\{G_n(I_{\boldsymbol{j}}) : \boldsymbol{j} \in [\mathbf{1} : \boldsymbol{J}]\} \lesssim J^{-d}$ 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.3. *Suppose J is deterministic and satisfies $J \to \infty$ and $J^d/n \to 0$. For $\boldsymbol{X}$ 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(β_1, β_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 $\theta_{0,\boldsymbol{j}} = N_{\boldsymbol{j}}^{-1} \sum_{i:X_i \in I_{\boldsymbol{j}}} f_0(X_i)$. 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(\boldsymbol{X}_i) - \theta_{0,\lceil \boldsymbol{X}_i J\rceil})^2 + \frac{1}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} N_{\boldsymbol{j}}(\theta_{0,\boldsymbol{j}} - \zeta_{\boldsymbol{j}})^2 \\
&+ \frac{2}{n}\sum_{i=1}^n \varepsilon_i (f_0(\boldsymbol{X}_i) - \theta_{0,\lceil \boldsymbol{X}_i J\rceil}) + \frac{2}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} N_{\boldsymbol{j}}\bar{\varepsilon}|_{I_{\boldsymbol{j}}}(\theta_{0,\boldsymbol{j}} - \zeta_{\boldsymbol{j}}) \\
&+ \frac{2}{n}\sum_{i=1}^n (f_0(\boldsymbol{X}_i) - \theta_{0,\lceil \boldsymbol{X}_i J\rceil})(\theta_{0,\lceil \boldsymbol{X}_i J\rceil} - \zeta_{\lceil \boldsymbol{X}_i J\rceil}) \\
&- \frac{1}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{N_{\boldsymbol{j}}^2(\theta_{0,\boldsymbol{j}} - \zeta_{\boldsymbol{j}})^2 + N_{\boldsymbol{j}}^2(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2 + 2N_{\boldsymbol{j}}^2\bar{\varepsilon}|_{I_{\boldsymbol{j}}}(\theta_{0,\boldsymbol{j}} - \zeta_{\boldsymbol{j}})}{N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}} \\
=& \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{1}{n}\sum_{i=1}^n (f_0(\boldsymbol{X}_i) - \theta_{0,\lceil \boldsymbol{X}_i J\rceil})^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i (f_0(\boldsymbol{X}_i) - \theta_{0,\lceil \boldsymbol{X}_i J\rceil}) \\
&+ \frac{2}{n}\sum_{i=1}^n (f_0(\boldsymbol{X}_i) - \theta_{0,\lceil \boldsymbol{X}_i J\rceil})(\theta_{0,\lceil \boldsymbol{X}_i J\rceil} - \zeta_{\lceil \boldsymbol{X}_i J\rceil}) \\
&+ \frac{1}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{\lambda_{\boldsymbol{j}}^{-2} N_{\boldsymbol{j}}(\theta_{0,\boldsymbol{j}} - \zeta_{\boldsymbol{j}})^2}{N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}} + \frac{2}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{\lambda_{\boldsymbol{j}}^{-2} N_{\boldsymbol{j}}\bar{\varepsilon}|_{I_{\boldsymbol{j}}}(\theta_{0,\boldsymbol{j}} - \zeta_{\boldsymbol{j}})}{N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}} \\
&+ \frac{1}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]} \frac{N_{\boldsymbol{j}}^2(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2}{N_{\boldsymbol{j}} + \lambda_{\boldsymbol{j}}^{-2}}.
\end{aligned}
$$

Note that $\lambda_{\boldsymbol{j}}^{-2}$, $\zeta_{\boldsymbol{j}}$ and f_0 are all bounded. Then we can bound $|\hat{\sigma}_n^2 - \sigma_0^2|$ up to

a constant by

$$\begin{aligned}&\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2-\sigma_0^2\Big|+\frac{1}{n}\sum_{i=1}^{n}|f_0(\boldsymbol{X}_i)-\theta_{0,\lceil \boldsymbol{X}_i J\rceil}|+\frac{1}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}(\theta_{0,\boldsymbol{j}}-\zeta_{\boldsymbol{j}})^2\\&+\frac{1}{n}\Big|\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}\frac{N_{\boldsymbol{j}}\bar{\varepsilon}|_{I_{\boldsymbol{j}}}(\theta_{0,\boldsymbol{j}}-\zeta_{\boldsymbol{j}})}{N_{\boldsymbol{j}}+\lambda_{\boldsymbol{j}}^{-2}}\Big|+\frac{1}{n}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}N_{\boldsymbol{j}}(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2.\end{aligned}\tag{A.1}$$

The first term of (A.1) is $O_{\mathrm{P}_0}(n^{-1/2})$. By the monotonicity of f_0, the second term is bounded by $n^{-1}\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}N_{\boldsymbol{j}}(f_0(\boldsymbol{j}/\boldsymbol{J})-f_0((\boldsymbol{j}-\mathbf{1})/\boldsymbol{J}))$. By Remark A.2, following the same argument of the proof of Lemma A.2, we have the second term is $O_{\mathrm{P}_0}(J^{-1})$ for random $\boldsymbol{X}$ and $O(J^{-1})$ for deterministic $\boldsymbol{X}$ under Assumption 1. The third term is bounded by a constant multiple of J^d/n since the hyperparameters $\zeta_{\boldsymbol{j}}$ and $\theta_{0,\boldsymbol{j}}$ are bounded. Noting that $\mathrm{E}[(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2|\boldsymbol{X}]=\sigma_0^2/N_{\boldsymbol{j}}$, by Markov inequality, we know that the last term is $O_{\mathrm{P}_0}(J^d/n)$. For the fourth term, by Cauchy–Schwarz inequality, we have

$$\Big|\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}\frac{N_{\boldsymbol{j}}\bar{\varepsilon}|_{I_{\boldsymbol{j}}}}{N_{\boldsymbol{j}}+\lambda_{\boldsymbol{j}}^{-2}}(\overline{f_0(\boldsymbol{X})}|_{I_{\boldsymbol{j}}}-\zeta_{\boldsymbol{j}})\Big|\lesssim J^{d/2}\sqrt{\sum_{\boldsymbol{j}\in[\mathbf{1}:\boldsymbol{J}]}(\bar{\varepsilon}|_{I_{\boldsymbol{j}}})^2}=O_{\mathrm{P}_0}(J^d).$$

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 [43]. □

Acknowledgments

The authors would like to thank two anonymous referees, the Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

References

[1] Akakpo, N., Balabdaoui, F. and Durot, C. (2014). Testing monotonicity via local least concave majorants. *Bernoulli* **20** 514–544. MR3178508

[2] 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. MR0326887

[3] Bellec, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383

[4] Brunk, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)* 177–197. Cambridge Univ. Press, London. MR0277070

[5] Chakraborty, M. and Ghosal, S. (2021). Convergence rates for Bayesian estimation and testing in monotone regression. *Electron. J. Stat.* **15** 3478–3503. MR4280172

[6] Chakraborty, M. and Ghosal, S. (2021). Coverage of credible intervals in nonparametric monotone regression. *Ann. Statist.* **49** 1011–1028. MR4255117

[7] Chakraborty, M. and Ghosal, S. (2022). Rates and coverage in Bayesian inference for monotone densities. *Bernoulli* **23** 1093–1019. MR4388931

[8] Chatterjee, S., Guntuboyina, A. and Sen, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. MR3706788

[9] Chipman, H. A., George, E. I., McCulloch, R. E. and Shively, T. S. (2022). mBART: multidimensional monotone BART. *Bayesian Anal.* **17** 515–544. MR4483229

[10] 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** 1–24.

[11] Deng, H. and Zhang, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs. *Ann. Statist.* **48** 3672–3698. MR4185824

[12] Durot, C. (2007). On the $\mathbb{L}_p$-error of monotonicity constrained estimators. *Ann. Statist.* **35** 1080–1104. MR2341699

[13] Durot, C., Kulikov, V. N. and Lopuhaä, H. P. (2012). The limit distribution of the L_∞-error of Grenander-type estimators. *Ann. Statist.* **40** 1578–1608. MR3015036

[14] Dykstra, R. L. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Ann. Statist.* **10** 708–716. MR663427

[15] 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** 6389–6402. MR4173546

[16] Ghosal, S., Sen, A. and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054–1082. MR1810919

[17] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge University Press, Cambridge. MR3587782

[18] Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957). MR0093415

[19] 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),* 539–555. Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA. MR0822052

[20] Groeneboom, P. (1989). Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields* **81** 79–109. MR981568

[21] Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.* **28** 20–39.

MR1762902
[22] Han, Q. (2021). Set structured global empirical risk minimizers are rate optimal in general dimensions. *Ann. Statist.* **49** 2642–2671. MR4338378
[23] Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2019). Isotonic regression in general dimensions. *Ann. Statist.* **47** 2440–2471. MR3988762
[24] Han, Q. and Zhang, C.-H. (2020). Limit distribution theory for block estimators in multiple isotonic regression. *Ann. Statist.* **48** 3251–3282. MR4185808
[25] Kulikov, V. N. and Lopuhaä, H. P. (2005). Asymptotic normality of the L_k-error of the Grenander estimator. *Ann. Statist.* **33** 2228–2255. MR2211085
[26] Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349
[27] Luss, R. and Rosset, S. (2014). Generalized isotonic regression. *J. Comput. Graph. Statist.* **23** 192–210. MR3173767
[28] Meyer, M. C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Comm. Statist. Simulation Comput.* **42** 1126–1139. MR3039672
[29] Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60** 398–406. MR2066274
[30] Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36. MR0267677
[31] 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. MR961262
[32] Saarela, O. and Arjas, E. (2011). A method for Bayesian monotonic multiple regression. *Scand. J. Stat.* **38** 499–513. MR2833843
[33] Salomond, J.-B. (2014). Adaptive Bayes test for monotonicity. In *The contribution of young researchers to Bayesian statistics. Springer Proc. Math. Stat.* **63** 29–33. Springer, Cham. MR3133254
[34] Scott, J. G., Shively, T. S. and Walker, S. G. (2015). Nonparametric Bayesian testing for monotonicity. *Biometrika* **102** 617–630. MR3394279
[35] Shively, T. S., Sager, T. W. and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **71** 159–175. MR2655528
[36] Spouge, J., Wan, H. and Wilbur, W. J. (2003). Least squares isotonic regression in two dimensions. *J. Optim. Theory Appl.* **117** 585–605. MR1989929
[37] Stout, Q. F. (2013). Isotonic regression via partitioning. *Algorithmica* **66** 93–112. MR3023808
[38] Stout, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica* **71** 450–470. MR3331888
[39] 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. MR1385671
[40] Vittorietti, M., Hidalgo, J., Sietsma, J., Li, W. and Jongbloed, G. (2022). Isotonic regression for metallic microstructure data: estimation and testing under order restrictions. *J. Appl. Stat.* **49** 2208–2227. MR4436260
[41] Wang, K. and Ghosal, S. (2022). Coverage of Credible Intervals in Bayesian Multivariate Isotonic Regression. *arXiv preprint.*
[42] Westling, T., Gilbert, P. and Carone, M. (2020). Causal isotonic regression. *J. Roy. Stat. Soc. Ser. B. Stat. Methodol.* **82** 719–747. MR4112782
[43] Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.* **44** 1069–1102. MR3485954
[44] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898