Electronic Journal of Statistics
Vol. 15 (2021) 3478–3503
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1861

Convergence rates for Bayesian estimation and testing in monotone regression[*]

Moumita Chakraborty

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

and

Subhashis Ghosal

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

Abstract: Shape restrictions such as monotonicity on functions often arise naturally in statistical modeling. We consider a Bayesian approach to the estimation of a monotone regression function and testing for monotonicity. We construct a prior distribution using piecewise constant functions. For estimation, a prior imposing monotonicity of the heights of these steps is sensible, but the resulting posterior is harder to analyze theoretically. We consider a "projection-posterior" approach, where a conjugate normal prior is used, but the monotonicity constraint is imposed on posterior samples by a projection map onto the space of monotone functions. We show that the resulting posterior contracts at the optimal rate $n^{-1/3}$ under the $\mathbb{L}_1$-metric and at a nearly optimal rate under the empirical $\mathbb{L}_p$-metrics for $0 < p \leq 2$. The projection-posterior approach is also computationally more convenient. We also construct a Bayesian test for the hypothesis of monotonicity using the posterior probability of a shrinking neighborhood of the set of monotone functions. We show that the resulting test has a universal consistency property and obtain the separation rate which ensures that the resulting power function approaches one.

Keywords and phrases: Monotonicity, posterior contraction, Bayesian testing, projection-posterior.

Received June 2020.

1. Introduction

We consider the nonparametric regression model $Y = f(X) + \varepsilon$ for a response variable Y with respect to a one-dimensional predictor variable X taking values

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

in a bounded interval, and ε a mean-zero random error with finite variance σ^2. Instead of the more commonly imposed smoothness condition, here we assume that f is a monotone increasing function on a bounded interval, which can be taken to be $[0,1]$ without loss of generality. We observe n independent replications $(X_1,Y_1),\dots,(X_n,Y_n)$, where the design points $X_1,\dots,X_n$ are either deterministic or are randomly sampled from a fixed distribution G. The error ε is assumed to be distributed independently of the predictor X.

The problem has been widely studied in the frequentist literature, and is commonly known as isotonic regression. Barlow and Brunk [5] obtained the greatest convex minorant (GCM) of a cumulative sum diagram as the least-square estimator under the monotonicity constraint. The Pool-Adjacent-Violators Algorithm (PAVA) describes a method of successive approximation to the GCM, and is the most commonly used algorithm for isotonic regression (see Ayer et al. [2], Barlow et al. [4], or De Leeuw et al. [19]). Brunk [12] showed that the estimated value of the regression function at a point converges at a rate $n^{-1/3}$, and obtained its asymptotic distribution. Durot [20] established the $n^{-1/3}$ rate of convergence of the isotonic regression estimator under the $\mathbb{L}_1$-metric. Testing for monotonicity of a regression function has been studied in the frequentist literature by Akakpo et al. [1], Hall and Heckman [25], Baraud et al. [3], Ghosal et al. [21] and Bowman et al. [11].

A Bayesian approach to the monotone regression problem involves putting a prior on functions under the monotonicity constraint. Since step-functions can approximate monotone functions, a natural approach is to put priors on step heights under the monotonicity constraint, and possibly also on the locations and the number of intervals. For smoother sample paths, higher-order splines can be used instead of the indicator functions of intervals. To put a prior on a monotone regression function, Neelon and Dunson [28] used a basis consisting of piecewise linear functions, put a prior on the coefficients, and developed Markov chain Monte Carlo (MCMC) methods for posterior computation. Shivley et al. [34] used a spline basis and a mixture of constrained normal distributions as a prior for the coefficients in the basis expansion. The resulting procedure can be approximated by a constrained regression spline prior, which leads to an MCMC algorithm on the space of coefficients. They also discussed posterior consistency. Bornkamp and Ickstadt [10] modeled a monotone function as a mixture of parametric probability distribution functions, used a general random probability measure as a prior for the mixing distribution and developed MCMC methods for drawing posterior samples. Chipman et al. [16] put a prior on multivariate monotone regression function through a Bayesian additive regression tree structure by restricting the stepheights that leads to multivariate monotonicity of the function, and devised an MCMC posterior sampling technique.

A Bayesian approach to testing monotonicity was proposed by Salomond [30, 32] based on the posterior probability of the event that the regression function is nearly monotone in the sense of a distance measure. The relaxation to near monotonicity instead of the exact monotonicity is needed to avoid the problem of falsely rejecting the null hypothesis of monotonicity because a monotone true function may be approximated by non-monotone functions, leading to a high probability of type I error (see Section 4). Scott et al. [33] used a more classical approach based on Bayes factors to test the hypothesis of monotonicity using an integrated Brownian motion or constrained regression spline prior on the regression function. They converted the hypothesis of monotonicity to a statement about the minimum value of the derivative function, and gave formula and results on the consistency of the Bayes factor. Coverage of Bayesian credible regions for monotone regression has been recently studied by Chakraborty and Ghosal [14]. They computed the limit and observed an interesting phenomenon that a posterior quantile interval for the value of the function has asymptotic coverage higher than the corresponding credibility level. This is the opposite of the phenomenon of less asymptotic coverage of Bayesian credible regions observed by Cox [17] for smooth function estimation. Moreover, they showed that starting with a suitable lower credibility level, which can be calculated and depends only on the target coverage, the intended asymptotic coverage can be obtained. Bayesian nonparametric methods have been developed also for other shape-constrained problems, such as monotone density and the current status censoring model. Salomond [31] established the near minimax rate $n^{-1/3}$ for a decreasing density using a mixture of uniform densities as a prior. Chakraborty and Ghosal [15] studied posterior contraction rate and the limiting coverage of a Bayesian credible interval for a monotone decreasing density, and constructed a Bayesian test for the hypothesis of monotonicity. For a monotone regression function, asymptotic coverage of a credible interval for a regression quantile was obtained by Chakraborty and Ghosal [13]. They also showed that the posterior contraction rate may be improved by sampling in two stages, where the second stage samples are collected from the credible interval obtained in the first stage.

A difficulty with the usual Bayesian approach to isotonic regression is that the monotonicity constraint on the coefficient makes both posterior computation and study of posterior concentration with increasing sample size a lot more challenging. This is because a neighborhood of the true regression function contains non-monotone functions, which must be discounted for posterior sampling and estimating the prior concentration near the true regression function. A very useful approach that can still utilize the conjugacy structure is provided by a "projection-posterior" distribution. In this approach, the monotonicity constraint on the step size is initially ignored, so that the coefficient may be given independent normal priors. Hence the posterior distribution is also normal, allowing easy sampling, and large sample analysis of posterior concentration. Then a posterior distribution is directly induced by a projection map that projects a step function to the nearest monotone function in terms of the $\mathbb{L}_1$-distance or some other metric. A similar idea based on a Gaussian process prior was used by Lin and Dunson [27] for monotone regression. Bhaumik and Ghosal [6, 7, 8] used this idea of embedding in an unrestricted space and then projecting a conjugate posterior in regression models driven by ordinary differential equations. Bhaumik et al. [9] extended their approach to generalized regression described by partial differential equations. In this paper, we pursue the projection-posterior approach and show that the resulting projection-posterior distribution concentrates at the optimal rate $n^{-1/3}$ in terms of the $\mathbb{L}_1$-distance.

We also obtain nearly optimal posterior concentration under an empirical $\mathbb{L}_p$-distance for $0 < p \leq 2$ using a different prior. In addition to being extremely convenient for theoretical studies, the approach is very convenient for posterior computation as well, allowing the use of the convenient conjugate prior and the resulting posterior sampling without needing to use more computationally expensive MCMC sampling. It may be mentioned that Chakraborty and Ghosal [14, 15, 13] also followed the projection-posterior approach, and indeed the prior used in the present paper was also used in Chakraborty and Ghosal [14, 13]. In this paper, we also construct a Bayesian test for the hypothesis of monotonicity based on the posterior distribution of the difference between the unrestricted posterior sample and its projection. We show that the resulting test is universally consistent, in that the Type I error probability goes to zero and the power goes to one at any fixed alternative, regardless of smoothness. For a sequence of smooth alternatives, we also compute the needed separation from the null region to obtain high power. Our proposed test is similar in spirit to Salomond's [32] test in that both are based on the posterior probability of a slightly extended null region, but our use of the $\mathbb{L}_1$-metric on the function or the Hellinger metric on the density of Y, leads to the universal consistency.

The paper is organized as follows. In the next section, we formally introduce the modeling assumptions and the prior and describe the projection-posterior approach. In Section 3, we present results on posterior contraction rates of the projection- posterior distribution. In Section 4, we derive asymptotic properties of the proposed Bayesian tests. A simulation study assessing the accuracy of the proposed Bayesian estimator and tests with other Bayesian and non-Bayesian completing methods is presented in Section 5. Proofs of the main results are given in Section 6 and those of the auxiliary results in Section 7.

2. Model, prior and projection-posterior

The following notations will be used throughout the paper. Let $\boldsymbol{I}_m$ stand for the $m \times m$ identity matrix. By $\boldsymbol{Z} \sim \mathrm{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we mean that $\boldsymbol{Z}$ has a J-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For a vector $\boldsymbol{x}$, the Euclidean norm will be denoted by $\|\boldsymbol{x}\|$. The transpose of a vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^{\mathrm{T}}$ and that of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^{\mathrm{T}}$. If f is a function and H is a measure, the $\mathbb{L}_p$-norm of f is given by $\|f\|_{p,H} = (\int |f|^p dH)^{1/p}$ for $1 \leq p < \infty$, and the $\mathbb{L}_p$-distance between two functions f and g is given by $d_{p,H}(f,g) = \|f - g\|_{p,H}$ for $1 \leq p < \infty$ and $d_{p,H}(f,g) = \int |f-g|^p dH$ for $0 < p < 1$. The indicator function will be denoted by $\mathbb{1}$ and # will stand for the cardinality of a finite set.

For two sequences of real numbers a_n and b_n, $a_n \lesssim b_n$ means that a_n/b_n is bounded, $a_n \asymp b_n$ means that both $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ means that $a_n/b_n \to 0$. For a random variable X and a sequence of random variables X_n, $X_n \to_P X$ means that X_n converges to X in P-probability.

Let $\mathcal{F}$ and $\mathcal{F}_+$ respectively denote the space of real-valued measurable functions and monotone increasing functions on $[0,1]$, and for $K > 0$, let $\mathcal{F}_+(K) = \{f \in \mathcal{F}_+ : |f| \leq K\}$. For $f : [0,1] \mapsto \mathbb{R}$ and d a distance on $\mathcal{F}$, let the projection of f on $\mathcal{F}_+$ be the function f^* that minimizes $d(f,h)$ over $h \in \mathcal{F}_+$, provided such a minimizer exists. The projection need not be unique, in which case any choice may be taken. If a minimizer does not exist, a near minimizer may be used as is commonly adopted for M-estimation; see van der Vaart [35], page 45. However, in this paper, we need to use the projection map only on piece-wise constant functions with respect to nicely behaved metrics like $\mathbb{L}_1$ or $\mathbb{L}_2$, for which the projection exists and is unique. The topological closure of $\mathcal{F}_+$ is denoted by $\bar{\mathcal{F}}_+$. The ϵ-covering number of a set A with respect to a metric d, denoted by $\mathcal{N}(\epsilon, A, d)$, is the minimum number of balls of radius ϵ needed to cover A.

Let $G_n(x) = n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i \leq x\}$, the empirical distribution of the predictors X.

A prior distribution Π on the regression function f will be given by a random step function $f(x) = \sum_{j=1}^J \theta_j \mathbb{1}\{x \in I_j\}$, $x \in [0,1]$, where $I_1, \ldots, I_J$ are disjoint intervals partitioning $[0,1]$ given by $I_1 = [\xi_0, \xi_1]$ and $I_j = (\xi_{j-1}, \xi_j]$, $j = 2, \ldots, J$. The knot points are $0 = \xi_0 < \xi_1 < \ldots < \xi_{J-1} < \xi_J = 1$. With a given set of J knots, the corresponding collection of step functions is denoted by $\mathcal{F}_J$. The counts of these intervals are denoted by $N_j = \sum_{i=1}^n \mathbb{1}\{X_i \in I_j\}$, $j = 1, \ldots, J$. For the prior, J or $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{J-1})$ or both may be given, or these may be distributed according to a prior. Depending on their choices, the following three types of prior distributions will be considered in this paper.

1. **Type 1 prior**: The number of steps J is deterministic (but depends on the sample size n), $\xi_j = j/J$, $j = 1, \ldots, J-1$.
2. **Type 2 prior**: The number of steps J is deterministic,
$$\mathrm{P}((\xi_1, \ldots, \xi_{J-1}) = S) = \frac{1}{\binom{n}{J-1}},\ S \subset \{X_1, \ldots, X_n\}, \#S = J-1,$$
that is, the knots are sampled randomly without replacement from the observed values of the predictor variable (only applicable for a deterministic X with distinct values).
3. **Type 3 prior**: The knots are equidistant and the number of steps J is given a prior satisfying
$$\exp[-b_1 j (\log j)^{t_1}] \leq \Pi(J = j) \leq \exp[-b_2 j (\log j)^{t_2}] \qquad (2.1)$$
for some $b_1, b_2 > 0$ and $0 \leq t_2 \leq t_1 \leq 1$.

In all three cases, given σ and J, the coefficients $\theta_1, \ldots, \theta_j$ are given independent normal priors $\theta_j|\sigma \sim \mathrm{N}(\zeta_j, \sigma^2\lambda_j^2)$, $B_1 < \lambda_j < B_2$ for some $B_1, B_2 > 0$ and bounded $|\zeta_1|, \ldots, |\zeta_J|$. We write $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_J^2)$, the diagonal matrix with entries $\lambda_1^2, \ldots, \lambda_J^2$. The choices of these parameters are not important for asymptotic properties as long as the stated boundedness conditions are satisfied. In finite samples, the choices may make some impact though. If a prior guess $\bar{f}$ about the monotone regression is available, ζ_j may be taken as the average $\int_{I_j} \bar{f}(x) dG(x)$ of $\bar{f}$ over I_j, while λ_j indicates the lack of faith in the prior belief in ζ_j, $j = 1, \ldots, J$. More commonly, in the absence of any reliable prior information, ζ_j, $j = 1, \ldots, J$, may be set to 0 and λ_j to relatively large value, for a low-information diffuse prior. The Type 1 prior will be used to obtain optimal posterior contraction in $\mathbb{L}_1$-distance, Type 2 prior for posterior contraction in terms of an empirical $\mathbb{L}_2$-distance while Type 3 prior will be used for testing monotonicity against smooth alternatives of unspecified smoothness.

The variance parameter σ^2 is either estimated by maximizing the marginal likelihood, or is given an inverse-gamma prior $\sigma^2 \sim \mathrm{IG}(\beta_1, \beta_2)$ with $\beta_1 > 2$ and $\beta_2 > 0$.

We write $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$, $D_n = (\boldsymbol{Y}, \boldsymbol{X})$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, $\boldsymbol{B} = ((\mathbb{1}\{X_i \in I_j\}))$, an $n \times J$ matrix, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_j)^{\mathrm{T}}$. Thus the model can be written as $\boldsymbol{Y} = \boldsymbol{B\theta} + \boldsymbol{\varepsilon}$, and the prior (given J and σ) as $\boldsymbol{\theta}|(J, \sigma) \sim \mathrm{N}_J(\boldsymbol{\zeta}, \sigma^2 \boldsymbol{\Lambda})$ with $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_J)^{\mathrm{T}}$. Then we have (see, eg., Hoff [26], page 155) that

$$\boldsymbol{\theta}|(D_n, J, \sigma, \boldsymbol{\xi}) \sim \mathrm{N}_J((\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \boldsymbol{\Lambda})^{-1}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\zeta}), \sigma^2(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} + \boldsymbol{\Lambda}^{-1})^{-1}).$$

Since in our context $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \mathrm{diag}(N_1, \ldots, N_J)$, it follows that θ_j are a posteriori independent with

$$\theta_j|(\boldsymbol{\xi}, \sigma, J, D_n) \sim \mathrm{N}\left(\frac{N_j \bar{Y}_j + \zeta_j/\lambda_j^2}{N_j + 1/\lambda_j^2}, \frac{\sigma^2}{N_j + 1/\lambda_j^2}\right). \tag{2.2}$$

The marginal distribution of the observations $\boldsymbol{Y}$ (given $\boldsymbol{X}$ and $J, \sigma^2, \boldsymbol{\xi}$) is

$$\boldsymbol{Y}|(\sigma, \boldsymbol{\xi}, J, \boldsymbol{X}) \sim \mathrm{N}_n\big(\boldsymbol{B\zeta}, \sigma^2(\boldsymbol{B\Lambda B}^{\mathrm{T}} + \boldsymbol{I}_n)\big). \tag{2.3}$$

As the coefficients $\boldsymbol{\theta}$ have not been restricted to the cone of monotone increasing values $\mathcal{Q} := \{(q_1, \ldots, q_J) : q_1 \leq q_2 \leq \cdots \leq q_J\}$, the resulting regression function $f = \sum_{j=1}^J \theta_j \mathbb{1}_{I_j}$, sampled randomly from the posterior distribution $\Pi(\cdot|D_n)$, may not be monotone. In order to comply with the monotonicity restriction, a sampled value of the function f from its posterior (obtained through the posterior sampling of $\boldsymbol{\theta}$) is projected on the set of monotone functions $\mathcal{F}_+$ on $[0, 1]$ to obtain $f^* \in \mathcal{F}_+$ nearest to f with respect to some distance d. The induced distribution of f^* will be called the projection-posterior distribution. It will be denoted by Π_n^* and will be the basis for inference on the regression function f. By its definition, the projection-posterior distribution is restricted to $\mathcal{F}_+$.

We also find that the projection f^* of a step function $f = \sum_{j=1}^J \theta_j \mathbb{1}_{I_j} \in \mathcal{F}_J$ is itself a step function $f = \sum_{j=1}^J \theta_j^* \mathbb{1}_{I_j} \in \mathcal{F}_J$, with $\theta_1^* \leq \cdots \leq \theta_J^*$. For the the $\mathbb{L}_2(G_n)$-distance, these values are obtained by the weighted isotonization procedure

$$\text{minimize } \sum_{j=1}^J N_j(\theta_j - \theta_j^*)^2 \text{ subject to } \theta_1^* \leq \cdots \leq \theta_J^*. \tag{2.4}$$

The optimizing values $\theta_1^*, \ldots, \theta_J^*$ can be computed using the PAVA and can be characterized as the left-derivative at the point $n^{-1}\sum_{k=1}^j N_k$ of the greatest convex minorant of the graph of the line segments connecting the points

$\big\{(0,0), \big(N_1/n, N_1\theta_1/n\big), \dots, \big(\sum_{k=1}^{J} N_k/n, \sum_{k=1}^{J} N_k\theta_k/n\big)\big\}$ (cf. Lemma 2.1 of Groeneboom and Jongbloed [24]). The same solution is obtained even if the $\mathbb{L}_2(G_n)$-distance is replaced by a wider class; see Theorem 2.1 of Groeneboom and Jongbloed [24].

We make one of the following design assumptions (DD) or (DR) on the predictor X and the assumption (E) on the error variables.

Condition (DD) (Deterministic predictor). The predictor variables X is deterministic assuming values $X_1, \dots, X_n$, and the counts $N_1, \dots, N_J$ of J equispaced intervals $I_1, \dots, I_J$ satisfy, for $J \to \infty$, $\max\{N_j : 1 \le j \le J\}/n = O(J^{-1})$.

The bound is clearly implied by the condition $\sup\{|G_n(x) - G(x)| : x \in [0,1]\} = o(J^{-1})$, where G has a positive and continuous density g on $[0,1]$.

Condition (DR) (Random predictor). The predictor X is sampled independently from a distribution G, having a density g, which is bounded and bounded away from zero on $[0,1]$.

The assumption of normality on the error is only a working hypothesis. We only construct the likelihood function using the model $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\mathrm{N}(0,\sigma^2)$ with an unknown $\sigma > 0$. We assume the following condition on the true distribution of the error.

Condition (E) (True error distribution). The error variables $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. sub-Gaussian with mean 0 and variance σ_0^2.

We denote the true value of the regression function by f_0 and write the vector of function values at the observed points by $\boldsymbol{F}_0 = (f_0(X_1), \dots, f_0(X_n))$. The true value of the variance σ^2 is thus σ_0^2. We denote true distribution of (X,Y) by P_0. Let $\mathrm{E}_0(\cdot)$ and $\mathrm{Var}_0(\cdot)$ be the expectation and variance operators taken under the true distribution P_0.

The error variance σ^2 may be estimated by maximizing the marginal likelihood of σ. From (2.3), it follows that the marginal maximum likelihood estimate of σ^2 is given by

$$\hat{\sigma}_n^2 = n^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\zeta})^{\mathrm{T}}(\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\zeta}). \tag{2.5}$$

The plug-in posterior distribution of f is then obtained by substituting $\hat{\sigma}_n$ for σ in (2.2). If instead, we equip σ^2 with inverse-gamma prior $\sigma^2 \sim \mathrm{IG}(\beta_1, \beta_2)$, then a fully Bayes procedure can be based on the posterior distribution

$$\sigma^2 | D_n \sim \mathrm{IG}(\beta_1 + n/2, \beta_2 + (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\zeta})^{\mathrm{T}}(\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\zeta})/2); \tag{2.6}$$

see Hoff [26], page 155.

3. Posterior contraction rates under monotonicity

3.1. Preliminaries

To establish posterior contraction rates for f with unknown σ, we need to effectively control the range of values of σ.

It will be shown in Lemma 7.2 that the maximum marginal likelihood estimator for σ^2 in the plug-in Bayes approach and the marginal posterior distribution of σ^2 in the fully Bayes approach, are consistent for any $f_0 \in \mathcal{F}_+$, and the convergence is also uniform over $\mathcal{F}_+(K)$, for any fixed $K > 0$. This allows us to treat σ as essentially known in studying the posterior contraction.

As mentioned in the last section, we impose monotonicity on f by projecting f onto $\mathcal{F}_+$ and use the projection-posterior distribution for inference. The following argument shows that the concentration property of the posterior at any monotone function is not weakened by this procedure.

Let Π_n^* stand for the projection-posterior distribution given by

$$\Pi_n^*(B) = \Pi(f : f^* \in B | D_n), \quad B \subset \mathcal{F}_+, \tag{3.1}$$

where f^* is the projection of f on $\mathcal{F}_+$ with respect to some metric d on the space of regression functions. Then for the true regression function $f_0 \in \mathcal{F}_+$ and $\epsilon > 0$, we have the following *concentration inequality* for the projection-posterior distribution:

$$\Pi_n^*(d(f^*, f_0) > 2\epsilon) \leq \Pi(f : d(f, f_0) > \epsilon | D_n). \tag{3.2}$$

Consequently, the contraction rate of the unrestricted posterior is inherited by the projection-posterior, giving a path to the derivation of the posterior contraction of the projection-posterior distribution. To see this, note that $d(f^*, f) \leq d(f_0, f)$ by the property of the projection. Hence, using the triangle inequality

$$d(f^*, f_0) \leq d(f^*, f) + d(f, f_0) \leq d(f_0, f) + d(f, f_0) = 2d(f, f_0). \tag{3.3}$$

For $p \geq 1$, the $\mathbb{L}_p$-projection of a step function is easily computable, by algorithms similar to the PAVA (see Section 3.1 of De Leeuw et al. [19]).

It may be noted that the concentration inequality for the projection-posterior applies well beyond the specific prior based on piece-wise constant functions used in the paper. For instance, if the regression function is also known to be smoother, we can use higher order B-splines (piece-wise constant functions are linear combinations of order 1, while piece-wise linear functions are linear combinations of order 2 B-splines) to construct a prior distribution that has better concentration property at smoother functions; see Section 10.4 of Ghosal and van der Vaart [23]. The only additional complications are that the posterior distributions of the coefficients are dependent normal and the projection cannot be computed by the PAVA. Nevertheless, the optimal posterior contraction rate obtained at a smooth monotone function may be passed to the projection-posterior distribution by (3.2). However, if only monotonicity is assumed, piece-wise linear or higher order B-splines, although can be used to construct prior, may not be useful for obtaining the contraction rate, due to the lack of optimal approximation property of such functions at an arbitrary monotone function.

3.2. Contraction rates under the $\mathbb{L}_1$-metric

In this subsection, we derive the posterior contraction rate with respect to the $\mathbb{L}_1$-metric. An important factor determining this rate is the approximation rate of monotone functions by step functions. For the $\mathbb{L}_1$-metric, step functions with regularly placed knots are adequate for the optimal approximation rate (see Lemma 7.3), and hence it is sufficient to consider a Type 1 prior. In the following theorem, we derive the contraction rate at a monotone function in the $\mathbb{L}_1$-metric by directly bounding posterior moments.

Theorem 3.1. *Let $f_0 \in \mathcal{F}_+$, and assume that Condition* (E) *holds. Let the prior on f be of Type* 1*, with $J \to \infty$ and $J \ll n$. Let σ^2 be estimated using the plug-in Bayes approach or endowed with the inverse-gamma prior using a fully Bayes approach. Assume that either X is deterministic and Condition* (DD) *holds, or X is random and Condition* (DR) *holds. Then for $\epsilon_n = \max\{J^{-1}, (J/n)^{1/2}\}$ and every $M_n \to \infty$,*

(a) $\mathrm{E}_0\, \Pi_n^* \left(\|f - f_0\|_{1,G_n} > M_n\epsilon_n\right) \to 0$ *for the fixed design;*
(b) $\mathrm{E}_0\, \Pi_n^* \left(\|f - f_0\|_{1,G} > M_n\epsilon_n\right) \to 0$ *for the random design.*

In particular, if we choose $J \asymp n^{1/3}$, the projection-posterior contracts at the minimax rate $\epsilon_n = n^{-1/3}$. Moreover, the convergence is uniform over $\mathcal{F}_+(K)$ for any $K > 0$.

Under Condition (DR), the $\mathbb{L}_1(G)$-distance is equivalent to the usual Lebesgue $\mathbb{L}_1$-metric on $[0, 1]$, and hence the contraction rate may be stated in terms of the latter. Conditions (DD) or (DR) on X in the theorem above is needed only to conclude, using Lemma 7.2, that the estimator (or the posterior) for σ is consistent. The conclusion is only used to get an upper bound for σ. If instead, we assume an upper bound for σ (and change the prior on σ to comply with the bound, if the fully Bayes procedure is used), we can remove these conditions.

3.3. Contraction rates under the empirical $\mathbb{L}_p$*-metric*

When the metric under consideration is $\mathbb{L}_p$ with $p > 1$, step functions based on equidistant knots do not have the optimal approximation property. To restore this ability, we need to allow arbitrary knots (see Lemma 7.3), and put a prior on these. Then the theory of posterior contraction for general (independent, not identically distributed) observations of Ghosal and van der Vaart [22] can be applied by computing the prior concentration rate near the truth and bounding the metric entropy of a suitable subset of the parameter space, called a sieve. However, due to their ordering requirement and possibly very uneven allocation of the knots $\boldsymbol{\xi}$ used for the construction of the optimal approximation, the concentration of the prior distribution of $\boldsymbol{\xi}$ near their values appearing in the optimal approximation may be low, and hence the posterior concentration rate may suffer. The problem can be avoided by choosing knots from the observed values of X when the predictor variable is deterministic and the empirical $\mathbb{L}_p$-norm $\|f\|_{p,G_n}$ is used. Then the optimal rate (up to a logarithmic factor) can be obtained.

Theorem 3.2. *Let X be deterministic assuming values $X_1, \dots, X_n$. Let $f_0 \in \mathcal{F}_+$ and the prior on f be of Type* 2*, with $\log J \asymp \log n$. Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. normal with mean zero and variance σ^2, which is estimated using the plug-in*

Bayes approach or is endowed with the inverse-gamma prior using a fully Bayes approach. Then for any $0 < p \leq 2$, $\mathrm{E}_0\ \Pi_n^*\left(\|f - f_0\|_{p,G_n} > M_n\epsilon_n\right) \to 0$, *where* $\epsilon_n = \max\{\sqrt{(J\log n)/n}, J^{-1}\}$. *In particular, the best rate* $\epsilon_n = (n/\log n)^{-1/3}$ *is obtained by choosing* $J \asymp (n/\log n)^{1/3}$. *Moreover, the convergence is uniform over* $\mathcal{F}_+(K)$ *for any* $K > 0$.

If, instead of choosing J*, we put a prior also on* J *following* (2.1)*, then the contraction rate is given by* $n^{-1/3}(\log n)^{(5-3t_2)/6}$.

Clearly, with a prior on J given by (2.1), the best rate $(n/\log n)^{-1/3}$ is obtained when $t_1 = t_2 = 1$. A Poisson (or a suitably truncated Poisson) prior meets the requirement. Again, Condition (DD) is used only to derive the consistency of the estimator (or the posterior) of σ, and the condition can be removed if σ is assumed to be bounded.

It would be interesting to obtain nearly optimal contraction rates for the continuous $\mathbb{L}_p$-metric, but we do not know an appropriate prior on the knot-locations that would allow sufficient prior concentration to yield the desired result. For the continuous metric $\mathbb{L}_p$-metric, the weak approximation with equal intervals allows only a sub-optimal approximation rate $J^{-1/p}$ (see Lemma 7.3), and consequently a suboptimal posterior contraction rate $(n/\log n)^{-1/(p+2)}$.

4. Bayesian testing for monotonicity of *f*

A natural test for the hypothesis of monotonicity is given by the posterior probability of $\mathcal{F}_+$: reject the hypothesis if $\Pi(f \in \mathcal{F}_+|D_n)$ is smaller than 1/2, say. The problem with this test is that for a true regression $f_0 \in \mathcal{F}_+$, even though the posterior is consistent at f_0, the posterior probability $\Pi(f \in \mathcal{F}_+|D_n)$ may be low because a large part of a neighborhood of f_0 may fall outside $\mathcal{F}_+$. In order to avoid such false rejections, one may quantify a test based on a discrepancy measure $d(f, \mathcal{F}_+)$ between f sampled from the posterior, and the set of monotone functions $\mathcal{F}_+$ (that is, a nonnegative function f that vanishes exactly on $\mathcal{F}_+$), or equivalently, based on $d(f, f^*)$, where f^* is the projection of f on $\mathcal{F}_+$. A reasonable test can be determined by the posterior probability $\Pi(f : d(f, \mathcal{F}_+) < \tau_n|D_n)$ for a sequence $\tau_n \to 0$ slowly. In other words, we reject for low values of the posterior probability $\Pi(\mathcal{F}_+^{\tau_n}|D_n)$ of the τ_n-neighborhood $\mathcal{F}_+^{\tau_n} = \{f : d(f, \mathcal{F}_+) < \tau_n\}$ of $\mathcal{F}_+$. This approach was also pursued by Salomond [30, 32], with a discrepancy measure given by $d(f, \mathcal{F}_+) = \max\{(\theta_j - \theta_i) : 1 \leq j \leq i \leq J\}$ for $f = \sum_{j=1}^J \theta_j \mathbb{1}_{I_j}$ (with equidistant knots) and a cut-off $\tau_n = \sqrt{(J\log n)/n}$. This test has the probability of Type I error going to zero and has high power against smooth alternatives, if appropriately separated from the null. However, the power of this test at a non-smooth alternative may not go to one. This prompts us to propose an alternative test, based on the $\mathbb{L}_1$-distance as the discrepancy measure, which has the property of universal consistency, that is, the power at any fixed alternative goes to one.

Let $\mathcal{H}(\alpha, L)$ be the Hölder space of α-smooth function with Hölder norm bounded by L (see Definition C.4 of Ghosal and van der Vaart [23]).

Theorem 4.1. *Consider a Type* 1 *prior with* $J \asymp n^{1/3}$. *Let* σ^2 *be estimated using the plug-in Bayes approach or endowed with the inverse-gamma prior using a fully Bayes approach. Assume that* X *is random and Condition* (DR) *holds, and the errors satisfy Condition* (E). *For the discrepancy measure* $d(f, \mathcal{F}_+) = \inf\{\|f-h\|_{1,G} : h \in \mathcal{F}_+\}$, *consider a test* $\phi_n = \mathbb{1}\{\Pi(d(f, \mathcal{F}_+) \leq M_n n^{-1/3} | D_n) < \gamma\}$, *where* $0 < \gamma < 1$ *is a predetermined constant and* $M_n \to \infty$ *is fixed slowly growing sequence. Then the following assertions hold.*

(a) (*Consistency under* H_0) : *For any fixed* $f_0 \in \mathcal{F}_+$, $\mathrm{E}_0 \phi_n \to 0$, *and further the convergence is uniform over* $\mathcal{F}_+(K)$.

(b) (*Universal Consistency*) : *For any fixed* f_0 *integrable on* $[0,1]$ *and* $f_0 \notin \bar{\mathcal{F}}_+$, $\mathrm{E}_0(1-\phi_n) \to 0$.

(c) (*High power at converging smooth alternatives*) : *For any* $0 < \alpha \leq 1$ *and* $L > 0$, $\sup\{\mathrm{E}_0(1-\phi_n) : f_0 \in \mathcal{H}(\alpha, L), d(f_0, \mathcal{F}_+) > \rho_n(\alpha)\} \to 0$, *where*

$$\rho_n(\alpha) = \begin{cases} C n^{-\alpha/3}, & \textit{for some } C > 0 \textit{ if } \alpha < 1, \\ C M_n n^{-1/3}, & \textit{for any } C > 1 \textit{ if } \alpha = 1. \end{cases}$$

In the above theorem, the $\mathbb{L}_1(G)$-distance may be replaced by the $\mathbb{L}_1$-distance under the Lebesgue measure, since under Condition (DR), these two metrics are equivalent. In this case, part (c) may be strengthened by replacing the Hölder space $\mathcal{H}(\alpha, L)$ by the Sobolev space with $(1, \alpha)$-Sobolev norm bounded by L (see Definition C.6 of Ghosal and van der Vaart [23]). Also, if G is replaced by the empirical distribution G_n (and assuming that Condition (DD) holds instead of Condition (DR) if X is deterministic), the conclusions in parts (a) and (c) will still hold. The proof is very similar. If σ has a known bound, then Condition (DD) or Condition (DR) is not needed.

The procedure involving the test ϕ_n is computationally simple as it does not involve a prior on J. The algorithm for median isotonic regression (see Robertson and Wright [29] and De Leeuw et al. [19]) allows us to compute $d(f, \mathcal{F}_+)$ very efficiently. However, with a deterministic choice of J, the posterior contraction is not adaptive on classes of functions with different smoothness α, where the optimal order is $n^{1/(1+2\alpha)}$. This is caused by the deterministic choice of $J \asymp n^{1/3}$ needed for the optimal rate at a monotone function, which differs from the optimal order of $n^{1/(1+2\alpha)}$ needed to match the minimax optimal rate $n^{-\alpha/(1+2\alpha)}$ (up to a logarithmic factor) at a function in $\mathcal{H}(\alpha, L)$. This has a consequence on the degree of separation needed for high asymptotic power at an alternative regression function $f_0 \in \mathcal{H}(\alpha, L)$ in testing for monotonicity. In order to make the power go to one, an $n^{-\alpha/3}$-sized neighborhood of f_0 should not contain a part of $\mathcal{F}_+$, that is, the order of separation from f_0 to $\mathcal{F}_+$ for high power needs to be at least $n^{-\alpha/3}$ in the above result. In other words, the order of separation $n^{-\alpha/3}$ (up to a logarithmic factor) between f_0 and $\mathcal{F}_+$ is needed. However, this is more than what should be ideally needed by a testing procedure. Since the contraction rate at a function in $\mathcal{H}(\alpha, L)$ is $n^{-\alpha/(1+2\alpha)}$ achievable by choosing $J \asymp n^{-\alpha/(1+2\alpha)}$, in order to make the power go to one, only an $n^{-\alpha/(1+2\alpha)}$-size neighborhood should be disjoint from $\mathcal{F}_+$. Thus the minimal order of separation from f_0 to $\mathcal{F}_+$ needed for high power by any testing procedure is $n^{-\alpha/(1+2\alpha)}$, or that, $n^{-\alpha/(1+2\alpha)}$ is the optimal order of separation. As the separation needed in the above result is larger than the minimal order except for $\alpha = 1$, the resulting test is not rate-adaptive. Adaptation can, however, be restored in a class of uniformly bounded regression functions by using a prior on J, and letting cut-off value for the discrepancy with $\mathcal{F}_+$ depend on J. The idea is similar to the one used in Salomond [32], except that we use the Hellinger distance between the underlying densities, instead of the maximum discrepancy measure on the coefficients used by him. This allows us to retain the universal consistency property.

Let $d_H(f_1, f_2)$ stand for the Hellinger distance between $p_{f_1,\sigma}$ and $p_{f_2,\sigma}$, where $p_{f,\sigma}$ stands for the joint density of (X, Y) following $Y|X \sim \mathrm{N}(f(X), \sigma^2)$ and $X \sim G$, with respect to the measure the product of G and the Lebesgue measure. By an easy calculation,

$$d_H^2(f_1, f_2) = 2\Big[1 - \int (\sigma\sqrt{2\pi})^{-1/2} \exp\{-(f_1(x) - f_2(x))^2/(8\sigma^2)\}\Big] g(x)dx.$$

It follows that $d_H(f_1, f_2) \lesssim \|f_1 - f_2\|_{2,G}$ and the reverse inequality also holds if f_1 and f_2 belong to a uniformly bounded class.

Theorem 4.2. *Let the prior on f be of Type* 3 *with J given a Poisson prior, and σ be bounded and be given a positive prior density with bounded support containing the true value σ_0. Assume that $X \sim G$ and G satisfies Condition* (DR). *Let $\phi_n = \mathbb{1}\{\Pi(d_H(f, \mathcal{F}_+) \le M_0\sqrt{(J \log n)/n}|D_n) < \gamma\}$, $0 < \gamma < 1$, be a predetermined constant and $M_0 > 0$ be a sufficiently large constant.*

(a) (*Consistency under H_0*) : *For any fixed $f_0 \in \mathcal{F}_+$, $\mathrm{E}_0\phi_n \to 0$, and the convergence is uniform over $\mathcal{F}_+(K)$.*
(b) (*Universal Consistency*) : *For any fixed f_0 integrable on* $[0, 1]$ *and $f_0 \notin \bar{\mathcal{F}}_+$,* $\mathrm{E}_0(1 - \phi_n) \to 0$.
(c) (*Adaptive power at converging smooth alternatives*) : *For $f_0 \notin \mathcal{F}_+$, $f_0 \in \mathcal{H}(\alpha, L)$, there exists C depending on α and L only such that*

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

In the theorem, G can be replaced by the uniform distribution in the definition of the test. In this case, the Hölder space $\mathcal{H}(\alpha, L)$ in part (c) can be replaced by the Sobolev space with $(2, \alpha)$-Sobolev norm bounded by L.

Unlike Theorem 4.1, the proof requires the application of the general theory of posterior contraction. The weaker Hellinger distance for separation is used so that a test required for the application of the theory is available automatically without requiring the regression functions to be bounded by a known constant, a condition that will rule out the conjugate normal prior needed in the proof. An alternative is to use the empirical $\mathbb{L}_1$-distance and conclude parts (a) and (c) only, assuming that $N_j \asymp n/J$ uniformly in $j = 1, \ldots, J$.

5. Simulation study

In this section, we compare the proposed method with some other Bayesian and non-Bayesian methods. For estimation, we compare the $\mathbb{L}_1$-distance of the Bayes estimator from the true function, with the $\mathbb{L}_1$-distance of other estimators from the true function. For testing, we compare the level and power of the Bayesian test with those of the competing tests for monotonicity.

5.1. *Simulation for contraction rates*

We perform a simulation study to demonstrate the behavior of our projection-posterior method of estimation in finite samples. We use the Bayesian algorithm for estimating monotone functions given by Bornkamp and Ickstadt [10], and the median isotonic regression estimator as benchmarks to compare our results with. We consider the following monotone functions on $[0,1]$:

1. $f_1(x) = 0$,
2. $f_2(x) = x^2 + x/5$,
3. $f_3(x) = \begin{cases} 0, & x \le 0.6 \\ 1, & x > 0.6, \end{cases}$
4. $f_4(x) = (\mathrm{Be}(x,1,1) + \mathrm{Be}(x,200,80) + \mathrm{Be}(x,80,200))/3$, where $\mathrm{Be}(x,a,b)$ is the distribution function of a $\mathrm{Beta}(a,b)$ random variable evaluated at x.

For every true regression function, we generate samples of size n from the model $Y = f(X) + \varepsilon$, with X as points equally distributed on $(0,1)$, ε generated from $\mathrm{N}(0,\sigma^2)$, with $\sigma = 0.1$. We take the number of pieces J as the greatest integer less than or equal to $n^{1/3}\log n$. We generate 1000 posterior samples for each dataset, project them onto the monotone class of functions and use the mean of projection-posterior samples as our estimator.

We compare the performance of our method with that of the median isotonic estimator and the algorithm by Bornkamp and Ickstadt [10]. The R package “isotone” efficiently computes the $\mathbb{L}_1$-projection of a step function on $\mathcal{F}_+$. We use 5000 burn-in samples and 20000 MCMC samples to estimate f using the method of Bornkamp and Ickstadt [10], implemented in the R package “bnpmr”.

To evaluate the performance of an estimator f, we find the $\mathbb{L}_1$-distance between f and the true function f_0 on a fine grid, calculated as

$$D(f,f_0) = \frac{1}{K}\sum_{k=1}^{K} \big|f(k/K) - f_0(k/K)\big|, \quad K = 100.$$

We report the results in Table 1. Each entry is $D(f,f_0)$ averaged over 1000 replications, and the standard deviation across these replications is reported in the bracket. We use “projection”, “BNPMR” and “isotone” respectively to denote the estimators obtained from our method, Bornkamp and Ickstadt [10], and median isotonic regression. We observe that while no single method performs better than others in all the cases considered, the projection-posterior method

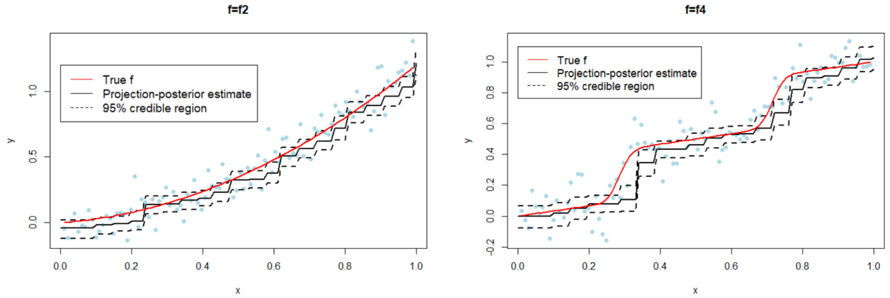

Fig 1. *Scatterplot,* 95% *projection-posterior credible interval, projection-posterior estimate and true function, displayed for one dataset generated from models with* $f = f_2$ *and* $f = f_4$*.*

seems to do better than Bornkamp and Ickstadt [10] for $f = f_2$ when $n = 100$ and $f = f_4$. For $f = f_1$, projection-posterior and isotonic regression give very similar results. In the experimental setup of Table 1, the approximate computing times needed to generate the posterior samples from one dataset of size 100 were 0.57 seconds for projection-posterior, and 1 second for Bornkamp and Ickstadt [10]. This is expected, as our method does not involve drawing MCMC samples and is therefore computationally simpler than MCMC-based Bayesian methods.

Table 1

Comparison of the $\mathbb{L}_1$*-distance of the true function from the proposed estimate. Each entry is the error averaged over* 1000 *replications and the number in bracket is the standard deviation across the replications.*

	$n = 100$			$n = 200$		
	projection	**BNPMR**	**isotone**	**projection**	**BNPMR**	**isotone**
f_1	0.019(0.007)	0.011(0.006)	0.019(0.008)	0.013(0.005)	0.008(0.004)	0.013(0.006)
f_2	0.059(0.008)	0.065(0.007)	0.036(0.005)	0.042(0.007)	0.018(0.019)	0.029(0.004)
f_3	0.065(0.006)	0.020(0.006)	0.024(0.007)	0.045(0.004)	0.016(0.004)	0.018(0.006)
f_4	0.056(0.008)	0.068(0.002)	0.033(0.007)	0.039(0.005)	0.067(0.001)	0.026(0.004)

We display the scatter-plot of a dataset of size 100 generated from $f = f_2$ and $f = f_4$ with $\sigma = 0.1$, along with our estimator and a 95% projection-posterior credible region in Figure 1. The estimator is seen to approximate the flat parts of f_4 and most parts of f_2 quite well.

5.2. Simulation for testing monotonicity

In this section we report the results from a simulation study comparing the performance of our test for monotonicity with that of two other methods. The first method is the one by Salomond [32], and the second is a non-Bayesian test proposed by Ghosal et. al. [21].

We generate samples of size n from the model $Y_i = f(X_i) + \varepsilon_i$, with X_i's equally spaced on $(0, 1)$. The errors ε_i are i.i.d. $\mathrm{N}(0, 0.1^2)$. We generate 500 such datasets for each test function f on $[0, 1]$. We choose three monotone functions: $m_1(x) = 0$, $m_2(x) = 0.2\mathbb{1}_{\{x>0.6\}}$, $m_3(x) = (\text{Be}(x, 1, 1) + \text{Be}(x, 200, 80) + \text{Be}(x, 80, 200))/3$. For non-monotone functions, we choose f as $m_4(x) = -0.1x$, $m_5(x) = -0.1 \exp\{-50(x - 0.5)^2\}$, $m_6(x) = 0.1 \cos(6\pi x)$, $m_7 = 0.2x + m_6(x)$, $m_8(x) = x + 0.415 \exp(-50x^2)$, $m_9(x) = x + 1 - 0.45 \exp\{-50(x - 0.5)^2\}$. The functions m_4 to m_8 have been considered by Scott [33], Salomond [32] and Ghosal et. al. [21] as examples of non-monotone functions with small departures from monotonicity. The test by Salomond [32] is for monotone decreasing functions, so we use his algorithm on the negative of our simulated Y-values.

We choose the number of knots J as the greatest integer less than or equal to $n^{1/3}$. The hyperparameters are chosen as $\zeta_j = 0$ and $\lambda_j^2 = 100$ for all $1 \le j \le J$. For the error variance, we use the marginal maximum likelihood estimate of σ^2. For our test and that of Salomond's [32], we use 1000 posterior samples for each dataset to make inference. The γ in Theorem 4.1 is chosen as 1/2.

To determine an appropriate cutoff point for our test, we find a slowly growing sequence M_n such that using $M_n n^{-1/3}$ as the cutoff in Theorem 4.1 results in low Type 1 error for $f = 0$. We let $M_n = M_0(\log n)^\kappa$ for $M_0, \kappa > 0$. We run our test across different combinations of (M_0, κ) on datasets generated from the model with $f = 0$ for several values of n, ranging from $n = 50$ to $n = 10000$, and choose the combinations that result in the misclassification error rate being less than 0.10. We then select the combination that maximizes the power when the test is used on datasets generated from non-monotone functions. From this analysis, we found $M_n = 0.8(\log n)^{0.1}$ to be an appropriate candidate for M_n.

The Type 1 errors of the tests are presented in Table 2. The error rate corresponding to m_1 gives us an idea about the level of the test. The power of the tests are displayed in Table 3. We find that our test has low Type 1 error rate and it outperforms Salomond's [32] test in samples of sizes 50 and 100 in datasets generated from monotone functions. In terms of power, our test does slightly worse than the other tests in small and moderate sample sizes. However, as n grows, the power approaches one.

The computing time for our procedure was found to be quite reasonable. We evaluated $d(f, \mathcal{F}_+)$ using the `gpava` function in the R package "isotone". For a dataset with $n = 500$, using 2500 posterior draws, our test took 0.75 seconds to execute. On the same dataset, Salomond's [32] test took 2.03 seconds to run on the same processor, for 2500 posterior samples. As mentioned before, our test is computationally simpler than Salomond's [32] as it does not draw posterior samples of J, which is possibly the reason for the faster computational time.

6. Proofs of the main results

Proof of Theorem 3.1. In view of (3.2), it is enough to obtain the contraction rate of the unrestricted posterior. We prove the result for the plug-in Bayes approach; the fully Bayes case can be dealt with similarly. From Lemma 7.2, get a shrinking neighborhood $\mathcal{U}_n$ of σ_0 with $P_0(\hat{\sigma}_n \in \mathcal{U}_n) \to 1$. Hence for the purpose of the proof, we may assume that $\hat{\sigma}_n \in \mathcal{U}_n$.

We first consider the case that X is deterministic. Let $f_{0J} = \sum_{j=1}^{J} \theta_{0j} \mathbb{1}_{I_j}$ with $\theta_{0j} = N_j^{-1} \sum_{i:X_i \in I_j} f_0(X_i)$ for all $1 \le j \le J$. By Lemma 7.3 (a), we

TABLE 2
Type 1 *error: the proportion of instances out of* 500 *when* H_0 *was rejected. Our test based on the* $\mathbb{L}_1$*-distance, the test by Salomond [32] and that of Ghosal et al. [21]*
are denoted by T_1, T_∞ and T_G respectively.

	$n = 50$			$n = 100$			$n = 200$			$n = 500$		
f	T_1	T_∞	T_G	T_1	T_∞	T_G	T_1	T_∞	T_G	T_1	T_∞	T_G
m_1	0.070	0.180	0.020	0.082	0.066	0.032	0.060	0.038	0.035	0.072	0.006	0.038
m_2	0.000	0.358	0.086	0.002	0.200	0.069	0.016	0.058	0.078	0.004	0.014	0.031
m_3	0.000	0.468	0.104	0.000	0.264	0.086	0.000	0.146	0.068	0.000	0.100	0.058

TABLE 3
Power: the proportion of instances out of 500 *when* H_0 *was rejected. Our test based on the* $\mathbb{L}_1$*-distance, the test by Salomond [32] and that of Ghosal et al. [21]*
are denoted by T_1, T_∞ and T_G respectively.

	$n = 200$			$n = 500$			$n = 600$			$n = 700$		
f	T_1	T_∞	T_G	T_1	T_∞	T_G	T_1	T_∞	T_G	T_1	T_∞	T_G
m_4	0.982	0.929	0.962	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
m_5	0.816	0.965	0.967	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
m_6	0.950	1.000	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000	1.000
m_7	0.232	1.000	0.898	1.000	1.000	0.994	1.000	1.000	1.000	1.000	1.000	1.000
m_8	0.004	1.000	0.936	0.858	1.000	0.995	0.934	1.000	1.000	0.964	1.000	1.000
m_9	0.334	1.000	0.903	0.858	1.000	0.979	0.980	1.000	1.000	1.000	1.000	1.000

have $\|f_{0J} - f_0\|_{1,G_n} \lesssim J^{-1}$ and the bound is also uniform for $f_0 \in \mathcal{F}_+(K)$. To complete the proof, we now show that

$$\mathrm{E}_0\Pi(\|f - f_{0J}\|_{1,G_n} > M_n\sqrt{J/n}\big|D_n) \to 0 \text{ for any } M_n \to \infty. \tag{6.1}$$

Since $f = \theta_j$ and $f_0 = \theta_{0j}$ on I_j, $\|f - f_{0J}\|_{1,G_n} = n^{-1}\sum_{j=1}^J N_j|\theta_j - \theta_{0j}|$. Hence by the Cauchy-Schwarz inequality followed by Markov's inequality,

$$\Pi(\|f - f_{0J}\|_{1,G_n} > M_n\sqrt{J/n}\big|D_n) \lesssim \frac{1}{M_n^2 J}\sum_{j=1}^J N_j \mathrm{E}(|\theta_j - \theta_{0j}|^2\big|D_n). \tag{6.2}$$

For $1 \le j \le J$, we bound $\mathrm{E}(|\theta_j - \theta_{0j}|^2\big|D_n) = \mathrm{Var}(\theta_j\big|D_n) + |\mathrm{E}(\theta_j\big|D_n) - \theta_{0j}|^2$, bound the expectation of both terms, and put in (6.2) to obtain the desired result. For the first term,

$$N_j\mathrm{Var}(\theta_j\big|D_n) \le \sup_{\sigma\in\mathcal{U}_n} \frac{N_j\sigma^2}{[N_j + \lambda_j^{-2}]} \lesssim 1. \tag{6.3}$$

We bound $\mathrm{E}_0[N_j|\mathrm{E}(\theta_j\big|D_n) - \theta_{0j}|^2]$ as

$$\begin{aligned}\mathrm{E}_0&\left[N_j\left|\frac{N_j\bar{Y}_j + \zeta_j/\lambda_j^2}{N_j + 1/\lambda_j^2} - \frac{\sum_{i:X_i\in I_j} f_0(X_i)}{N_j}\right|^2\right] \\ &\lesssim 1 + \mathrm{E}_0\left[N_j^{-1}\left|\sum_{i:X_i\in I_j}(Y_i - f_0(X_i))\right|^2\right].\end{aligned}$$

This follows because, by the boundedness of ζ_j and λ_j^{-2}, $|\mathrm{E}(\theta_j|D_n) - \bar{Y}_j| \lesssim N_j^{-1}$. The second term in the last expression is σ_0^2, and hence the expression is bounded above by a constant.

For random predictors, we use the $\|\cdot\|_{1,G}$-distance, which involves another integration with respect to $X_1, \dots, X_n$ on the left side of (6.1). □

Proof of Theorem 3.2. Because of (3.2), it suffices to obtain the contraction rate of the unrestricted posterior. Since for $0 < p < 2$, the $\mathbb{L}_p(G_n)$-distance is dominated by the $\mathbb{L}_2(G_n)$-distance, it suffices to prove the result for $p = 2$. We shall apply the general theory of posterior contraction (Ghosal and van der Vaart [23], Chapter 8) using the sieve

$$\mathcal{P}_n = \big\{f = \sum_{j=1}^{J} \theta_j \mathbb{1}_{[\xi_{j-1},\xi_j)},\ \xi_1, \dots, \xi_{J-1} \in \boldsymbol{X}, \max_j |\theta_j| \le n\big\}. \tag{6.4}$$

Let $p_{f,\sigma}^{(n)}$ denote the joint density of $Y_1, \dots, Y_n$ for a regression function f. We verify the conditions of Theorem 8.26 of Ghosal and van der Vaart [23] for $\epsilon_n = \max\{\sqrt{(J \log n)/n}, J^{-1}\}$. Note that by Lemma 7.2, we can restrict σ to an arbitrarily small neighborhood of σ_0, so the test construction in Lemma 8.27 of Ghosal and van der Vaart [23] is applicable.

By direct calculations, the Kullback-Leibler divergence and the squared Kullback-Leibler variation are respectively equal to

$$K(p_{f_0,\sigma_0}^{(n)}; p_{f,\sigma}^{(n)}) = \mathrm{E}_0\Big(\log \frac{p_{f_0,\sigma_0}^{(n)}}{p_{f,\sigma}^{(n)}}\Big) = \frac{n}{2\sigma^2}\|f - f_0\|_{2,G_n}^2 + \frac{n}{2}\big[\frac{\sigma_0^2}{\sigma^2} - 1 - \log\frac{\sigma_0^2}{\sigma^2}\big],$$

$$V_{2,0}(p_{f_0,\sigma_0}^{(n)}; p_{f,\sigma}^{(n)}) = \mathrm{Var}_0\Big(\log \frac{p_{f_0,\sigma_0}^{(n)}}{p_{f,\sigma}^{(n)}}\Big) = \frac{n}{4}\big(\frac{\sigma_0^2}{\sigma^2} - 1\big)^2 + \frac{n\sigma_0^2}{\sigma^4}\|f - f_0\|_{2,G_n}^2.$$

Therefore for a sufficiently small ϵ, there exists $C_1 > 0$ such that

$$\begin{aligned} B_{n,0}((f_0,\sigma_0),\epsilon) &:= \{(f,\sigma) : K(p_{f_0,\sigma_0}^{(n)}, p_{f,\sigma}^{(n)}) \le n\epsilon^2, V_{2,0}(p_{f_0,\sigma_0}^{(n)}; p_{f,\sigma}^{(n)}) \le n\epsilon^2\} \\ &\supset \{(f,\sigma) : \|f - f_0\|_{2,G_n}^2 \le C_1\epsilon^2, |\sigma^2 - \sigma_0^2|^2 \le C_1\epsilon^2\}. \end{aligned}$$

By Lemma 7.3, there exists f_{0J} such that $f_{0J}(\cdot) = \sum_{j=1}^{J} \theta_{0j}\mathbb{1}_{I_j}$, where $I_1, \dots, I_J$ are an interval partition with knots $\{\xi_{0,1}, \dots, \xi_{0,J-1}\} \subset \{X_1, \dots, X_n\}$ and $\|f_{0J} - f_0\|_{2,G_n}^2 \lesssim \epsilon_n^2$. By the prior independence of f and σ, and because $-\log \Pi(|\sigma - \sigma_0|^2 \le C\epsilon_n^2) \lesssim \log(1/\epsilon_n) \lesssim \log n$, it suffices that

$$\begin{aligned} \Pi(\|f - f_{0J}\|_{2,G_n}^2 \le C_2\epsilon_n^2) &= \Pi\big(\sum_{i=1}^{n} N_j(\theta_j - \theta_{0j})^2 \le C_2 n\epsilon_n^2 | \boldsymbol{\xi} = \boldsymbol{\xi}_0\big)\Pi(\boldsymbol{\xi} = \boldsymbol{\xi}_0) \\ &\ge \Pi\big(\bigcap_{j=1}^{J} \{|\theta_j - \theta_{0,j}| \le \sqrt{C_2}\epsilon_n\}\big)\frac{1}{\binom{n}{J-1}}, \end{aligned}$$

since $\sum_{i=1}^n (f(X_i) - f_{0J}(X_i))^2 = \sum_{j=1}^J N_j |\theta_j - \theta_{0j}|^2$ and $\sum_{j=1}^J N_j = n$. The last expression is at least of the order $(C_3\epsilon_n)^J\ n^{-(J-1)}$ for some $C_3 > 0$. Putting these together, we have $-\log \Pi(B_{n,0}((f_0,\sigma_0),\epsilon_n)) \lesssim J[\log(1/\epsilon_n) + \log J] \lesssim J \log n \lesssim n\epsilon_n^2$ by the definition of ϵ_n, fullfilling the condition of prior probability concentration needed for posterior contraction rate ϵ_n.

Observe that the metric entropy $\log \mathcal{N}(\epsilon, \mathcal{P}_n, \|\cdot\|_{p,G_n})$ of the sieve $\mathcal{P}_n$ in (6.4) is bounded above by $J\log(n/\epsilon_n) \lesssim J \log n \lesssim n\epsilon_n^2$. Finally, the prior probability $\Pi(\mathcal{P}_n^c)$ of the complement of the sieve $\mathcal{P}_n$ is bounded by $Je^{-n^2/2} \ll e^{-cn\epsilon_n^2}$ for any $c > 0$, establishing the condition (8.33) of Ghosal and van der Vaart [23]. This leads to the rate $\epsilon_n = \max(\sqrt{(J\log n)/n}, J^{-1})$ when J is chosen deterministically. Clearly, the best choice is $J \asymp (n/\log n)^{1/3}$, giving the nearly optimal rate $(n/\log n)^{-1/3}$.

When J is given a prior, to lower bound $\Pi(B_{n,0}((f_0,\sigma_0),\epsilon))$, we intersect the set with $\{J = J_0\}$, where $J_0 \asymp (n/\log n)^{1/3}$. This gives an additional factor $e^{-b_1 J_0 (\log J_0)^{t_1}}$, which is absorbed in $e^{-cn\bar\epsilon_n^2}$ by adjusting the constant for a pre-rate $\bar\epsilon_n = (n/\log n)^{-1/3}$, because $t_1 \le 1$. Modify the sieve in (6.4) by intersecting with $\{J \le J_1\}$, where J_1 is to be determined. The prior probability of the complement $\mathcal{P}_n^c$ then contributes an extra factor a constant multiple of $e^{-b_2 J_1 (\log J_1)^{t_2}}$ to $J_1 e^{-n^2/2}$. To obtain the final rate, we need to choose J_1 such that $J_1(\log n)^{t_2}$ exceeds a sufficiently large multiple of $n\bar\epsilon_n^2$, and then the rate is given by $\sqrt{(J_1 \log n)/n} = n^{-1/3}(\log n)^{(5-3t_2)/6}$. □

Proof of Theorem 4.1. (a) Let $f_0 \in \mathcal{F}_+$. Using the definition of the projection,

$$\mathrm{E}_0\Pi(\|f - f^*\|_{1,G} > M_n n^{-1/3} | D_n) \le \mathrm{E}_0\Pi(\|f - f_0\|_{1,G} > M_n n^{-1/3} | D_n) \to 0$$

for $J \asymp n^{1/3}$ by Theorem 3.1. Then it follows that $\mathrm{E}_0\phi_n = \mathrm{P}_0(\Pi(d(f,\mathcal{F}_+) \le M_n n^{-1/3}|D_n) < \gamma) \to 0$. Further, the convergence is uniform over $f_0 \in \mathcal{F}_+(K)$ for any $K > 0$.

(b) Let $f_0 \notin \bar{\mathcal{F}}_+$ be fixed and integrable. Using the properties of the projection, $d(f_0,\mathcal{F}_+) = \|f_0 - f_0^*\|_{1,G}$ is bounded by $\|f_0 - f^*\|_{1,G}$, which, by the triangle inequality, is further bounded above by

$$\|f_0 - f\|_{1,G} + \|f - f^*\|_{1,G} = \|f - f_0\|_{1,G} + d(f,\mathcal{F}_+).$$

This leads to $d(f,\mathcal{F}_+) \ge d(f_0,\mathcal{F}_+) - \|f - f_0\|_{1,G}$, and hence

$$\Pi(d(f,\mathcal{F}_+) \le M_n n^{-1/3}\big|D_n) \le \Pi(\|f_0 - f\|_{1,G} + M_n n^{-1/3} \ge d(f_0,\mathcal{F}_+)\big|D_n).$$

Let $\theta_{0j} = \int_{I_j} f_0 dG/G(I_j)$, $1 \le j \le J$. Then as shown in the proof of Theorem 3.1, $\Pi(\|f - f_{0J}\|_{1,G} > M_n\sqrt{J/n}\big|D_n) \to_{P_0} 0$, and hence for $J \asymp n^{1/3}$, we have $\Pi(\|f - f_{0J}\|_{1,G} > M_n n^{-1/3}\big|D_n) \to_{P_0} 0$. Next, since f_0 is integrable, by the martingale convergence theorem, $\|f_0 - f_{0J}\|_{1,G} \to 0$. Hence

$$\begin{aligned}&\mathrm{E}_0\Pi(\|f - f_0\|_{1,G} + M_n n^{-1/3} \ge d(f_0,\mathcal{F}_+)|D_n)\\ &\le \mathrm{E}_0\Pi\left(\|f - f_{0J}\|_{1,G} \ge d(f_0,\mathcal{F}_+) - \|f_{0J} - f_0\|_{1,G} - M_n n^{-1/3}\big|D_n\right) \to 0\end{aligned}$$

because $d(f_0, \mathcal{F}_+)$ is fixed and positive. This implies that the probability of Type 2 error $\mathrm{P}_0(\Pi(d(f, \mathcal{F}_+) \leq M_n n^{-1/3} | D_n) \geq \gamma) \to 0$.

(c) Let $f_0 \notin \mathcal{F}_+$ and $f_0 \in \mathcal{H}(\alpha, L)$ such that $d(f_0, \mathcal{F}) \geq \rho_n(\alpha)$. Consider the step function f_{0J} of f_0 as in part (b). By a well-known fact from approximation theory, we have that $\|f_0 - f_{0J}\|_{1,G} \leq C(L)J^{-\alpha}$ for some constant $C(L)$ depending only on L. For instance, the bound follows from de Boor [18] as step functions with equidistant points are B-splines of order 1. Hence for $J \asymp n^{1/3}$, we have $\Pi(\|f - f_0\|_{1,G} > M_n n^{-1/3} + C(L)n^{-\alpha/3} | D_n) \to_{P_0} 0$, uniformly for all $f_0 \in \mathcal{H}(\alpha, L)$. Thus by the triangle inequality, $d(f, \mathcal{F}_+) = \|f - f^*\|_{1,G}$ is bounded below by

$$\|f_0 - f^*\|_{1,G} - \|f - f_0\|_{1,G} \geq d(f_0, \mathcal{F}_+) - \|f - f_0\|_{1,G} \geq \rho_n(\alpha) - \|f - f_0\|_{1,G},$$

so that

$$\Pi(d(f, \mathcal{F}_+) \leq M_n n^{-1/3} | D_n) \leq \Pi(\|f - f_0\|_{1,G} \geq \rho_n(\alpha) - M_n n^{-1/3} | D_n) \to_{P_0} 0$$

because for $\alpha < 1$,

$$\rho_n(\alpha) - M_n n^{-1/3} \geq M_n n^{-1/3} + C(L)n^{-\alpha/3}$$

for $C > C(L)$, while for $\alpha = 1$,

$$\rho_n(\alpha) - M_n n^{-1/3} \geq M_n n^{-1/3} + C(L)n^{-\alpha/3}$$

for $C > 1$; the last follows because $M_n \to \infty$. □

Proof of Theorem 4.2. Let f_0 be a bounded, measurable true regression function (irrespective of monotonicity or smoothness). For a given J, consider $f_{0J} = \sum_{j=1}^{J} \theta_{0j} \mathbb{1}_{I_j}$ with $\theta_{0j} = \int_{I_j} f_0 dG$, $j = 1, \ldots, J$. First, we show that for a given $\gamma' > 0$ and sufficiently large constant M_0 and sample size n,

$$\mathrm{E}_0 \Pi(\|f - f_{0J}\|_{2,G} \geq M_0 \sqrt{(J \log n)/n}, J \leq J_n | D_n) < \gamma', \tag{6.5}$$

provided that $\log J_n \asymp \log n$. We write the expression inside the expectation as

$$\sum_{J=1}^{J_n} \Pi(J|D_n) \Pi\Big(\sum_{j=1}^{J} (\theta_j - \theta_{0j})^2 G(I_j) \geq M_0^2 J(\log n)/n \big| D_n\Big), \tag{6.6}$$

and bound

$$\begin{aligned} &\Pi\Big(\sum_{j=1}^{J} (\theta_j - \theta_{0j})^2 G(I_j) \geq M_0^2 J(\log n)/n \big| D_n\Big) \\ &\leq \frac{n \sum_{j=1}^{J} G(I_j)[\mathrm{Var}(\theta_j|D_n) + (\mathrm{E}(\theta|D_n) - \theta_{0j})^2]}{M_0^2 J \log n}. \end{aligned} \tag{6.7}$$

In view of Condition (DR), $G(I_j)$ are of the order $1/J$, and by Lemma 7.1, N_j are of the order n/J in probability uniformly in $j = 1, \ldots, J$. Under the boundedness assumption on the prior parameters and the sampling variance, $\mathrm{Var}(\theta_j|D_n) \lesssim 1/N_j \lesssim J/n$ with high probability, from the standard expressions for normal-normal conjugate setting (see the proof of Theorem 3.1).

To estimate $(\mathrm{E}(\theta|D_n) - \theta_{0j})^2$, with $\bar{Y}_j$ standing for $N_j^{-1}\sum_{i:X_i\in I_j} Y_i$ and $\bar{\varepsilon}_j$ standing for $N_j^{-1}\sum_{i:X_i\in I_j}\varepsilon_i$, we first observe that $|\bar{\varepsilon}_j|^2 \leq N_j^{-1}\log n \lesssim (J\log n)/n$ with high probability. Here we have used the maximal norm estimate using the squared-exponential Orlicz norm (see Lemma 2.2.2 of van der Vaart and Wellner [36]) and $\#\{\bar{\varepsilon}_j : j \leq J \leq J_n\} \lesssim J_n^2$. By the same argument and the boundedness of f_0, we also have

$$|N_j^{-1}\sum_{i:X_i\in I_j} f(X_i) - \theta_{0j}|^2 \lesssim N_j^{-1}\log n \lesssim (J\log n)/n$$

with high probability. Also, $|\bar{Y}_j|$ is uniformly bounded with high probability, because $Y_i = f_0(X_i) + \varepsilon_i$. Putting in the expression for $\mathrm{E}(\theta_j|D_n)$, we conclude that $\sum_{j=1}^{J}(\mathrm{E}(\theta_j|D_n) - \theta_{0j})^2 \leq (J\log n)/n$.

Putting these estimates in (6.7), we find that the expression is bounded by M_0^{-2} with high probability simultaneously for all $J \leq J_n$. Hence by (6.6), it follows that (6.5) holds.

We also observe that, if the posterior contracts at the rate ϵ_n at f_0 in the sense that $\mathrm{E}_0\Pi(f : d_H(f, f_0) > M_0\epsilon_n|D_n) \to 0$ for some $M_0 > 0$, then for some other constant $M_0' > 0$,

$$\mathrm{E}_0\Pi(J : d_H(f_{0J}, f_0) > M_0'\epsilon_n|D_n) \to 0. \tag{6.8}$$

To see this, let $\tilde{f}_{0J}$ stand for element of $\mathcal{F}_J$ closest to f_0 in terms of d_H. Recall that f_{0J} is the element of $\mathcal{F}_J$ closest to f_0 in terms of the $\|\cdot\|_{2,G}$-distance. The function $\tilde{f}_{0J}$ may be different from f_{0J}, but clearly $|f_{0J}| \leq K$ and $|\tilde{f}_{0J}| \leq K$, whenever $|f_0| \leq K$. Thus from the definitions of the respective minimizers and the equivalence of the metrics d_H and $\|\cdot\|_{2,G}$ on uniformly bounded functions, it follows that

$$d_H(\tilde{f}_{0J}, f_0) \leq d_H(f_{0J}, f_0) \lesssim \|f_{0J} - f_0\|_{2,G} \lesssim \|\tilde{f}_{0J} - f_0\|_{2,G} \lesssim d_H(\tilde{f}_{0J}, f_0).$$

Note that, as $\tilde{f}_{0J}$ is the closest to f_0 in $\mathcal{F}_J$ in terms of the metric d_H, so if for a J_0, $d_H(\tilde{f}_{0J_0}, f_0) > M_0\epsilon_n$, then $\Pi(J = J_0|D_n) \leq \Pi(J : d_H(\tilde{f}_{0J}, f_0) > M_0\epsilon_n|D_n)$. In view of the last display, f_{0J} can replace $\tilde{f}_{0J}$ in the assertion at the expense of changing the constant from M_0 to some appropriate M_0', giving (6.8).

(a) If $f_0 \in \mathcal{F}_+$, then $f_{0J} \in \mathcal{F}_+$. By Lemma 7.3, the $\mathbb{L}_2$-approximation rate of $\mathcal{F}_J$ with equidistant intervals at a monotone function is $J^{-1/2}$. Stated differently, in order to achieve an error bound ϵ, the number of intervals J should be chosen to be of the order ϵ^{-2}. Then standard arguments as in the proof of Theorem 3.2 show that the prior probability of a Kullback-Leibler neighborhood of size ϵ^2 is bounded below by $\exp\{-C_1\epsilon^{-2}\log(1/\epsilon)\}$. The required test with respect to d is automatically available, while the sieve can be chosen as in Theorem 3.2 and its entropy can be bounded in the same way by noting that d is bounded by the

$\mathbb{L}_2(G)$-metric, leading to a (suboptimal) contraction rate $\epsilon_n = (n/\log n)^{-1/4}$. It also follows that for J_n a large constant multiple of ϵ_n^{-1}, the prior probability of $J > J_n$ is exponentially small compared with the prior concentration, and hence $\{J > J_n\}$ has also exponentially small posterior probability (see Theorem 8.11 (iii′) of Ghosal and van der Vaart [23]). Since $\log J_n \lesssim \log n$, it follows that (6.5) holds.

(b) Let $f_0 \notin \bar{\mathcal{F}}_+$ be fixed and bounded. By the martingale convergence theorem, $\|f_{0J} - f_0\|_{2,G} \to 0$ as $J \to \infty$, so for a given $\epsilon > 0$, we can get J_0 (depending on ϵ but not depending on n) such that $\|f_{0J_0} - f_0\|_{2,G} < \epsilon/2$. Then for some $\delta > 0$, we have

$$\Pi(\|f - f_0\|_{2,G} < \epsilon) \geq \Pi(J = J_0)\Pi(\max\{|\theta_j - \theta_{0j}| : 1 \leq j \leq J_0\} < \delta) > 0.$$

Further, for J_{1n} an arbitrarily small multiple of $n/\log n$, the excess prior probability $\Pi(J > J_{1n})$ can be bounded by e^{-bn} for some $b > 0$ depending on c. Considering a sieve $\mathcal{P}_n = \left\{f = \sum_{j=1}^{J} \theta_j \mathbb{1}_{I_j}, \max_j |\theta_j| \leq n, J \leq J_{1n}\right\}$, standard estimates gives a bound for its $\mathbb{L}_\infty$-metric entropy an arbitrarily small multiple of n. Therefore it follows that (see Theorem 6.17 of Ghosal and van der Vaart [23]) that $\mathrm{E}_0\Pi(J > J_{1n}|D_n) \to 0$ and the posterior is consistent at f_0 with respect to d_H, because $d_H(f_1, f_2) \lesssim \|f_1 - f_2\|_\infty$.

Observe that for any $f \in \mathcal{F}_J$, $h \in \mathcal{F}_+$, by the triangle inequality

$$d_H(f, h) \geq d_H(f_0, h) - d_H(f, f_{0J}) - d_H(f_{0J}, f_0). \tag{6.9}$$

Since $f_0 \notin \bar{\mathcal{F}}_+$, the first term is bounded below by a fixed positive number for any $h \in \mathcal{F}_+$. The second term is bounded above by $\sqrt{(J \log n)/n}$ with high posterior probability, and J can be restricted to be at most J_{1n}, which can be taken to be an arbitrarily small multiple of $n/\log n$. Hence we can make the second terms as small as we like, with high posterior probability. By (6.8) and posterior consistency, the third term can also be made arbitrarily small with high posterior probability; note that here the posterior variation is due to the randomness of J only. Minimizing the left-hand side with respect to $h \in \mathcal{F}_+$, this shows that $d_H(f, \mathcal{F}_+)$ is larger than some fixed positive number with high posterior probability in true probability. If $J \leq J_{1n}$, this separation will exceed any fixed multiple of $\sqrt{(J \log n)/n}$ with high posterior probability for all $J \leq J_{1n}$, while the posterior probability of $J > J_{1n}$ is small in true probability. Hence the posterior probability of the event $\{d_H(f, \mathcal{F}_+) > M_0\sqrt{(J \log n)/n}\}$ tends to one in true probability, prompting the test to reject the null hypothesis of monotonicity with true probability tending to one.

(c) Let $f_0 \notin \mathcal{F}_+$ and $f_0 \in \mathcal{H}(\alpha, L)$ such that $d_H(f_0, \mathcal{F}_+) \geq \rho_n(\alpha)$. The proof is very similar to that of part (b) with the following changes. First, by the well-known $\mathbb{L}_\infty$-approximation rate $J^{-\alpha}$ at functions in $\mathcal{H}(\alpha, L)$ by step functions, and standard arguments as used in part (a) and (b), giving prior concentration and $\mathbb{L}_\infty$-metric entropy bounds, the posterior contraction rate at f_0 with respect to d_H is $\epsilon_n = (n/\log n)^{-\alpha/(2\alpha+1)}$, since the $\mathbb{L}_\infty$-metric is stronger than d_H. Also, with high posterior probability, J can be restricted to less than $J_{2n} \asymp n\epsilon_n^2/\log n = (n/\log n)^{1/(2\alpha+1)}$. This bounds the second term by a multiple of $(n/\log n)^{-\alpha/(2\alpha+1)}$ with high posterior probability. Finally, by (6.8), the third term is also bounded by a multiple of $(n/\log n)^{-\alpha/(2\alpha+1)}$ with high posterior probability with the true probability tending to one. Therefore, the expression on the right side of (6.9) is larger than $M_0\sqrt{(J\log n)/n}$ with high posterior probability. Thus the test rejects the null hypothesis of monotonicity with true probability tending to one, that is, the type II error probability goes to zero. □

7. Auxiliary results

Lemma 7.1. *If the predictors are random, Condition* (DR) *holds with $A_1 \le g \le a_2$, and $n/J \gg \log J$, then for $A_n = \{a_1 n/(2J) \le \min(N_1,\dots,N_J) \le \max(N_1,\dots,N_J) \le 2a_2 n/J\}$, we have $P_0(A_n)\to 1$. In other words, $N_1,\dots,N_J$ are simultaneously of the order n/J in probability.*

Proof. From $N_j \sim \mathrm{Bin}(n; G(I_j))$ and $a_1/J \le G(I_j) \le a_2/J$ for every $1\le j\le J$, a standard large deviation estimate for $\mathrm{P}(N_j \ge 2a_2 n/J)$ is $2e^{-Cn/J}$ for some constant $C>0$, and similarly for $\mathrm{P}(N_j \le a_1 n/(2J))$. Adding these probabilities J times, we get the desired result because a factor $\log J$ can be absorbed in n/J in the exponential. □

Lemma 7.2. *Let the predictors be deterministic satisfying Condition* (DD) *or be random satisfying Condition* (DR). *Let $f_0 \in \mathcal{F}_+$, the prior on f of Type* 1, *and Condition* (E) *holds. Then for $J\to\infty$ such that $J \ll n$, we have*

- *(a) the maximum marginal likelihood estimator $\hat\sigma_n^2$ converges in probability to σ_0^2 at the rate* $\max\{n^{-1/2}, n^{-1}J\}$.
- *(b) If $\sigma^2 \sim \mathrm{IG}(\beta_1,\beta_2)$ with $\beta_1>2$, $\beta_2>0$, then the marginal posterior distribution of σ^2 contracts at the rate* $\max\{n^{-1/2}, n^{-1}J\}$.

Proof. (a) Let $f_0\in\mathcal{F}_+$. We first show that there exists $\boldsymbol{\theta}_{0J} = (\theta_{01},\dots,\theta_{0J})$ such that $n^{-1}\|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_{0J}\|^2 \lesssim J^{-1}$ for deterministic $\boldsymbol{X}$, and $n^{-1}\mathrm{E}_G\|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_{0J}\|^2 \lesssim J^{-1}$ for random $\boldsymbol{X}$.

On a set with $\min\{N_j : 1\le j\le J\}>0$, let $\theta_{0j} = N_j^{-1}\sum_{i:X_i\in I_j} f_0(X_i)$. Using the monotonicity of f_0, we write $n^{-1}\|\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_{0J}\|^2$ as

$$\frac{1}{n}\sum_{j=1}^{J}\sum_{i:X_i\in I_j}(f_0(X_i)-\theta_{0j})^2 \le \frac{1}{n}\sum_{j=1}^{J}\sum_{i:X_i\in I_j}(f_0(j/J)-f_0((j-1)/J))^2$$
$$= \sum_{j=1}^{J}\frac{N_j}{n}\left(f_0(j/J)-f_0((j-1)/J)\right)^2. \quad (7.1)$$

For deterministic X, by Condition (DD) and the monotonicity of f_0, (7.1) is bounded by

$$\max_{1\le j\le J}\frac{N_j}{n}\sum_{j=1}^{J}[f_0(j/J)-f_0((j-1)/J)]^2 \le \max_{1\le j\le J}\frac{N_j}{n}(f_0(1)-f_0(0))^2 \lesssim J^{-1}. \quad (7.2)$$

For random X, using the fact that $N_j \sim \mathrm{Bin}(n; G(I_j))$, the expectation of (7.1) under G equals to $\sum_{j=1}^{J} G(I_j)\,(f_0(j/J) - f_0((j-1)/J))^2$, which, in view of Condition (DR), has the bound $\max_{1\le j\le J} G(I_j)(f_0(1) - f_0(0))^2 \lesssim J^{-1}$.

For the rest of the proof, we assume that X is fixed, satisfying Condition (DD); the random case can be dealt with similarly, by taking expectation with respect to G and using Condition (DR). We imitate the proof of Proposition 4.1 (a) of Yoo and Ghosal [37] but assuming that f_0 is monotone instead of smooth. Define $\boldsymbol{U} = (\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{\mathrm{T}} + \boldsymbol{I}_n)^{-1}$. We write

$$|\mathrm{E}_0(\hat{\sigma}_n^2) - \sigma_0^2| = |n^{-1}\sigma_0^2 \mathrm{tr}(\boldsymbol{U}) - \sigma_0^2| + n^{-1}(\boldsymbol{F}_0 - \boldsymbol{B\zeta})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{B\zeta})$$

and bound it by a constant multiple of

$$\begin{aligned} n^{-1}[\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{U}) + (\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_{0J})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{F}_0 - \boldsymbol{B}\boldsymbol{\theta}_{0J}) \\ +(\boldsymbol{B}\boldsymbol{\theta}_{0J} - \boldsymbol{B\zeta})^{\mathrm{T}}\boldsymbol{U}(\boldsymbol{B}\boldsymbol{\theta}_{0J} - \boldsymbol{B\zeta})]. \end{aligned} \tag{7.3}$$

Among these terms, only the middle term arising out of the approximation of the true function by step functions, is different from Yoo and Ghosal [37] — the other two terms are bounded by J/n considering step functions as B-splines of order 1 in one dimension. The second term can also be bounded by a multiple of J^{-1} in the same way Yoo and Ghosal [37] did using the $\mathbb{L}_2$-approximation rate $J^{-1/2}$ for monotone function, leading the upper bound a multiple of $J/n + J^{-1}$ for the expression in (7.3).

To complete the proof of part (a), we bound $\mathrm{Var}_0(\hat{\sigma}_n^2)$ by a multiple of n^{-1}. Again, we can follow the same steps in the proof of Proposition 4.1 (a) of Yoo and Ghosal [37] with the approximate rate for a smooth function replaced by the $\mathbb{L}_2(G_n)$-approximation rate J^{-1} for a monotone function. We also observe that the bounds obtained in the proof are uniform over $f_0 \in \mathcal{F}_+(K)$ for any $K > 0$.

Given part (a), the proof of part (b) follows exactly as in the proof of Proposition 4.1 (a) of Yoo and Ghosal [37]. □

Lemma 7.3. *Let $1 \le p < \infty$ and $K > 0$. Then for every $f \in \mathcal{F}_+(K)$ and $J > 1$, there exist $\theta_1 \le \cdots \le \theta_J$ from $[-K, K]$ such that the following assertions hold.*

(a) For any partition intervals $I_1, \dots, I_J$ and probability measure H satisfying $H(I_j) \le M/J$, with $f_J = \sum_{j=1}^{J} \theta_j \mathbb{1}_{I_j} \in \mathcal{F}_+(K)$, we have that $\int |f - f_J|^p dH \le MK^p/J$.

(b) For any probability measure H and $1 \le p < \infty$, there exist knots $0 = \xi_0 < \xi_1 < \cdots < \xi_{J-1} < \xi_J = 1$ from the topological support of H such that for any $f \in \mathcal{F}_+(K)$, there exits a function of the form $f_J = \sum_{j=1}^{J} \theta_j \mathbb{1}_{I_j} \in \mathcal{F}_+(K)$ satisfying $\int |f - f_J|^p dH \le K^p/J^p$, where $I_1 = [\xi_0, \xi_1]$, and $I_j = [\xi_{j-1}, \xi_j)$, $j = 2, \dots, J$.

Proof. Decomposing the integral in integrals over $I_1, \dots, I_J$, we have the discrepancy $\int |f - f_J|^p dH = \sum_{j=1}^{J} \int_{I_j} |f - f_J|^p dH$. By the monotonicity of f and the constancy of f_J, we have $f(x) - f_J(x) \le f(j/J) - f((j-1)/J)$ for all $x \in I_j$, and similarly $f_J(x) - f(x) \le f(j/J) - f((j-1)/J)$. Hence $\int |f - f_J|^p dH$ is bounded by

$$\sum_{j=1}^{J} H(I_j)|f(j/J) - f((j-1)/J)|^p \le MJ^{-1} \sum_{j=1}^{J} |f(j/J) - f((j-1)/J)|^p,$$

because $H(I_j) \le M/J$ for all $j = 1, \ldots, J$. Using the monotonicity of f, the last expression is bounded by $|f(1) - f(0)|^p$ by the estimate $\sum a_k^p \le (\sum a_k)^p$ for positive numbers $a_1, \ldots, a_k$ and $p \ge 1$.

The proof of part (b) is essentially contained in the proof of Theorem 2.7.5 of van der Vaart and Wellner [36], although their theorem is about a bound for the bracketing or metric entropy. Implicit in their construction is that, given $\epsilon > 0$, there exists a $J = J(\epsilon) \lesssim \epsilon^{-1}$, $0 \le \xi_1 < \cdots < \xi_{J-1} \le 1$ and $\theta_1, \ldots, \theta_J$ such that $f_J = \sum_{j=1}^{J} \theta_j \mathbb{1}_{I_j}$ satisfies $\|f - f_J\|_{p,H} < \epsilon$, where $I_1, \ldots, I_J$ form an interval partition of $[0, 1]$ with knots $0 = \xi_0 < \xi_1 < \cdots < \xi_{J-1} \le \xi_J = 1$. For instance, one of the lower brackets in their construction of an ϵ-bracketing will satisfy the approximation property. The role of ϵ and J can be reversed, in that, given J, we can first obtain $\epsilon > 0$ such that the corresponding $J(\epsilon)$ is within J.

Finally, we need to conclude that the knot points $\xi_1 < \cdots < \xi_{J-1}$ can be chosen from the support of H. The construction in van der Vaart and Wellner [36] assumed, without loss of generality, that H is uniform. For a general H, the quantile transform is applied, transforming the jth knot ξ_j to $H^{-1}(\xi_j)$, which belongs to the support of H. □

References

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

[2] 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. MR0073895

[3] Baraud, Y., Huet, S. and Laurent, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.* **33** 214–257. MR2157802

[4] 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.* John Wiley & Sons, London-New York-Sydney Wiley Series in Probability and Mathematical Statistics. MR0326887

[5] Barlow, R. E. and Brunk, H. D. (1972). The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* **67** 140–147. MR0314205

[6] Bhaumik, P. and Ghosal, S. (2015). Bayesian two-step estimation in differential equation models. *Electron. J. Statist.* **9** 3124–3154. MR3453972

[7] Bhaumik, P. and Ghosal, S. (2017). Efficient Bayesian estimation and uncertainty quantification in ordinary differential equation models. *Bernoulli* **23** 3537–3570. MR3654815

[8] Bhaumik, P. and Ghosal, S. (2017). Bayesian inference for higher-order ordinary differential equation models. *J. Multivariate Anal.* **157** 103–114. MR3641739

[9] Bhaumik, P., Shi, W. and Ghosal, S. (2021+). Bayesian generalized regression in partial differential equation models. To appear in Bernoulli.

[10] Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics* **65** **1** 198-205. MR2665861

[11] Bowman, A. W., Jones, M. C. and Gijbels, I. (1998). Testing Monotonicity of Regression. *J. Comput. Graph. Statist.* **7** 489–500.

[12] 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

[13] Chakraborty, M. and Ghosal, S. (2020). Bayesian inference on monotone regression quantile: coverage and rate acceleration. *Preprint.*

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

[15] Chakraborty, M. and Ghosal, S. (2021+). Rates and coverage for monotone densities using projection-posterior. *To appear in Bernoulli.*

[16] Chipman, H. A., George, E. I., McCulloch, R. E. and Shively, T. S. (2021). mBART: Multidimensional Monotone BART. *Bayesian Analysis* **1** 1–30. MR2758172

[17] Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[18] de Boor, C. (2001). *A Practical Guide to Splines*, Revised ed. Springer-Verlag New York, Inc. MR1900298

[19] De Leeuw, J., Kurt, H. and Mair, P. (2009). Isotone optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and active set methods. *J. Stat. Softw.* **32**.

[20] Durot, C. (2002). Sharp asymptotics for isotonic regression. *Probab. Theory Relat. Fields* **122** 222–240. MR1894068

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

[22] Ghosal, S. and Van Der Vaart, A. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35** 192–223. MR2332274

[23] 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

[24] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge University Press, New York Estimators, algo-

rithms and asymptotics. MR3445293

[25] 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

[26] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods* **580**. Springer. MR2648134

[27] Lin, L. and Dunson, D. B. (2014). Bayesian monotone regression using Gaussian process projection. *Biometrika* **101** 303–317. MR3215349

[28] Neelon, B. and Dunson, D. B. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60** 398–406. MR2066274

[29] Robertson, T. and Wright, F. T. (1973). Multiple isotonic median regression. *Ann. Statist.* **1** 422–432. MR0378224

[30] 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

[31] Salomond, J.-B. (2014). Concentration rate and consistency of the posterior distribution for selected priors under monotonicity constraints. *Electron. J. Stat.* **8** 1380–1404. MR3263126

[32] Salomond, J.-B. (2018). Testing un-separated hypotheses by estimating a distance. *Bayesian Anal.* **13** 461–484. MR3780431

[33] Scott, J. G., Shively, T. S. and Walker, S. G. (2015). Nonparametric Bayesian testing for monotonicity. *Biometrika* **102** 617–630. MR3394279

[34] 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* **71** 159–175. MR2655528

[35] van der Vaart, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press. MR1652247

[36] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Process With Applications to Statistics*. Springer-Verlag New York, Inc. MR1385671

[37] Yoo, W. W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.* **44** 1069–1102. MR3485954