POSTERIOR CONSISTENCY IN NONPARAMETRIC REGRESSION PROBLEMS UNDER GAUSSIAN PROCESS PRIORS

By Taeryon Choi and Mark J. Schervish*
Carnegie Mellon University

Posterior consistency can be thought of as a theoretical justification of the Bayesian method. One of the most popular approaches to nonparametric Bayesian regression is to put a nonparametric prior distribution on the unknown regression function using Gaussian processes. In this paper, we study posterior consistency in nonparametric regression problems using Gaussian process priors. We use an extension of the theorem of Schwartz (1965) for nonidentically distributed observations, verifying its conditions when using Gaussian process priors for the regression function with normal or double exponential (Laplace) error distributions. We define a metric topology on the space of regression functions and then establish almost sure consistency of the posterior distribution. Our metric topology is weaker than the popular L^1 topology. With additional assumptions, we prove almost sure consistency when the regression functions have L^1 topologies. When the covariate (predictor) is assumed to be a random variable, we prove almost sure consistency for the joint density function of the response and predictor using the Hellinger metric.

1. Introduction. In this paper, we verify almost sure consistency for posterior distributions in nonparametric regression problems when the prior distribution on the regression function is a Gaussian process. Such problems involve infinite-dimensional parameters, and consistency of posterior distributions is a much more challenging problem than in the finite-dimensional case. There are several reviews on nonparametric Bayesian methods and posterior consistency such as Wasserman (1998), Ghosal, Ghosh and Ramamoorthi (1999), Hjort (2002), Ghosh and Ramamoorthi (2003) and Choudhuri, Ghosal and Roy (2003). In addition, there have been many results giving general conditions under which features of posterior distributions are consistent in infinite-dimensional spaces. For examples, see Doob (1949), Schwartz (1965), Barron, Schervish and Wasserman (1999), Amewou-Atisso et al. (2003), Walker (2003) and Choudhuri, Ghosal and Roy (2004a,b).

Early results on posterior consistency have focused mainly on density estimation, that is on estimating a density function for a random sample without assuming the density belongs to a finite-dimensional parametric family. More recently, attention has turned to posterior consistency in nonparametric and semiparametric regression problems. Some popular nonparametric Bayesian regression methods are the techniques of orthogonal basis expansion, free-knot splines and Gaussian process priors. An orthogonal basis expansion for a regression function $\eta(x)$ is a representation as $\eta(x) = \sum_{i=1}^{\infty} \theta_i \varphi_i(x)$ where, $\{\varphi_i(x)\}_{i=1}^{\infty}$ is the orthonormal basis for an L_2 space. Asymptotic properties of these expansions have been studied by re-expressing the regression model as a problem of estimating the infinitely parameters $\{\theta_i\}_{i=1}^{\infty}$. This approach is called the *infinitely many normal means problem* and it has been studied extensively by Cox (1993), Freedman (1999), Zhao (2000)

*This paper is based on the first author's thesis work under the supervision of the second author in the Department of Statistics at Carnegie Mellon University.

AMS 2000 *subject classifications.* Primary 62G20, 62G08; secondary 60G15.

Key words and phrases. Almost sure consistency, Convergence in probability, Hellinger metric, L^1 metric, Laplace distribution.

and Shen and Wasserman (2001). Briefly, one models the θ_i's as independent normal random variables with mean 0 and variance τ_i^2. Freedman (1999) studied the nonlinear functional $\|\eta - \hat{\eta}\|^2$, where $\hat{\eta}$ is the Bayes estimator, both from the Bayesian and the frequentist perspectives. His main results imply consistency of the Bayes estimator for all $\{\theta_i\}_{i=1}^{\infty} \in \ell_2$ if $\sum_{i=1}^{\infty} \tau_i^2 < \infty$. Zhao (2000) showed a similar consistency result and that the Bayes estimator attains the minimax rate of convergence for certain class of priors. Shen and Wasserman (2001) investigated asymptotic properties of posterior distribution and obtained convergence rates. Huang (2004) also considered convergence rates of posterior distributions using sieve-based priors in the adaptive estimation, where the smoothness parameter is unknown. Denison, Mallick and Smith (1998) and DiMatteo, Genovese and Kass (2001) model η using free-knot splines. Specifically, they modeled η as a polynomial spline of fixed order, while putting a prior on the number of the knots, the locations of the knots and the coefficients of the polynomials. Generally, consistency in spline models can be shown using the the same methods as those used for orthonormal basis expansions as in Huang (2004). However, in free-knot spline models, consistency has not been investigated yet.

The approach on which we focus in this paper is to model η as a Gaussian processes a priori. Gaussian processes are a natural way of defining prior distributions over spaces of functions, which are the parameter spaces for nonparametric Bayesian regression models. O'Hagan (1978) and Wahba (1978) suggested the use of Gaussian processes as nonparametric regression priors, and essentially the same model has long been used in spatial statistics under the name of "kriging". Gaussian processes have been used successfully for regression and classification, particularly in machine learning (Seeger, 2004). Neal (1996) has shown that many Bayesian regression models based on neural networks converge to Gaussian processes in the limit as the number of nodes becomes infinite. This has motivated examination of Gaussian process models for the high-dimensional applications to which neural networks are typically applied (Rasmussen, 1996). Applications of Gaussian processes as priors in spatial statistics applications include Higdon, Swall and Kern (1998), Fuentes and Smith (2001), Paciorek (2003) and Paciorek and Schervish (2004).

Posterior consistency in nonparametric regression problems with Gaussian process priors has been studied mainly in the orthogonal basis expansion framework mentioned earlier. Brown and Low (1996), Freedman (1999) and Zhao (2000) have exploited an asymptotic equivalence between white-noise problems and nonparametric regression to prove that the existence of a consistent estimator in a white-noise problem implies the existence of a corresponding consistent estimator in a nonparametric regression problem. The white-noise problem is one in which an observation process $Y(x)$ is modeled as

$$Y(x) = \eta(x) + n^{-1/2}\epsilon(x),$$

where $\epsilon(x)$ is a Brownian motion. They use an infinitely many normal means prior for η and show that the posterior mean of η is consistent in the white-noise problem under certain conditions on the prior. The corresponding estimator in the nonparametric regression problem might not be the posterior mean of η, however. In addition, to use a prior distribution in the white-noise problem requires either knowing the eigenvalue decomposition of the desired covariance function (analytically, not numerically) or letting the covariance function be determined by the orthogonal basis. In many applications, such as those described by Neal (1996), Rasmussen (1996) and Seeger (2004), one uses a particular form of covariance function (like squared exponential) in order to model desired forms of dependence. In such cases, one would like to be able to verify consistency for the particular prior distribution that one is using. For example, Choudhuri, Ghosal and Roy (2004b) prove in-probability consistency of posterior distributions in binary regression problems with mild conditions on the Gaussian process prior. They do this by extending a result of Schwartz (1965) to the case of independent, not identically distributed observations.

We follow the approach of Choudhuri, Ghosal and Roy (2004b) for nonparametric regression problems with normal and other types of errors and unknown error variance while extending the results to almost sure consistency. First, we show almost sure consistency in a topology that is weaker than the L^1 topology used by Choudhuri, Ghosal and Roy (2004b). The need for a weaker topology arises from the fact that our regression functions can be unbounded. With the weaker topology, we can handle both random and nonrandom design points. When the design points are fixed, we strengthen the result to the case of the L^1 topology. When the design points are random, we prove almost sure consistency of the posterior probabilities of Hellinger neighborhoods of the joint density of the response and design point. Finally, if we assume that the regression functions are uniformly bounded, we prove almost sure consistency of L^1 neighborhoods of the true regression function.

The rest of the paper is organized as follows. In Section 2, we describe the model that we are using. In Section 3, we define our metric topology on the set of regression functions to establish almost sure consistency. In Section 4, we state the extension of Choudhuri, Ghosal and Roy (2004a). In Section 5, we state the assumptions needed to prove consistency in nonparametric regression and verify almost sure consistency. In Section 6, we give examples of covariance functions for Gaussian processes that both satisfy the smoothness conditions of our theorems and are used in practice. In Section 7, we discuss some directions on future work.

2. The model. Consider a random response Y corresponding to a single covariate X taking values in a bounded interval $T \subset \mathbb{R}$. We are interested in estimating the regression function, $\eta(x) = \mathrm{E}(Y|X = x)$ based on independent observations of (X, Y). We do not assume a parametric form for the regression function, but rather we assume some smoothness conditions. We model the unknown function η as a random process with a Gaussian process (GP) prior distribution. A Gaussian process is a stochastic process parameterized by its mean function $\mu : T \to \mathbb{R}$ and its covariance function $R : T^2 \to \mathbb{R}$ which we denote $GP(\mu, R)$. To say that $\eta \sim GP(\mu, R)$ means that, for all n and all $t_1, \ldots, t_n \in T$,, the joint distribution of $(\eta(t_1), \ldots, \eta(t_n))$ is an n-variate normal distribution with mean vector $(\mu(t_1), \ldots, \mu(t_n))$ and covariance matrix Σ whose (i, j) entry is $R(t_i, t_j)$.

To be specific, the GP regression model we consider here, is the following.

$$\begin{aligned}
Y_i &= \eta(X_i) + \epsilon_i \ , \qquad i = 1, \ldots, n, \\
\epsilon_i &\sim N(0, \sigma^2) \text{ or } DE(0, \sigma) \text{ given } \sigma, \\
\sigma &\sim \nu, \\
\eta(\cdot) &\sim \mathrm{GP}(\mu(\cdot), R(\cdot, \cdot)), \text{ independent of } \sigma \text{ and } (\epsilon_1, \ldots, \epsilon_n),
\end{aligned}$$

where ν is a probability measure with support $\mathbb{R}^+$, and $DE(0, \sigma)$ stands for the double exponential (or Laplace) distribution with median 0 and scale factor σ. The objective of this paper is to identify conditions on the GP prior distribution of η and the sequence of predictors $\{X_i\}_{i=1}^{\infty}$ that guarantee almost sure consistency of the posterior distribution under the model described above.

Suppose that the true response function, $\eta_0(x)$ as a function of the covariate X, is a continuously differentiable function on a bounded interval T. Without loss of generality, we will assume that $T = [0, 1]$ for the remainder of this paper. Our work is similar to that of Choudhuri, Ghosal and Roy (2004b) who give conditions for in-probability consistency of posteriors in general non-identically distributed data problems. We identify conditions on the GP prior that allow us to verify the conditions of their theorem, and we prove an extension of their theorem to the case of almost sure consistency. Under somewhat different assumptions about the GP prior, we are able to verify the conditions of this extension.

3. Topologies on the set of regression functions. First, we need to be clear on what we mean by the expression "almost surely consistent". Let $\mathcal{F}$ be the set of Borel measurable functions defined on T. For now, assume that we have chosen a topology on $\mathcal{F}$. For each neighborhood N of the true regression function η_0 and each sample size n, we compute the posterior probability

$$p_{n,N}(Y_1,\ldots,Y_n,X_1,\ldots,X_n) = \Pr(\{\eta \in N\}|Y_1,\ldots,Y_n,X_1,\ldots,X_n),$$

as a function of the data. To say that the posterior distribution of η is almost surely consistent means that, for every neighborhood N, $\lim_{n\to\infty} p_{n,N} = 1$ a.s. with respect to the joint distribution of the infinite sequence of data values. Similarly, in-probability consistency means that for all N, $p_{n,N}$ converges to 1 in probability.

To make these definitions precise, we must specify the topology on $\mathcal{F}$. This topology can be chosen independently of whether one wishes to consider almost sure consistency or in-probability consistency of the posterior. Popular choices of topology on $\mathcal{F}$ include the L^p topologies related to a probability measure Q on the domain T of the regression functions. For $1 \le p < \infty$, the $L^p(Q)$ distance between two functions η_1 and η_2 is $\|\eta_1 - \eta_2\|_p = \left[\int_T |\eta_1 - \eta_2|^p dQ\right]^{1/p}$. For $p = \infty$, the $L^\infty(Q)$ distance is

$$\|\eta_1 - \eta_2\|_\infty = \inf_{A:Q(A)=1} \sup_{x\in A} |\eta_1(x) - \eta_2(x)|.$$

For example, Choudhuri, Ghosal and Roy (2004b) use the L^1 topology related to Lebesgue measure and prove in-probability consistency in the binary regression setting. Another topology on $\mathcal{F}$ is the topology of in-probability convergence related to a probability Q, and we prove almost sure consistency. This topology is weaker than the $L^p(Q)$ topologies. As with the $L^p(Q)$ topologies, we must count as identical all functions that equal each other a.s. $[Q]$. Lemma 1 gives a metric representation of the topology of in-probability convergence.

LEMMA 1. *Let $(T, \mathcal{B}, \mathcal{Q})$ be a probability space, and let $\mathcal{F}$ be the set of all real-valued measurable functions defined on T. Define*

$$d_Q(\eta_1,\eta_2) = \inf\{\epsilon : Q(\{x : |\eta_1(x) - \eta_2(x)| > \epsilon\}) < \epsilon\}.$$

Then d_Q is a metric on the set of equivalence classes under the relation $\eta_1 \sim \eta_2$ if $\eta_1 = \eta_2$ a.s. $[Q]$. If X has distribution Q, then $\eta_n(X)$ converges to $\eta(X)$ in probability if and only if $\lim_{n\to\infty} d_Q(\eta_n,\eta) = 0$.

It is well known that $L^p(Q)$ convergence implies in-probability convergence, so that $L^p(Q)$ neighborhoods must be smaller than d_Q neighborhoods in some sense. It is not difficult to show that for every $\epsilon \in (0,1)$, the ball of radius ϵ under d_Q contains the ball of radius $\epsilon^{1+1/p}$ in $L^p(Q)$ for all $1 \le p \le \infty$. In addition, for every p and every $\delta > \epsilon^{1+1/p}$, there exist functions in the ball of radius δ under $L^p(Q)$ that are not in the ball of radius ϵ under d_Q. When the random variables are all bounded, in-probability convergence implies L^p convergence for all finite p.

When the values of the predictor X are chosen deterministically (and satisfy a condition relative to Lebesgue measure λ) we prove almost sure consistency of posterior probabilities of $L^1(\lambda)$ neighborhoods of the true regression function. When the values of the predictor X have a distribution Q, we prove almost sure consistency of posterior probabilities of d_Q neighborhoods of the true regression function. If we make an additional assumption that the regression function is uniformly bounded by a known constant, we can prove almost sure consistency of posterior probabilities of $L^1(Q)$ neighborhoods even when X is random.

An alternative to topologizing regression functions is to place a topology on the set of distributions. For the case in which the predictor X is random, we shall consider this alternative as well as the topologies mentioned above. In particular, we shall use the Hellinger metric on the collection of joint distributions of (X, Y). Suppose that X has distribution Q and Y has a density $f(y|x)$ with respect to Lebesgue measure λ given $X = x$. Then $f(y|x)$ is a joint density of (X, Y) with respect to $\nu = Q \times \lambda$. Under the true regression function η_0 with noise scale parameter σ_0, we denote the conditional density of Y by $f_0(y|x)$. In this case, the Hellinger distance between the two distributions corresponding to (η, σ) and (η_0, σ_0) is the following:

$$d_H(f, f_0) = \int \left[\sqrt{f(y|x)} - \sqrt{f_0(y|x)}\right]^2 d\nu(x, y).$$

It is easy to show that $d_H(f, f_0)$ is unchanged if one chooses a different dominating measure instead of ν. For the case just described, we will show that, under conditions similar to those of our other theorems, the posterior probability of each Hellinger neighborhood of f_0 converges almost surely to 1.

4. Consistency theorems for non-i.i.d. observations. Schwartz (1965) proved a theorem that gave conditions for consistency of posterior distributions of parameters of the distributions of independent and identically distributed random variables. These conditions include the existence of tests with sufficiently small error rates and the prior positivity of certain neighborhoods of η_0. Choudhuri, Ghosal and Roy (2004a) extend the theorem of Schwartz to a triangular array of independent non-identically distributed observations for the case of convergence in-probability. We provide another extension of Schwartz's theorem to almost sure convergence. Our extension is based on both Amewou-Atisso et al. (2003) and Choudhuri, Ghosal and Roy (2004a). We present this extension as Theorem 1 and also verify the conditions for a wide class of GP priors. The proofs of all theorems stated in the body of this paper are given in an appendix at the end.

THEOREM 1. *Let $\{Z_i\}_{i=1}^{\infty}$ be independently distributed with densities $\{f_i(\cdot;\theta)\}_{i=1}^{\infty}$, with respect to a common σ-finite measure, where the parameter θ belongs to an abstract measurable space Θ. The densities $f_i(\cdot;\theta)$ are assumed to be jointly measurable. Let $\theta_0 \in \Theta$ and let P_{θ_0} stand for the joint distribution of $\{Z_i\}_{i=1}^{\infty}$ when θ_0 is the true value of θ. Let $\{U_n\}_{n=1}^{\infty}$ be a sequence of subsets of Θ. Let θ have prior Π on Θ. Define*

$$\begin{aligned}
\Lambda(\theta_0, \theta) &= \log \frac{f_i(Z_i;\theta_0)}{f_i(Z_i;\theta)}, \\
K_i(\theta_0, \theta) &= \mathrm{E}_{\theta_0}(\Lambda(\theta_0, \theta)), \\
V_i(\theta_0, \theta) &= \mathrm{Var}_{\theta_0}(\Lambda(\theta_0, \theta)).
\end{aligned}$$

(A1) Prior positivity of neighborhoods.
Suppose that there exists a set B with $\Pi(B) > 0$ such that

(i) $\displaystyle\sum_{i=1}^{\infty} \frac{V_i(\theta_0, \theta)}{i^2} < \infty, \ \forall\, \theta \in B,$

(ii) *For all $\epsilon > 0$, $\Pi(B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon$ for all $i\}) > 0$.*

(A2) Existence of tests
Suppose that there exist test functions $\{\Phi_n\}_{n=1}^{\infty}$, sets $\{\Theta_n\}_{n=1}^{\infty}$ and constants $C_1, C_2, c_1, c_2 > 0$ such that

(i) $\displaystyle\sum_{n=1}^{\infty} \mathrm{E}_{\theta_0}\Phi_n < \infty,$

(ii) $\sup_{\theta \in U_n^C \bigcap \Theta_n} \mathrm{E}_\theta(1-\Phi_n) \le C_1 e^{-c_1 n}$,

(iii) $\Pi(\Theta_n^C) \le C_2 e^{-c_2 n}$.

Then

$$\Pi(\theta \in U_n^C | Z_1, \ldots, Z_n) \to 0 \quad a.s.[P_{\theta_0}]. \tag{1}$$

The first condition (A1) assumes that there are sets with positive prior probabilities, which could be regarded as neighborhoods of the true parameter θ_0. We assume that the true value of the parameter is included in the Kullback-Leibler neighborhood according to the prior Π. The second condition (A2) assumes the existence of certain tests of the hypothesis $\theta = \theta_0$. We assume that tests with vanishingly small type I error probability exist. We also assume that these tests have exponentially small type II error probability on part of the complement of a set U_n containing θ_0, namely $\Theta_n \cap U_n^C$.

5. Consistency in nonparametric regression. In this section, we apply Theorem 1 to cases in which the prior Π is a GP distribution as described in Section 2. We must make assumptions about the smoothness of the GP prior as well as about the rate at which the design points x_i's, $i = 1, \ldots, n$ fill out the interval $[0,1]$. For the latter, we consider two versions of the assumption on design points, one for random covariates and one for nonrandom (fixed) covariates.

Assumption RD. The design points (covariates) $\{X_n\}_{n=1}^\infty$ are independent and identically distributed with probability distribution Q on $[0,1]$.

Assumption NRD. Let $x_1 \le x_2 \le \cdots \le x_n$ be the design points on $[0,1]$ and let $S_i = x_{i+1} - x_i$ $i = 1, \ldots, n-1$ denote the spacings between them. There is a constant $0 < K_1 < 1$ such that the $\max_{1 \le i < n} S_i < 1/(K_1 n)$.

Assumption NRD is obviously satisfied by the equally spaced design.

Our smoothness condition on the GP prior is slightly weaker than that of Choudhuri, Ghosal and Roy (2004b).

Assumption P. The Gaussian process $\eta(x)$ has a continuously differentiable mean function $\mu(x)$ and the covariance function $R(x, x')$ has continuous fourth partial derivatives. In addition, ν assigns positive probability to every neighborhood of the true variance σ_0^2, i.e. for every $\epsilon > 0$, $\nu\left\{\left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\} > 0$.

Assumption P can be verified for many popular covariance functions of Gaussian processes. These include both stationary and nonstationary covariance functions. We will give some examples in Section 6.

To apply Theorem 1 to the nonparametric regression problem, we use the following notation. The parameter θ in Theorem 1 is (η, σ) with $\theta_0 = (\eta_0, \sigma_0)$. The density $f_i(\cdot; \theta)$ is the normal density with mean $\eta(x_i)$ and variance σ^2 or the double exponential density with location parameter $\eta(x_i)$ and scale parameter σ. The parameter space Θ is a product space of a function space Θ_1 and $\mathbb{R}^+$. Let θ have prior Π, a product measure, $\Pi_1 \times \nu$, where Π_1 is a Gaussian process prior for η and ν is a prior for σ. A sieve, Θ_n is constructed to facilitate finding uniformly consistent tests. Finally, we define the sets that play the roles of U_n and contain θ_0 in terms of the various topologies that we will use, one for d_Q, one for L^1, and one for Hellinger. In our theorems, these sets are the same for all n.

$$U_\epsilon \;=\; \left\{(\eta, \sigma) \;:\; d_Q(\eta, \eta_0) < \epsilon, \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}, \tag{2}$$

$$
\begin{aligned}
W_\epsilon &= \left\{(\eta,\sigma) : \|\eta-\eta_0\|_1 < \epsilon, \left|\frac{\sigma}{\sigma_0}-1\right| < \epsilon\right\}, \\
H_\epsilon &= \{f : d_H(f,f_0) < \epsilon\}.
\end{aligned}
$$

A point (η,σ) is in U_ϵ so long as σ is close to σ_0 and η differs greatly from η_0 only on a set of small Q measure. It doesn't matter by how much η differs from η_0 on that set of small Q measure. Although U_ϵ is not necessarily open in any familiar topology, it does contain familiar open subsets. For each $1 \le p \le \infty$, U_ϵ contains $W_{\epsilon^{1+1/p}}$. For the cases in which the noise terms have normal or Laplace distributions, we will also show that for each ϵ there is a δ such that H_ϵ contains U_δ.

In summary, the main theorems that we prove in the appendix are the following, in which the data $\{Y_n\}_{n=1}^\infty$ are assumed to be conditionally independent with either normal or Laplace distributions given η, σ and the covariates.

Theorem 2. *Suppose that the values of the covariate in* $[0,1]$ *arise according to a nonrandom design satisfying Assumption NRD. Assume that the prior satisfies Assumption P. Let* P_0 *denote the joint conditional distribution of* $\{Y_n\}_{n=1}^\infty$ *assuming that* η_0 *is the true response function and* σ_0^2 *is the true noise variance. Assume that the function* η_0 *is continuously differentiable. Then for every* $\epsilon > 0$,

$$\Pi\left\{ W_\epsilon^C \middle| Y_1,\ldots,Y_n,x_1,\ldots,x_n\right\} \to 0 \quad a.s.[P_0].$$

Theorem 3. *Suppose that the values of the covariate in* $[0,1]$ *arise according to a design satisfying Assumption RD. Assume that the prior satisfies Assumption P. Let* P_0 *denote the joint conditional distribution of* $\{Y_n\}_{n=1}^\infty$ *given the covariate assuming that* η_0 *is the true response function and* σ_0^2 *is the true noise variance. Assume that the function* η_0 *is continuously differentiable. Then for every* $\epsilon > 0$,

$$\Pi\left\{ U_\epsilon^C \middle| Y_1,\ldots,Y_n,x_1,\ldots,x_n\right\} \to 0 \quad a.s.[P_0].$$

Theorem 4. *Suppose that the values of the covariate in* $[0,1]$ *arise according to a design satisfying Assumption RD. Assume that the prior satisfies Assumption P. Let* P_0 *denote the joint distribution of each* (X_n,Y_n) *and let* f_0 *denote the joint density assuming that* η_0 *is the true response function and* σ_0 *is the true noise scale parameter. Assume that the function* η_0 *is continuously differentiable. Then for every* $\epsilon > 0$,

$$\Pi\left\{ H_\epsilon^C \middle| (X_1,Y_1),\ldots,(X_n,Y_n)\right\} \to 0 \quad a.s.[P_0].$$

Finally, when we deal with random covariates, we can prove consistency of posterior probabilities of L^1 neighborhoods for the case in which the support of the prior distribution contains only uniformly bounded regression functions.

Assumption B. Let Π_1' and ν be a Gaussian process and a prior on σ satisfying Assumption P. Let $\Omega = \{\eta : \|\eta\|_\infty < M\}$ with $M > \|\eta_0\|_\infty$. Assume that $\Pi_1(\cdot) = \Pi_1'(\cdot \cap \Omega)/\Pi_1'(\Omega)$.

Theorem 5. *Suppose that the values of the covariate in* $[0,1]$ *arise according to a fixed design satisfying Assumption RD. Assume that the prior satisfies Assumption B. Let* P_0 *denote the joint conditional distribution of* $\{Y_n\}_{n=1}^\infty$ *given the covariate assuming that* η_0 *is the true response function and* σ_0^2 *is the true noise variance. Assume that the function* η_0 *is continuously differentiable. Then for every* $\epsilon > 0$,

$$\Pi\left\{ W_\epsilon^C \middle| (X_1,Y_1),\ldots,(X_n,Y_n)\right\} \to 0 \quad a.s.[P_0],$$

6. Smoothness conditions on Gaussian process priors. In Assumption P, we required some smoothness conditions on the covariance function in the Gaussian process as a prior distribution for η. The important consequence of Assumption P is that there exists a constant K_2 such that

$$\Delta_h \Delta_h R_{11}(t,t) \leq K_2 |h|^2,$$

where

$$\Delta_h \Delta_h R_{11} \equiv R_{11}(t+h, t+h) - R_{11}(t+h, t) - R_{11}(t, t+h) + R_{11}(t,t)$$

and

$$R_{11}(s,t) \equiv \partial^2 R(s,t)/\partial s \partial t$$

This condition guarantees the existence of continuous sample derivative $\eta'(\cdot)$ with probability 1. (See Lemma 5 in the appendix.)

Many covariance functions of Gaussian processes, which are widely used in the literature mentioned earlier, satisfy Assumption P. We give some illustrations of these covariance functions in this section.

6.1. *Stationary Gaussian process $X(t)$ with isotropic covariance function.* The covariance function $R(x,x')$ depends on distance between x and x' alone, i.e. $R(x,x') = R(|x-x'|)$

- squared-exponential covariance function

$$R(h) = \exp(-h^2) = 1 - h^2 + O(h^4), \quad \text{as } h \to 0$$

- Cauchy covariance function

$$R(h) = \frac{1}{1+h^2} = 1 - h^2 + O(h^4), \quad \text{as } h \to 0$$

- Matérn covariance function with $\nu > 2$

$$R(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}} (\alpha h)^\nu \mathcal{K}_\nu(\alpha h),$$

  where $\alpha > 0$ and $\mathcal{K}_\nu(x)$ is a modified Bessel function of order ν.

  It is known from Abrahamsen (1997, p. 43) that for $n < \nu$ then

$$\left. \frac{d^{2n-1} R(h)}{dh^{2n-1}} \right|_{h=0} = 0$$

  and

$$\left. \frac{d^{2n} R(h)}{dh^{2n}} \right|_{h=0} \in (-\infty, 0)$$

  Consequently, it is straightforward that if $\nu > 2$, then there exists a constant $\xi > 0$ such that

$$R(h) = R(0) - \xi h^2 + O(h^4), \quad \text{as } h \to 0$$

Clearly, in all three of the above cases,

$$\begin{aligned} |\Delta_h \Delta_h R_{11}(t,t)| &= |R_{11}(t+h, t+h) - 2R_{11}(t, t+h) + R_{11}(t,t)| \\ &= |2R_{11}(0) - 2R_{11}(h)| \\ &\leq K_2 h^2 \end{aligned}$$

6.2. *Nonstationary Gaussian process.* Let $Y(t) = \sigma(t)X(t), \quad t \in [0,1]$ where $\sigma(t)$ is a twice continuously differentiable function and $X(t)$ is one of the stationary Gaussian processes listed above

$$\begin{aligned}
\mathrm{Cov}\{Y(s), Y(t)\} \equiv R^Y(s,t) &= \sigma(t)\sigma(s)\mathrm{Cov}\{X(s), X(t)\} \\
&= \sigma(t)\sigma(s)R(|t-s|)
\end{aligned}$$

It can be shown that for some positive constants, $K_2, K_3, K_4 > 0$

$$\begin{aligned}
\Delta_h\Delta_h R^Y_{11}(t,t) &= K_2\left\{\sigma'(t+h) - \sigma'(t)\right\}^2 + 2\sigma'(t+h)\sigma'(t)K_2[\lambda h^2 - O(h^4)] \\
&\leq K_2 \sup_{t\in[0,1]} |\sigma''(t)|^2|h|^2 + 2K_3 \sup_{t\in[0,1]} |\sigma'(t)|^2 h^2 \\
&\leq K_4 h^2
\end{aligned}$$

6.3. *Convolution of white noise process with convolution kernel.*

$$Z(s) = \int_{\mathrm{I\!R}} \mathbf{K}_s(u)X(u)du, \tag{3}$$

where $X(s)$ is a white-noise process and $\mathbf{K}_s(\cdot)$ is a kernel. (The integral in (3) is understood in the mean-square sense.)

By Fubini's theorem,

$$\mathrm{E}\{Z(s)\} = \int_{\mathrm{I\!R}} \mathbf{K}_s(u)\mathrm{E}(X(u))du$$

$$\begin{aligned}
\mathrm{Cov}\{Z(s), Z(t)\} &= \mathrm{E}\{Z(s)Z(t)\} - \mathrm{E}\{Z(s)\}\mathrm{E}\{Z(t)\} \\
&= \int_{\mathrm{I\!R}}\int_{\mathrm{I\!R}} \mathrm{E}\{\mathbf{K}_s(u)X(u)\mathbf{K}_t(w)X(w)\}dudw \\
&= \int_{\mathrm{I\!R}} \mathbf{K}_s(u)\mathbf{K}_t(u)du \equiv R^Z(s,t)
\end{aligned}$$

For example, take

$$\mathbf{K}_s(u) = \phi(s-u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(s-u)^2\right)$$

Then

$$\begin{aligned}
R^Z(s,t) &= \int_R \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(s-u)^2\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(t-u)^2\right)du \\
&= \frac{1}{2\pi}\exp\left(-\frac{1}{4}(s-t)^2\right),
\end{aligned}$$

which belongs to the previous stationary Gaussian processes case

6.4. *Nonstationary processes proposed by Higdon, Swall and Kern (1998) or Paciorek and Schervish (2004).*

$$\begin{aligned}
R(s,t) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{1}{2\sigma_s^2}(s-u)^2\right) \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left(-\frac{1}{2\sigma_t^2}(t-u)^2\right) du \\
&= \frac{1}{\sqrt{2\pi(\sigma_s^2+\sigma_t^2)}} \exp\left(-\frac{1}{2(\sigma_s^2+\sigma_t^2)}(s-t)^2\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
R_{10}(s,t) &= \frac{\partial}{\partial s} R(s,t) \\
&= -\frac{2\sigma_s \sigma_s'}{(\tau_s^2+\sigma_t^2)\sqrt{2\pi(\sigma_s^2+\sigma_t^2)}} \exp\left(-\frac{1}{2(\sigma_s^2+\sigma_t^2)}(s-t)^2\right) \\
&\quad + \frac{1}{\sqrt{2\pi(\sigma_s^2+\sigma_t^2)}} \exp\left(-\frac{1}{2(\sigma_s^2+\sigma_t^2)}(s-t)^2\right) \frac{1}{2}\left(\frac{-2(s-t)(\sigma_s^2+\sigma_t^2)+(s-t)^2 2\sigma_s\sigma_s'}{(\sigma_s^2+\sigma_t^2)^2}\right)
\end{aligned}$$

and by the tedious calculation, it turns out that if σ_s and σ_t are continuously differentiable, $R_{11}(s,t)$ can be written as

$$R_{11}(s,t) \le \left\{K_1(s-t)^2 + K_2(s-t) + K_3\right\} \exp\left(-K_4(s-t)^2\right)$$

for some positive constants $K_1, \ldots, K_4$. Consequently, there exsts a positive constant K_5 such that

$$\begin{aligned}
|\Delta_h \Delta_h R_{11}(t,t)| &= |R_{11}(t+h,t+h) - 2R_{11}(t,t+h) + R_{11}(t,t)| \\
&\le K_5 h^2,
\end{aligned}$$

which also satisfies Assumption P.

7. Discussion. We have provided almost sure consistency of posterior probabilities of various metric neighborhoods of the true regression function in nonparametric regression problems using Gaussian process priors. We have also verified that the conditions for consistency hold for several classes of priors that are already used in practice.

We found that the case of random covariates is more challenging than nonrandom covariates. This is due to the fact that it is difficult to insure that the covariates will spread themselves uniformly enough to obtain consistency at all smooth true regression functions. The problem arises when we consider regression functions with arbitrarily large upper bound. In this case, we can find functions η that are far from the true regression function η_0 in L^1 distance, but differ from η_0 very little over almost all of the covariate space. A random sample of covariates will not have much chance of containing sufficiently many points x such that $|\eta(x) - \eta_0(x)|$ is large. The metric d_Q declares such η functions to be close to η_0 while the L^1 metric might declare them to be far apart. Also, when the noise distribution is normal or Laplace, functions that are close to η_0 in d_Q distance produce similar joint distributions for the covariate and response in terms of Hellinger distance. These distinctions disappear when the space of possible regression functions is known to be uniformly bounded a priori.

There are several open issues that are worth further consideration. First, we need to treat the case of multidimensional covariates. There are some subtle issues concerning almost sure smoothness of sample paths of Gaussian processes with multidimensional index set. Second, we have said nothing about rates of convergence. Ghosal and van der Vaart (2004) present general results on convergence rates for non i.i.d observations which include nonparametric regression cases. They mention the general results for nonparametric regression but do not consider specific prior distributions. Third, we need to think about the case in which the covariance function of the Gaussian process has (finitely many) parameters that need to be estimated. That is, we assume that η has a distribution $GP(\mu, R_\vartheta)$ conditional on ϑ. This is a typical case in applications where the various parameters that govern the smoothness of the GP prior are not sufficiently well understood to be chosen with certainty. It is true that every result that holds with probability 1 conditional on ϑ for all ϑ holds with probability 1 marginally. However, the posterior distribution of η that is computed when ϑ is treated as a parameter is not the same as the conditional posterior given ϑ, but rather it is the mixture of those posteriors with respect to the posterior distribution of ϑ. Additional work will be required to deal with this case. Fourth, Assumption NRD is perhaps a bit strong. Choudhuri, Ghosal and Roy (2004b), in a problem with uniformly bounded regression functions, use a condition on the design points that is weaker than Assumption NRD. We have not tried to find the weakest condition that guarantees almost sure consistency. Fianlly, we have assumed that the form of the error distribution is known (normal or Laplace). However, the results of Kleijn and Van der Vaart (2002) suggests that misspecification of the error distribution does not matter for regression with uniformly bounded regression function. It would be interesting to investigate the extent to which misspecification of the error distribution matters in Gaussian process regression.

APPENDIX

A.1. Overview of proofs. This appendix is organized as follows. In Section A.2, we prove the general Theorem 1. Section A.3 contains the proof of Lemma 1. The rest of the appendix contains the proofs of the main consistency results. We stated several theorems with different conditions on the design (random and nonrandom designs) and different topologies (L^1, d_Q, and Hellinger). The proofs of these results all rely on Theorem 1, and thereby have many steps in common. Section A.4 contains the proof of condition (A1) of Theorem 1, which is virtually the same for all of the main theorems. Section A.5 shows how we construct the sieve that is used in condition (A2). We also verify subcondition (iii) in that section. In Section A.6, we show how to construct uniformly consistent tests. This is done by piecing together finitely many tests, one for each element of a covering of the sieve by L^∞ balls. This section contains two separate results concerning the spacing of design points in the random and nonrandom covariate cases (Lemmas 8 and 10). Section A.7 explains why regression functions that are close in d_Q metric lead to joint distributions of (X, Y) that are close in Hellinger distance. This proves Theorem 4. Finally, we verify that Assumption B leads to consistency of posterior probabilities of L^1 neighborhoods in Section A.8.

A.2. Proof of Theorem 1. The posterior probability (1) can be written as

$$\begin{aligned}
\Pi(\theta \in U_n^C | Z_1, \ldots, Z_n) &= \frac{\int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta)}{\int_\Theta \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta)} \\
&\leq \Phi_n + \frac{(1-\Phi_n)\int_{U_n^C \cap \Theta_n} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta) + \int_{U_n^C \cap \Theta_n^C} \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta)}{\int_\Theta \prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)} d\Pi(\theta)} \\
&= \Phi_n + \frac{I_{1n}(Z_1, \ldots, Z_n) + I_{2n}(Z_1, \ldots, Z_n)}{I_{3n}(Z_1, \ldots, Z_n)}. \qquad (4)
\end{aligned}$$

The remainder of the proof consists of proving the following results:

$$\begin{array}{lrcl}
(5) & \Phi_n & \to & 0 \quad \text{a.s.}[P_{\theta_0}], \\
(6) & e^{\beta_1 n} I_{1n}(Z_1, \ldots, Z_n) & \to & 0 \quad \text{a.s.}[P_{\theta_0}] \text{ for some } \beta_1 > 0, \\
(7) & e^{\beta_2 n} I_{2n}(Z_1, \ldots, Z_n) & \to & 0 \quad \text{a.s.}[P_{\theta_0}] \text{ for some } \beta_2 > 0, \\
(8) & e^{\beta n} I_{3n}(Z_1, \ldots, Z_n) & \to & \infty \quad \text{a.s.}[P_{\theta_0}] \text{ for all } \beta > 0.
\end{array}$$

Letting $\beta > \max\{\beta_1, \beta_2\}$ will imply (1).

The first term on the right hand side of (4) goes to 0 with probability 1, by the first Borel-Canteli lemma from (A2) (i). We show that the two terms in the numerator, I_{1n} and I_{2n} are exponentially small and for some $\beta_1 > 0$ and $\beta_2 > 0$, $e^{\beta_1 n} I_{1n}$ and $e^{\beta_2 n} I_{2n}$ goes to 0 from (A2) (ii) and (iii) with $P_{\theta_0}^n$ probability 1. Finally, we show $e^{\beta r} I_{3n} \to \infty$, with $P_{\theta_0}^n$ probability 1 for all $\beta > 0$, using Kolmogorov's strong law of large numbers for independent but not identically distributed random variables under the condition (A1).

First, we prove (5). By the Markov inequality, for every $\epsilon > 0$, $P_{\theta_0}(|\Phi_n| > \epsilon) \leq \mathrm{E}_{\theta_0}(|\Phi_n|)$. By (i) of (A2), we have $\sum_{n=1}^{\infty} P_{\theta_0}(|\Phi_n| > \epsilon) < \infty$. By the first Borel-Cantelli lemma, $P_{\theta_0}(|\Phi_n| > \epsilon \text{ i.o.}) = 0$. Since this is true for every $\epsilon > 0$, we have (5).

Next, we prove (6). For every nonnegative function ψ,

$$\mathrm{E}_{\theta_0}\left[\psi_n(Z_1, \ldots, Z_r) \int_C \prod_{i=1}^{n} \frac{f(Z_i, \theta)}{f(Z_i, \theta_0)} d\Pi(\theta)\right] = \int_C \mathrm{E}_{\theta}(\psi_n) d\Pi(\theta), \tag{9}$$

by Fubini's theorem. Let $\psi = 1 - \Phi_n$ and get

$$\begin{aligned}
\mathrm{E}_{\theta_0} I_{1n}(Z_1, \ldots, Z_n) &= \mathrm{E}_{\theta_0}\left[(1 - \Phi_n) \int_{\Theta_n \cap U_n^C} \prod_{i=1}^{n} \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta)\right] \\
&= \int_{\Theta_n \cap U_n^C} \mathrm{E}_{\theta}[(1 - \Phi_n)] \\
&\leq \sup_{\theta \in \Theta_n \cap U_n^C} \mathrm{E}_{\theta}(1 - \Phi_n) \\
&\leq C_1 e^{-c_1 n},
\end{aligned}$$

where the final inequality follows from condition (ii) of (A2). Thus,

$$P_{\theta_0}\left\{(1 - \Phi_n) \int_{\Theta_n \cap U_n^C} \prod_{i=1}^{n} \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) \geq e^{-c_1 \frac{n}{2}}\right\} \leq C_1 e^{c_1 \frac{n}{2}} e^{-c_1 n} = C_1 e^{-c_1 \frac{n}{2}}$$

An application of the first Borel-Cantelli Lemma yields

$$(1 - \Phi_n) \int_{\Theta_n \cap U_n^C} \prod_{i=1}^{n} \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta) \leq e^{-c_1 \frac{n}{2}}$$

all but finitely often with P_{θ_0} probability 1. Therefore,

$$e^{c_1 \frac{n}{4}} I_{1n} \to 0 \quad \text{a.s.}[P_{\theta_0}]$$

Next, we prove (7). Applying (9) and condition (iii) of (A2), we get

$$\begin{aligned}
\mathrm{E}_{\theta_0} I_{2n}(Z_1,\ldots,Z_n) &= \mathrm{E}_{\theta_0}\left[\int_{U_n^C\cap\Theta_n^C}\prod_{i=1}^n \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}d\Pi(\theta)\right] \\
&\le \Pi(\Theta_n^C) \\
&\le C_2 e^{-c_2 n},
\end{aligned}$$

Again, the first Borel-Cantelli Lemma implies

$$e^{c_2\frac{n}{4}} I_{2n} \to 0 \quad \text{a.s.}[P_{\theta_0}].$$

Next, we prove (8). Define $\log_+(x) = \max\{0, \log(x)\}$ and $\log_-(x) = -\min\{0, \log(x)\}$. Also, define

$$\begin{aligned}
W_i &= \log_+ \frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}, \\
K_i^+(\theta_0,\theta) &= \int f_i(z,\theta_0)\log_+\frac{f_i(z,\theta_0)}{f_i(z,\theta)}dz, \\
K_i^-(\theta_0,\theta) &= \int f_i(z,\theta_0)\log_-\frac{f_i(z,\theta_0)}{f_i(z,\theta)}dz.
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}_{\theta_0}(W_i) &= \mathrm{E}(W_i^2) - \{K_i^+(\theta_0,\theta)\}^2 \\
&\le \mathrm{E}(W_i^2) - \{K_i^+(\theta_0,\theta) - K_i^-(\theta_0,\theta)\}^2 \\
&= \mathrm{E}(W_i^2) - \{K_i(\theta_0,\theta)\}^2 \\
&\le \int f_i(Z_i,\theta_0)\left(\log_+\frac{f_i(Z_i,\theta_0)}{f_i(Z_i,\theta)}\right)^2 + \int f_i(Z_i,\theta_0)\left(\log_-\frac{f_i(Z_i,\theta_0)}{f_i(Z_i,\theta)}\right)^2 - \{K_i(\theta_0,\theta)\}^2 \\
&= \int f_i(Z_i,\theta_0)\left(\log_+\frac{f_i(Z_i,\theta_0)}{f_i(Z_i,\theta)} - \log_-\frac{f_i(Z_i,\theta_0)}{f_i(Z_i,\theta)}\right)^2 - \{K_i(\theta_0,\theta)\}^2 \\
&= V_i(\theta_0,\theta),
\end{aligned}$$

where the next-to-last equality follows from the fact that $\log_+(x)\log_-(x) = 0$ for all x. It follows that $\sum_{i=1}^\infty \mathrm{Var}_{\theta_0}(W_i)/i^2 < \infty$ for all $\theta \in B$, the set define in condition (A1).

According to Kolmogorov's strong law of large numbers for independent non-identically distributed random variables,

$$\frac{1}{n}\sum_{i=1}^n \left(W_i - K_i^+(\theta_0,\theta)\right) \to 0, \quad \text{a.s.}[P_{\theta_0}]. \tag{10}$$

For each $\theta \in B$, with P_{θ_0} probability 1

$$\begin{aligned}
\liminf_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n \log\frac{f_i(Z_i,\theta)}{f_i(Z_i,\theta_0)}\right) &\ge -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n \log_+\frac{f_i(Z_i,\theta_0)}{f_i(Z_i,\theta)}\right) \\
&= -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n K_i^+(\theta_0,\theta)\right) \\
&\ge -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n K_i(\theta_0,\theta) + \frac{1}{n}\sum_{i=1}^n\sqrt{K_i(\theta_0,\theta)/2}\right) \\
&\ge -\limsup_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n K_i(\theta_0,\theta) + \sqrt{\frac{1}{n}\sum_{i=1}^n K_i(\theta_0,\theta)/2}\right),
\end{aligned}$$

where the second line follows from (10), the third line follows from Amewou-Atisso et al. (2003, Lemma A.1), and the fourth follows from Jensen's inequality.

Let $\beta > 0$, and choose ϵ so that $\epsilon + \sqrt{\epsilon/2} \leq \beta/8$. Let $C = B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon \text{ for all } i\}$. For $\theta \in C$, $n^{-1} \sum_{i=1}^n K_i(\theta_0, \theta) < \epsilon$, so for each $\theta \in C$,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \log \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} \geq -(\epsilon + \sqrt{\epsilon/2}).$$

Now,

$$I_{3n} \geq \int_C \prod_{i=1}^n \frac{f_i(Z_i, \theta)}{f_i(Z_i, \theta_0)} d\Pi(\theta),$$

it follows from Fatou's lemma that

$$e^{n\beta/4} I_{3n} \to \infty, \text{a.s.}[P_{\theta_0}], \text{ for all } \beta > 0.$$

A.3. Proof of Lemma 1. Clearly, $d_Q(f,g) = d_Q(g,f)$, $d_Q(f,g) \geq 0$, and $d_Q(f,f) = 0$. If $d_Q(f,g) = 0$ then $f = g$ a.s. $[Q]$. All that remains for the proof that d_Q is a metric is to verify the triangle inequality.

For each $f, g \in \mathcal{F}$, define $B_{f,g} = \{\epsilon : Q(\{x : |f(x) - g(x)| > \epsilon\}) < \epsilon\}$. Then $d_Q(f,g) = \inf B_{f,g}$. We need to verify

$$\inf B_{f,g} \leq \inf B_{f,h} + \inf B_{h,g}. \tag{11}$$

We will show that if $\epsilon_1 \in B_{f,h}$ and $\epsilon_2 \in B_{h,g}$ then $\epsilon_1 + \epsilon_2 \in B_{f,g}$, which implies (11). Let $\epsilon_1 \in B_{f,h}$ and $\epsilon_2 \in B_{h,g}$. Then

$$\{x : |f(x) - g(x)| > \epsilon_1 + \epsilon_2\} \subseteq \{x : |f(x) - h(x)| > \epsilon_1\} \bigcup \{x : |h(x) - g(x)| > \epsilon_2\}.$$

It follows that

$$\begin{aligned} Q\left(\{x : |f(x) - g(x)| > \epsilon_1 + \epsilon_2\}\right) &\leq Q\left(\{x : |f(x) - h(x)| > \epsilon_1\}\right) + Q\left(\{x : |h(x) - g(x)| > \epsilon_2\}\right) \\ &\leq \epsilon_1 + \epsilon_2. \end{aligned}$$

Hence $\epsilon_1 + \epsilon_2 \in B_{f,g}$.

To prove the equivalence of d_Q convergence and convergence in probability, assume that X has distribution Q. First, assume that $\eta_n(X)$ converges to $\eta(X)$ in probability. Then, for every $\epsilon > 0$, $\lim_{n\to\infty} Q(\{x : |\eta_n(x) - \eta(x)| > \epsilon\}) = 0$. So, for every $\epsilon > 0$ there exists N such that for all $n \geq N$, $Q(\{x : |\eta_n(x) - \eta(x)| > \epsilon\}) < \epsilon$. In other words, for every $\epsilon > 0$, there exists N such that for all $n \geq N$, $d_Q(\eta_n, \eta) \leq \epsilon$. This is what it means to say $\lim_{n\to\infty} d_Q(\eta_n, \eta) = 0$. Finally, assume that $\lim_{n\to\infty} d_Q(\eta_n, \eta) = 0$. Then, for every $\epsilon > 0$ there exists N such that for all $n \geq N$, $d_Q(\eta_n, \eta) \leq \epsilon$, which is equivalent to $Q(\{x : |\eta_n(x) - \eta(x)| > \epsilon\}) < \epsilon$. Hence, $\eta_n(X)$ converges to $\eta(X)$ in probability.

A.4. Prior positivity conditions. In this section, we state and prove those results that allows us to verify condition (A1) of Theorem 1.

LEMMA 2. *Let $\epsilon > 0$ and define*

$$B = \left\{ (\eta, \sigma) \ : \ \|\eta - \eta_0\|_\infty < \epsilon \ , \left| \frac{\sigma}{\sigma_0} - 1 \right| < \epsilon \right\}.$$

Then

(i) *For all* $\epsilon > 0$, $K_i(\theta_0, \theta) < \epsilon$ *for all* i

(ii) $\displaystyle\sum_{i=1}^{\infty} \frac{V_i(\theta_0, \theta)}{i^2} < \infty, \ \forall\ \theta \in B,$

PROOF. We break the proof into two main parts, and each main part is split into two subparts. The main parts correspond to the noise distribution. First, we deal with normal noise and later with Laplace noise. The subparts deal with the nonrandom and random designs separately.

1. If $Y_i \sim N(\eta_0(x_i), \sigma_0^2)$

   (a) Nonrandom design:

$$\begin{aligned}
K_i(\theta_0; \theta) &= \mathrm{E}_{\theta_0}(\Lambda(\theta_0; \theta)) \\
&= \mathrm{E}_{\theta_0} \log \frac{f_i(Z_i; \theta_0)}{f_i(Z_i; \theta)} \\
&= \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2} + \mathrm{E}_{\theta_0}\left[-\frac{1}{2}\frac{(Y_i - \eta_0(x_i))^2}{\sigma_0^2}\right] - \mathrm{E}_{\theta_0}\left[-\frac{1}{2}\frac{(Y_i - \eta(x_i))^2}{\sigma^2}\right] \\
&= \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\frac{[\eta_0(x_i) - \eta(x_i)]^2}{\sigma^2}.
\end{aligned}$$

   It follows from the assumptions of Lemma 2 that, for all i,

$$\begin{aligned}
K_i(\theta_0; \theta) &\leq \log \frac{\sigma}{\sigma_0} + \frac{1}{2}\frac{(\sigma^2 - \sigma_0^2)}{\sigma_0^2}\left|\frac{\sigma_0^2}{\sigma^2}\right| + \frac{1}{2}\frac{\|\eta_0 - \eta\|_\infty^2}{\sigma_0^2}\left|\frac{\sigma_0^2}{\sigma^2}\right| \\
&\leq C_0\epsilon, \ \text{where } C_0 \text{ is some constant.}
\end{aligned}$$

   Let $Z = [Y_i - \eta_0(x_i)]/\sigma_0$, which has standard normal distribution. Then

$$\begin{aligned}
V_i(\theta_0; \theta) &= \mathrm{Var}_{\theta_0}(\Lambda(\theta_0; \theta)) \\
&= \mathrm{Var}_{\theta_0}\left(-\left[\frac{(Y_i - \eta_0(x_i))^2}{2\sigma_0^2}\right] + \frac{1}{2}\left[\frac{\sigma_0}{\sigma}\frac{Y_i - \eta_0(x_i) + \eta_0(x_i) - \eta(x)}{\sigma_0}\right]^2\right) \\
&= \mathrm{Var}\left(\left[-\frac{1}{2} + \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\right]Z^2 + \frac{\sigma_0^2}{\sigma^2}\left[\eta(x_i) - \eta_0(x_i)\right]Z\right) \\
&= \left[-\frac{1}{2} + \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\right]^2 \mathrm{Var}(Z^2) + \left[\frac{\sigma_0^2}{\sigma^2}\left[\eta(x_i) - \eta_0(x_i)\right]\right]^2 \mathrm{Var}(Z), \\
&= 2 \cdot \left[-\frac{1}{2} + \frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\right]^2 + \left[\frac{\sigma_0^2}{\sigma^2}\left[\eta(x_i) - \eta_0(x_i)\right]\right]^2 \\
&< \infty, \ \text{uniformly in } i.
\end{aligned}$$

   (b) Random design:

$$\begin{aligned}
K_i(\theta_0; \theta) &= \mathrm{E}\left(\mathrm{E}_{\theta_0}(\Lambda(\theta_0; \theta))|X_i\right) \\
&= \frac{1}{2} \log \frac{\sigma^2}{\sigma_0^2} - \frac{1}{2}\left(1 - \frac{\sigma_0^2}{\sigma^2}\right) + \frac{1}{2}\int \frac{[\eta_0(x_i) - \eta(x_i)]^2}{\sigma^2} dQ. \\
&\leq C_0\epsilon, \ \text{where } C_0 \text{ is some constant.}
\end{aligned}$$

$$
\begin{aligned}
V_i(\theta_0;\theta) &= \mathrm{E}\left[\mathrm{Var}_{\theta_0}(\Lambda(\theta_0;\theta))|X_i\right] + \mathrm{Var}\left[\mathrm{E}_{\theta_0}(\Lambda(\theta_0;\theta))|X_i\right] \\
&= 2\cdot\left[-\frac{1}{2}+\frac{1}{2}\frac{\sigma_0^2}{\sigma^2}\right]^2 + \int\left[\frac{\sigma_0^2}{\sigma^2}\left[\eta(x_i)-\eta_0(x_i)\right]\right]^2 dQ \\
&< \infty, \text{ uniformly in } i.
\end{aligned}
$$

2. If $Y_i \sim DE(\eta_0(x_i),\sigma_0)$, similar calculation verifies

   (a) Nonrandom design:

$$
\begin{aligned}
K_{i,n}(\theta_0;\theta) &= \log\frac{\sigma}{\sigma_0} + \mathrm{E}_{\theta_0}\left[-\frac{|y-\eta_0(x_i)|}{\sigma_0}\right] - \mathrm{E}_{\theta_0}\left[-\frac{|y-\eta(x_i)|}{\sigma}\right] \\
&\leq \log\frac{\sigma}{\sigma_0} + \left|\frac{\sigma_0}{\sigma}\right|\left|1-\frac{\sigma}{\sigma_0}\right| + \left|\frac{\sigma_0}{\sigma}\right|\left(\frac{\|\eta_0(x_i)-\eta(x_i)\|_\infty}{\sigma_0}\right) \\
&\leq C_0'\epsilon, \text{ where } C_0' \text{ is some constant.}
\end{aligned}
$$

$$
\begin{aligned}
V_{i,n}(\theta_0;\theta) &= \mathrm{Var}_{\theta_0}\left(-\left|\frac{y-\eta_0(x_i)}{\sigma_0}\right| + \left|\frac{y-\eta(x_i)}{\sigma}\right|\right) \\
&\leq \mathrm{E}_{\theta_0}\left(\left|\frac{y-\eta_0(x_i)}{\sigma_0}\right|^2\right) + \mathrm{E}_{\theta_0}\left(\left|\frac{y-\eta(x_i)}{\sigma}\right|^2\right) \\
&\leq 2 + \frac{\sigma_0^2}{\sigma^2}\left(1+\frac{|\eta_0(x_i)-\eta(x_i)|^2}{\sigma_0^2} + 2\frac{|\eta_0(x_i)-\eta(x_i)|}{\sigma_0^2}\right) \\
&< \infty,, \text{ uniformly in } i.
\end{aligned}
$$

   It follows that $\displaystyle\sum_{i=1}^{\infty}\frac{V_i(\theta_0;\theta)}{i^2} < \infty.$

   (b) Random design:

$$
\begin{aligned}
K_{i,n}(\theta_0;\theta) &\leq \log\frac{\sigma}{\sigma_0} + \left|\frac{\sigma_0}{\sigma}\right|\left|1-\frac{\sigma}{\sigma_0}\right| + \left|\frac{\sigma_0}{\sigma}\right|\left(\frac{\|\eta_0(x_i)-\eta(x_i)\|_\infty}{\sigma_0}\right) \\
&\leq C_0'\epsilon, \text{ where } C_0' \text{ is some constant.}
\end{aligned}
$$

$$
\begin{aligned}
V_{i,n}(\theta_0;\theta) &\leq 2 + \frac{\sigma_0^2}{\sigma^2}\left(1+\int\frac{|\eta_0(x_i)-\eta(x_i)|^2}{\sigma_0^2}dQ + 2\int\frac{|\eta_0(x_i)-\eta(x_i)|}{\sigma_0^2}dQ\right) \\
&< \infty,, \text{ uniformly in } i.
\end{aligned}
$$

□

LEMMA 3. *Let $\epsilon > 0$ and define*

$$
B = \left\{(\eta,\sigma) \ : \ \|\eta-\eta_0\|_\infty < \epsilon \ , \left|\frac{\sigma}{\sigma_0}-1\right| < \epsilon\right\}.
$$

Then, $\Pi(B) > 0$ *under Assumption P.*

PROOF. Under Assumption P, we know that $\nu\left\{\left|\frac{\sigma}{\sigma_0}-1\right|<\epsilon\right\}>0$. Thus, to verify $\Pi(B)>0$, it suffices to show that $\Pi_1(\eta:\|\eta-\eta_0\|_\infty<\epsilon)>0$. The prior distribution of η is $\eta\sim GP(\mu,R)$, where $\|\mu\|_\infty<C_3$ and $\|\mu'\|_\infty<C_4$ for some constants C_3 and C_4.

To show $\Pi_1(\eta:\|\eta-\eta_0\|_\infty<\epsilon)>0$, we follow the same approach as in Choudhuri, Ghosal and Roy (2004b).

Without loss of generality we assume $\mu\equiv 0$. Otherwise we can work with $\eta^*=\eta-\mu$ and $\eta_0^*=\eta_0-\mu$. Because η_0 is a uniformly continuous function on $[0,1]$, there exists δ_0 such that $|\eta_0(s)-\eta_0(t)|<\epsilon/3$ whenever $|s-t|<\delta_0$. Consider an equi-spaced partition $0=s_0<s_1<\ldots<s_k=1$ with $|s_j-s_{j-1}|<\delta_0$ for all j. Define $I_j=[s_{j-1},s_j)$ for $j=1,2,\ldots,k$.

For each $0\le s\le 1$,

$$|\eta(s)-\eta_0(s)|\le|\eta(s)-\eta(s_j)|+|\eta_0(s)-\eta_0(s_j)|+|\eta(s_j)-\eta_0(s_j)|, \tag{12}$$

where s_j is the partition point closest to s. By design, the middle term on the right side of (12) is at most $\epsilon/3$ for all s by choossing δ smaller than some δ_0. Consider the sets

$$\begin{aligned} E &= \left\{\sup_{s\in[0,1]}|\eta(s)-\eta_0(s)|<\epsilon\right\}\\ E_1 &= \left\{\max_{1\le j\le k}\sup_{s\in I_j}|\eta(s)-\eta(s_j)|<\epsilon/3\right\}\\ E_2 &= \left\{\max_{1\le j\le k}|\eta(s_j)-\eta_0(s_j)|<\epsilon/3\right\}\end{aligned}$$

It follows that $E_1\cap E_2\subset E$. Therefore it is enough to show that $\Pi_1(E_1\cap E_2)>0$.

We can write $\Pi_1(E_1\cap E_2)=\Pi_1(E_2)\Pi_1(E_1|E_2)$. Let $u^k=(u(s_0),u(s_1),\ldots,u(s_k))^T$ where $u(s_i)=\eta(s_i)-\eta_0(s_i)$. Then u^k has a multivariate normal distribution with a mean vector $\mathbf{0^{(k)}}$ and a nonsingular covariance matrix Σ_k whose (i,j) element is $R(s_i,s_j)$. Then

$$\Pi_1(E_2)=\Pi_1\left(\max_{0\le j\le k}|u(s_j)|<\frac{\epsilon}{3}\right).$$

Because k-dimensional Lebesgue measure is absolutely continuous with respect to the distribution of u^k and $\{(u_1,\ldots,u_k):\max_j|u_j|<\epsilon/3\}$ has positive Lebesgue measure, it follows that $\Pi(E_2)>0$.

In order to estimate $\Pi_1(E_1|E_2)$ we shall use the sub-Gaussian inequality in van der Vaart and Wellner (1996), (Corollary 2.2.8, page 101). Consider the Gaussian process $w(\cdot)$ whose distribution is the conditional distribution of η given $\eta^k=(\eta(s_0),\eta(s_1),\ldots,\eta(s_k))$. By Assumption P, the intrinsic semimetric for the η process is given by

$$\begin{aligned}\rho(s,t)=\sqrt{\mathrm{Var}(\eta(s)-\eta(t))} &= \sqrt{R(s,s)-R(s,t)-R(t,s)+R(t,t)}\\ &\le \sqrt{|t-s|\{R_{01}(t,\xi_1)-R_{01}(u,\xi_1)\}}\\ &\le \sup_{t_1,t_2\in[0,1]}\sqrt{R_{11}(t_1,t_2)}|s-t|\le C_5|s-t|\end{aligned}$$

using the mean value theorem for the covariance function $R(s,t)$, where $R_{01}(s,t)\equiv\partial R(s,t)/\partial t$ and $0<\xi_1,\xi_2<1$.

Thus, the process $w(\cdot)$ is sub-Gaussian with respect to the distance $d(s,t)\le C_5|s-t|$ because the conditional variance for $w(s)-w(t)$ given η^k is smaller than the variance of $\eta(s)-\eta(t)$. Then by Corollary 2.2.8 of van der Vaart and Wellner (1996), we have

$$\begin{aligned}\Pi_1\left(\max_{1\le j\le k}\sup_{s\in I_j}|w(s)-w(s_j)|>\frac{\epsilon}{3}\right) &\le \frac{3}{\epsilon}\mathrm{E}\left(\max_{1\le j\le k}\sup_{s\in I_j}|w(s)-w(s_j)|\right)\\ &\le \frac{3C_6}{\epsilon}\int_0^\delta\sqrt{\log\frac{\delta}{u}}du \le \frac{C_7\delta}{\epsilon}\end{aligned}$$

for some constant C_6 and C_7. We can choose δ such that $1-C_7\delta/\epsilon > 1/2$. Therefore,

$$\begin{aligned}\Pi_1(E_1|E_2) &= \Pi_1\left(\max_{1\le j\le k}\sup_{s\in I_j}|\eta(s)-\eta(s_j)|<\epsilon/3\,\middle|\,E_2\right)\\ &= \int\ldots\int_{E_2}\Pi_1\left(\max_{1\le j\le k}\sup_{s\in I_j}|\eta(s)-\eta(s_j)|<\epsilon/3\,\middle|\,\eta^{(k)}\right)d\Pi_1(\eta^{(k)})\\ &\ge \left(1-\frac{C_7\delta}{\epsilon}\right)>0.\end{aligned}$$

It follows that $\Pi_1(E_1\cap E_2)>0$, hence $\Pi_1(E)>0$. $\square$

A simple corollary to Lemma 3 is that, in Assumption B, $\Pi_1'(\Omega)>0$, so that Π_1 is well-defined. Also, it is clear that Π_1 in Assumption B also satisfies the conclusion of Lemma 3.

A.5. Constructing the sieve. To verify (A2) of Theorem 1, we first construct a sieve and then construct a test for each element of the sieve.

Let $M_n = O(n^{1/2})$, and define $\Theta_n = \Theta_{1n}\times \mathbb{R}^+$, where

$$\Theta_{1n} = \{\eta \ :\ \|\eta\|_\infty < M_n\ ,\ \ \|\eta'\|_\infty < M_n\}.$$

The nth test is constructed by combining a collection of tests, one for each of finitely many elements of Θ_n. Those finitely many elements come from a covering of Θ_{1n} by small balls. The following lemma is straightforward from Theorem 2.7.1 of van der Vaart and Wellner (1996).

Lemma 4. *The ϵ-covering number $N(\epsilon,\Theta_{1n},\|\cdot\|_\infty)$ of Θ_{1n} in the supremum norm satisfies*

$$\log N(\epsilon,\Theta_{1n},\|\cdot\|_\infty) \le \frac{K_4M_n}{\epsilon}.$$

Proof. The proof follows from Theorem 2.7.1. of van der Vaart and Wellner (1996) or from Lemma 2.3 of van de Geer (2000). We choose the former approach.

According to Theorem 2.7.1. of van der Vaart and Wellner (1996), let $\mathcal{X}$ be a bounded convex subset of $\mathbb{R}$ with nonempty interior and let $C_1^1(\mathcal{X})$ be the set of all continuous functions $f:\mathcal{X}\mapsto\mathbb{R}$ with $\|f\|_1 \equiv \sup_x|f(x)| + \sup_{x,y}\frac{|f(x)-f(y)|}{|x-y|} \le 1$. Then, there exists a constant K such that

$$\log N(\epsilon, C_1^1(\mathcal{X}),\|\cdot\|_\infty) \le K\lambda(\mathcal{X})\left(\frac{1}{\epsilon}\right),$$

for every $\epsilon>0$, where $\lambda(\mathcal{X})$ is the Lebesgue measure of the set $\{x : \|x-\mathcal{X}\|<1\}$.

For the proof of Lemma 4, we replace $f(x)$ with $\eta(x)/2M_n$ when $\eta(x) \in \Theta_n$, then

$$\begin{aligned}
\|f\|_1 &= \sup_x |f(x)| + \sup_{x,y} \frac{|f(x) - f(y)|}{|x-y|} \\
&= \sup_x \left|\frac{\eta(x)}{2M_n}\right| + \sup_{x,y} \frac{|\eta(x) - \eta(y)|}{2M_n|x-y|} \\
&\leq \frac{1}{2} + \frac{\sup_x |\eta'(x)|}{2M_n} \leq 1
\end{aligned}$$

In addition, the ϵ-covering number for η is identical to the $\epsilon/2M_n$-covering number for f and $\mathcal{X}$ is a interval of $[0,1]$.

Therefore,

$$\log N(\epsilon, \Theta_n, \|\cdot\|_\infty) \leq K' \left(\frac{M_n}{\epsilon}\right),$$

with a constant $K' > 0$. $\square$

For the proof of subcondition (iii) of (A2), we make use of the assumed smoothness in Assumption P. Lemma 5 below shows that under Assumption P, the sample paths of the Gaussian process are almost surely continuously differentiable and the first derivative process is also Gaussian. Furthermore, the the probability of being outside of the sieve becomes exponentially small.

LEMMA 5. *Let $\eta(\cdot)$ be a mean zero Gaussian process on [0,1] with a covariance kernel $R(\cdot,\cdot)$ which satisfy Assumption P. Then $\eta(\cdot)$ has continuously differentiable sample paths and the first derivative process $\eta'(\cdot)$ is also a Gaussian process.*

Further, there exist constants A and d such that

$$\Pr\{\sup_{0\leq s\leq 1} |\eta(s)| > M\} \leq A\exp(-dM^2)$$
$$\Pr\{\sup_{0\leq s\leq 1} |\eta'(s)| > M\} \leq A\exp(-dM^2)$$

PROOF. First, we show that the process has continuously differentiable sample paths. By Section 9.4 of Cramer and Leadbetter (1967), the sample derivative $\eta'(t)$ is continuous with probability one if $\Delta_h\Delta_h R_{11}(t,t) \leq \dfrac{C}{|\log|h||^a}, \quad a > 3$, where

$$\Delta_h\Delta_h R_{11} \equiv R_{11}(t+h, t+h) - R_{11}(t+h, t) - R_{11}(t, t+h) + R_{11}(t,t)$$

and

$$R_{11}(s,t) \equiv \partial^2 R(s,t)/\partial s\partial t.$$

Under the Assumption P, there exists a constant K_2 such that

$$\Delta_h\Delta_h R_{11}(t,t) \leq K_2|h|^2,$$

because

$$\begin{aligned}
\Delta_h\Delta_h R_{11} &= R_{11}(t+h, t+h) - R_{11}(t+h, t) - R_{11}(t, t+h) + R_{11}(t,t) \\
&= \sup_{(t,t)\in T^2} |R_{22}(t,t)|h^2
\end{aligned}$$

where

$$R_{22}(s,t) \equiv \partial^4 R(s,t)/\partial^2 s \partial^2 t.$$

Further, if $\Delta_h \Delta_h R_{11}(t,t) \leq K_2|h|^2$, then $\Delta_h \Delta_h R_{11}(t,t) \leq \dfrac{C}{|\log|h||^a}$, $a > 3$ because $h^2 = O(1/|\log|h||^a)$ for every a.

Secondly, the limit of a sequence of multivariate normal vectors is again a multivariate normal if and only if the means and covariance matrices converge,[1] and $\eta'(t) = \lim_{h\to 0}(\eta(t+h) - \eta(t))/h$. It follows that $\eta'(\cdot)$ is again a Gaussian process because the covariance kernel $R(\cdot,\cdot)$ is four times continuously differentiable.

Moreover, $\mathrm{E}(\eta'(t) - \eta'(s))^2$ may be obtained as $\mathrm{E}\left(\lim_{h\to 0} \dfrac{\eta(s+h) - \eta(s) - \eta(t+h) + \eta(t)}{h}\right)^2$. This follows by the uniform integrability of $(\eta(t+h) - \eta(t))^2/h^2$, which is a consequence of the fact that

$$\mathrm{E}\left(\frac{\eta(t+h) - \eta(t)}{h}\right)^4 = \frac{3(\Delta_h\Delta_h R(t,t))^2}{h^4} \leq 3 \sup_{t\in[0,1]} R_{11}(t,t) < \infty,$$

because,

$$\begin{aligned}
\Delta_h \Delta_h R(t,t) &= R(t+h,t+h) - R(t+h,t) - R(t,t+h) + R(t,t) \\
&= h \cdot R_{01}(t+h, t+\xi h) - h \cdot R_{01}(t, t+\xi h) \\
&= h^2 \cdot R_{11}(t+\xi h, t+\xi h)
\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{E}(\eta'(s) - \eta'(t))^2 &= \lim_{h\to 0} \mathrm{E}\{\eta(s+h) - \eta(s) - \eta(t+h) + \eta(t)\}^2/h^2 \\
&= \lim_{h\to 0} [\Delta_h\Delta_h R(s,s) - 2\Delta_h\Delta_h R(s,t) + \Delta_h\Delta_h R(t,t)]/h^2 \\
&= \lim_{h\to 0} \left(\frac{\Delta_h\Delta_h R(s,s) - \Delta_h\Delta_h R(s,t)}{h^2} + \frac{\Delta_h\Delta_h R(t,t) - \Delta_h\Delta_h R(s,t)}{h^2}\right) \\
&= R_{11}(s,s) - R_{11}(s,t) + (R_{11}(t,t) - R_{11}(s,t)) \\
&= \Delta_{s-t}\Delta_{s-t} R_{11}(t,t) \\
&\leq K_2|s-t|^2
\end{aligned}$$

because,

$$\begin{aligned}
\lim_{h\to 0} \frac{\Delta_h\Delta_h R(t,t) - \Delta_h\Delta_h R(s,t)}{h^2} &= \lim_{h\to 0} \left\{ \frac{R(t+h,t+h) - 2R(t,t+h) + R(t,t)}{h^2} \right. \\
&\quad \left. + \frac{-R(t+h,s+h) + R(t+h,s) + R(s,t+h) - R(t,s)}{h^2} \right\} \\
&= \lim_{h\to 0} \left(\frac{R_{10}(t,t+h) - R_{10}(t,t)}{h} - \frac{R_{01}(t+h,s) - R_{01}(t,s)}{h} \right) \\
&= R_{11}(t,t) - R_{11}(t,s)
\end{aligned}$$

Similar calculations show that the variance of $\eta'(s)$ is $R_{11}(s,s)$ for all s. Hence the covariance kernel for $\eta'(\cdot)$ is given by

$$\mathrm{Cov}(\eta'(s), \eta'(t)) = R_{11}(s,t).$$

[1] In the Gaussian case on $\mathbb{R}^n$, the derivative processes are also Gaussian processes and the joint distributions of all of these processes are Gaussian (Adler, 1981, p. 32).

because,

$$\begin{aligned}
\mathrm{E}(\eta'(s))^2 &= \lim_{h\to 0}\mathrm{E}\{\eta(s+h)-\eta(s)\}^2/h^2 \\
&= \lim_{h\to 0}[\Delta_h\Delta_h R(s,s)]/h^2 \\
&= R_{11}(s,s)
\end{aligned}$$

$$\mathrm{E}(\eta'(s)) = \lim_{h\to 0}\mathrm{E}\{\eta(s+h)-\eta(s)\}/h = 0$$

Without loss of generality we can assume the process to have zero mean. Otherwise

$$\begin{aligned}
\Pr(\eta : \|\eta\|_\infty > M) &\leq \Pr(\eta : \|\eta(\cdot)-\mu(\cdot)\|_\infty > M - \|\mu\|_\infty) \\
&\leq \Pr(\eta : \|\eta(\cdot)-\mu(\cdot)\|_\infty > M/2).
\end{aligned}$$

Also without loss of generality $\sigma(0) = 1$. Then for $K > 1$

$$N(\epsilon, [0,1], |\cdot|) \leq K/\epsilon,$$

where N is the ϵ-covering number.

Then by applying Theorem 5.3. of Adler (1990, page 43) and Mill's ratio, we have

$$\begin{aligned}
\Pr(\sup_s |\eta(s)| > M) &\leq 2\Pr(\sup_s \eta(s) > M) \\
&\leq C_\alpha M\Psi(M/\sigma_T) \\
&\leq \exp(-dM^2),
\end{aligned}$$

where C_α is a constant and $\Psi(\cdot) = \int_x^\infty \phi(x)dx$, provided that $\sup_{s\in T}\mathrm{Var}\{\eta(s)\} \equiv \sigma_T^2 < \infty$.[2] □

LEMMA 6. *For a given $\alpha > 0$, there exists a constant K_5 such that if $M_n \geq K_5 n^\alpha$, then $\Pi(\Theta_n^C) \leq C_8\exp(-c_8 n^{2\alpha})$ for some positive constants C_8 and c_8.*

PROOF. Since

$$\begin{aligned}
\Pi(\Theta_n^C) &= \Pi((\Theta_{1n}^C \times R^+) \cup (\Theta_{1n} \times (R^+)^C) \\
&= \Pi((\Theta_{1n}^C \times R^+)) \\
&= \Pi_1(\Theta_{1n}^C) \times \nu(R^+) \\
&= \Pi_1(\Theta_{1n}^C),
\end{aligned}$$

it suffices to show that there exist constants A and d such that

$$\Pr\{\sup_{0\leq s\leq 1} |\eta(s)| > M\} \leq A\exp(-dM^2)$$
$$\Pr\{\sup_{0\leq s\leq 1} |\eta'(s)| > M\} \leq A\exp(-dM^2),$$

which clearly follows from Lemma 5. □

[2]Since sample path of X_T is continuous a.s and T is a compact, the boundedness is achieved.

A.6. Construction of tests. For each n and each ball in the covering of Θ_{1n}, we find a test with small type I and type II error probabilities. Then we combine the tests and show that they satisfy subconditions (i) and (ii) of (A2). The following relatively straightforward result is useful in the construction.

PROPOSITION 1. *(a) Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random variables. If*

$$\Pr(X_i \le a) \le \Pr(Y_i \le a), \quad \forall a \in I\!R$$

then, $\forall c \in I\!R$,

$$\Pr\left(\sum_{i=1}^{n} X_i \le c\right) \le \Pr\left(\sum_{i=1}^{n} Y_i \le c\right).$$

(b) For every random variable X with unimodal distribution symmetric around 0 and every $c \in I\!R$,

$$\Pr(|X| \le x) \ge \Pr(|X + c| \le x).$$

The main part of test construction is contained in Lemma 7. For the random design cases, we first condition on the observed values of the covariate. In Lemma 7, understand all probability statements as conditional on the covariate values $X_1 = x_1, \ldots, X_n = x_n$ in the random design case.

LEMMA 7. *Let η_1 be a continuous function on T and define $\eta_{ij} = \eta_i(x_j)$ for $i = 0, 1$ and $j = 1, \ldots, n$. Let $\epsilon > 0$, and let $r > 0$. Let $c_n = n^{3/7}$. Let $b_j = 1$ if $\eta_{1j} \ge \eta_{0j}$ and -1 otherwise. Let Ψ_{1n} and Ψ_{2n} be respectively the indicators of the following two sets:*

1. If $Y_j \sim N(\eta_{0j}, \sigma_0^2)$

$$\left\{\sum_{j=1}^{n} b_j\left(\frac{Y_j - \eta_{0j}}{\sigma_0}\right) > 2c_n\sqrt{n}\right\}, \text{ and } \left\{\sum_{j=1}^{n} \frac{(Y_j - \eta_{0j})^2}{\sigma_0^2} > n(1+\epsilon) \text{ or } < n(1-\epsilon)\right\},$$

2. If $Y_j \sim DE(\eta_{0j}, \sigma_0)$

$$\left\{\sum_{j=1}^{n} b_j\left(\frac{Y_j - \eta_{0j}}{\sigma_0}\right) > 2c_n\sqrt{n}\right\}, \text{ and } \left\{\sum_{j=1}^{n} \left|\frac{Y_j - \eta_{0j}}{\sigma_0}\right| > n(1+\epsilon) \text{ or } < n(1-\epsilon)\right\},$$

Define

$$\Psi_n[\eta_1, \epsilon] = \Psi_{1n} + \Psi_{2n} - \Psi_{1n}\Psi_{2n}.$$

Then there exists a constant C_3 such that for all η_1 that satisfy

$$\sum_{j=1}^{n} |\eta_{1j} - \eta_{0j}| > rn, \tag{13}$$

$\mathrm{E}_{P_0}(\Psi_n[\eta_1, \epsilon]) < C_3 \exp(-2c_n^2)$. *Also, there exist constants C_4 and C_9 such that for all sufficiently large n and all η and σ satisfying $|\sigma/\sigma_0 - 1| > \epsilon$ and $\|\eta - \eta_1\|_\infty < r/4$,*

$$\mathrm{E}_P(1 - \Psi_n[\eta_1, \epsilon]) \le C_4 \exp(-nC_9\epsilon),$$

where P is the joint distribution of $\{Y_n\}_{n=1}^\infty$ assuming that $\theta = (\eta, \sigma)$.

PROOF.

1. Normal data:

   (1) Type I error:

   $\mathrm{E}_{P_0}(\Psi_n[\eta_1, \epsilon]) \le \mathrm{E}_{P_0}(\Psi_{1n}) + \mathrm{E}_{P_1}(\Psi_{2n})$.

$$\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{1n}) &= P_0 \left\{ \sum_{j=1}^{n} b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n\sqrt{n} \right\} \\
&= P_0 \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} b_j \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right) > 2c_n \right\} \\
&= 1 - \Phi(2c_n) \\
&\le \frac{\phi(2c_n)}{2c_n} \\
&= \frac{1}{2\sqrt{2\pi}} \frac{\exp(-2c_n^2)}{c_n}.
\end{aligned}$$

   Let $W \sim \chi^2_n$. Then, for all $0 < t_1 < 1/2$ and $t_2 < 0$,

$$\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{2n}) &= P_0 \left(\sum_{j=1}^{n} \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right)^2 > n(1+\epsilon) \right) + P_0 \left(\sum_{j=1}^{n} \left(\frac{Y_j - \eta_{0j}}{\sigma_0} \right)^2 < n(1-\epsilon) \right) \\
&= \Pr\left(W > n(1+\epsilon)\right) + \Pr\left(W < n(1-\epsilon)\right) \\
&\le \exp\left(-n(1+\epsilon)t_1\right) \mathrm{E}\left(\exp(t_1 W)\right) + \exp\left(-n(1-\epsilon)t_2\right) \mathrm{E}\left(\exp(t_2 W)\right) \\
&= \exp\left(-n(1+\epsilon)t_1\right) (1-2t_1)^{-n/2} + \exp\left(-n(1-\epsilon)t_2\right) (1-2t_2)^{-n/2}.
\end{aligned}$$

   Take

$$t_1 = \frac{1}{2}\left(1 - \frac{1}{1+\epsilon}\right) \quad \text{and} \quad t_2 = \frac{1}{2}\left(1 - \frac{1}{1-\epsilon}\right).$$

   Then,

$$\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{2n}) &\le \exp\left(-\frac{n\epsilon}{2} + \frac{n}{2}\log\left[1+\epsilon\right]\right) + \exp\left(\frac{n\epsilon}{2} + \frac{n}{2}\log\left[1-\epsilon\right]\right) \\
&\le \exp\left(-n\left[\frac{\epsilon^2}{4} - \frac{\epsilon^3}{6}\right]\right) + \exp\left(-n\frac{\epsilon^2}{4}\right),
\end{aligned}$$

   where the last line follows from the fact that $\log(1+x) \le x - x^2/2 + x^3/3, \quad x > 0$ and $\log(1-x) \le -x - x^2/2, \ x > 0$.

   Therefore, $\mathrm{E}_{P_0}(\Psi_n) \le \exp(-2c_n^2)$ for sufficiently large n.

   (2) Type II error:

   We know that $\mathrm{E}_P(1-\Psi_n[\eta_1, \epsilon]) \le \min\{\mathrm{E}_P(1-\Psi_{1n}), \mathrm{E}_P(1-\Psi_{2n})\}$. Hence, we need only show that at least one of the Type II error probabilities for Ψ_{1n} and Ψ_{2n} is exponentially small. There are three types of alternatives: (i) $\|\eta - \eta_1\|_\infty < r/4$, $\sigma = \sigma_0$, (ii) $\eta = \eta_0$, $|\sigma/\sigma_0 - 1| > \epsilon$ and (iii) $\|\eta - \eta_1\|_\infty < r/4$, $|\sigma/\sigma_0 - 1| > \epsilon$.

First, assume that $\sigma \le (1+\epsilon)\sigma_0$, and n is large enough so that $c_n/\sqrt{n} < r/(4\sigma_0)$. This will handle alternative (i) and part of alternative (iii). Let $\eta_{*j} = \eta(x_j)$ for $j = 1, \ldots, n$. In this case,

$$\begin{aligned}
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) &\le \mathrm{E}_P(1-\Psi_{1n}) \\
&= P\left\{ \sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right) \le 2c_n\sqrt{n} \right\} \\
&= P\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{*j}}{\sigma}\right) + \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{\eta_{*j}-\eta_{1j}}{\sigma}\right) \right. \\
&\qquad \left. + \frac{1}{\sqrt{n}} \sum_{j=1}^n \left|\frac{\eta_{1j}-\eta_{0j}}{\sigma}\right| \le 2c_n \frac{\sigma_0}{\sigma} \right\} \\
&\le P\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{*j}}{\sigma}\right) \le \frac{r\sqrt{n}}{4\sigma} - \frac{r\sqrt{n}}{\sigma} + 2c_n\frac{\sigma_0}{\sigma} \right\} \\
&\le P\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{*j}}{\sigma}\right) \le -\frac{r\sqrt{n}}{4\sigma_0(1+\epsilon)} \right\} \\
&= \Phi\left(-\frac{r\sqrt{n}}{4\sigma_0(1+\epsilon)}\right) \\
&\le \frac{4\sigma_0(1+\epsilon)}{r\sqrt{2\pi n}} \exp\left(-\frac{nr^2}{32\sigma_0^2(1+\epsilon)^2}\right),
\end{aligned}$$

where the last inequality is by Mill's ratio.

For the next case, assume that $\sigma > (1+\epsilon)\sigma_0$. This handles the rest of alternative (iii) and half of alternative (ii). Let $W \sim \chi_n^2$ and let W' have a noncentral χ^2 distribution with n degrees of freedom and noncentrality parameter $\sum_{j=1}^n(\eta_{*j}-\eta_{0j})^2$. Then, for all $t < 0$,

$$\begin{aligned}
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) &\le \mathrm{E}_P(1-\Psi_{2n}) \\
&= P\left\{ n[1-\epsilon] \le \sum_{j=1}^n \left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right)^2 \le n[1+\epsilon] \right\} \\
&\le P\left\{ \sum_{j=1}^n \left(\frac{Y_j-\eta_{0j}}{\sigma_1}\right)^2 \frac{\sigma^2}{\sigma_0^2} \le n[1+\epsilon] \right\} \\
&= \Pr\left(W' \le n\frac{\sigma_0^2}{\sigma^2}[1+\epsilon]\right) \\
&\le \Pr\left(W \le n\frac{\sigma_0^2}{\sigma^2}[1+\epsilon]\right), \\
&\le \Pr\left(W \le \frac{n}{1+\epsilon}\right) \\
&= \Pr\left\{ \exp(Wt) \ge \exp\left(\frac{nt}{1+\epsilon}\right) \right\}
\end{aligned}$$

$$\leq \exp\left(-\frac{nt}{1+\epsilon}\right)(1-2t)^{-n/2}.$$

Let $t = -\epsilon/2$ to get

$$\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) \leq \exp\left(\frac{n}{2}\left[\frac{\epsilon}{1+\epsilon} - \log(1+\epsilon)\right]\right) \leq \exp\left(-n\frac{\epsilon^2-\epsilon^3}{4(1+\epsilon)}\right),$$

where the last inequality follows from the fact that $\log(1+x) > x - x^2/2$.

Finally, assume that $\sigma < (1-\epsilon)\sigma_0$ to handle the rest of alternative (ii). Let W be as in the previous case. Then, for all $t > 0$,

$$\begin{aligned}
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) &\leq \mathrm{E}_P(1-\Psi_{2n}) \\
&= P\left\{n[1-\epsilon] \leq \sum_{j=1}^n \left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right)^2 \leq n[1+\epsilon]\right\} \\
&\leq P\left\{n[1-\epsilon] \leq \sum_{j=1}^n \left(\frac{Y_j-\eta_{0j}}{\sigma_1}\right)^2 \frac{\sigma^2}{\sigma_0^2}\right\} \\
&\leq \Pr\left(n\frac{\sigma_0^2}{\sigma^2}[1-\epsilon] \leq W\right), \\
&\leq \Pr\left(\frac{n}{1-\epsilon} \leq W\right) \\
&= \Pr\left\{\exp(Wt) \geq \exp\left(\frac{nt}{1-\epsilon}\right)\right\} \\
&\leq \exp\left(-\frac{nt}{1-\epsilon}\right)(1-2t)^{-n/2}.
\end{aligned}$$

Let $t = \epsilon/2$ to get

$$\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) \leq \exp\left(\frac{n}{2}\left[-\frac{\epsilon}{1-\epsilon} - \log(1-\epsilon)\right]\right) \leq \exp\left(-n\frac{\epsilon^2}{2(1-\epsilon)^2}\left\{\frac{3-5\epsilon}{3(1-\epsilon)}\right\}\right),$$

where the last inequality follows from the fact that $\log\left(\frac{1}{1-x}\right) = \log\left(1+\frac{x}{1-x}\right)$ and $\log(1+x) < x - x^2/2 + x^3/3,\ x > 0$.

2. Laplace data:

(1) Type I error:

$$\mathrm{E}_{P_0}(\Psi_n[\eta_1,\epsilon]) \leq \mathrm{E}_{P_0}(\Psi_{1n}) + \mathrm{E}_{P_1}(\Psi_{2n}).$$

$$\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{1n}) &= P_0\left\{\sum_{j=1}^n b_j\left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right) > 2c_n\sqrt{n}\right\} \\
&= P_0\left\{t\cdot\sum_{j=1}^n b_j\left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right) > t\cdot 2c_n\sqrt{n}\right\} \quad 0<t<1
\end{aligned}$$

$$
\begin{aligned}
&\leq \exp\left(-t\cdot 2c_n\sqrt{n}\right)\left(1-t^2\right)^{-n} \\
&= \exp\left(\left[-2t\frac{c_n}{\sqrt{n}}+\log\left(1+\frac{t^2}{1-t^2}\right)\right]\cdot n\right) \\
&\leq \exp\left(\left[-2t\frac{c_n}{\sqrt{n}}+\frac{t^2}{1-t^2}\right]\cdot n\right)
\end{aligned}
$$

Take $t=\dfrac{2c_n}{\sqrt{n}}$. Then,

$$
\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{1n}) &\leq \exp\left(\left[-4\frac{c_n^2}{n}+\frac{\frac{4c_n^2}{n}}{1-\frac{4c_n^2}{n}}\right]\cdot n\right) \\
&\leq \exp\left(\left[-4+\frac{1}{\frac{n}{4c_n^2}-1}\right]\cdot c_n^2\right) \\
&\leq \exp\left(-2c_n^2\right)
\end{aligned}
$$

because it is clear that $\dfrac{4c_n^2}{n}<\dfrac{2}{3}$ if n is large.

Let $V\sim \mathit{Gamma}(n,1)$. Then, for all $0<t_1<1$ and $t_2<0$,

$$
\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{2n}) &= P_0\left(\sum_{j=1}^{n}\left|\frac{Y_j-\eta_{0j}}{\sigma_0}\right|>n\sqrt{1+\epsilon}\right)+P_0\left(\sum_{j=1}^{n}\left|\frac{Y_j-\eta_{0j}}{\sigma_0}\right|<n\sqrt{1-\epsilon}\right) \\
&= \Pr\left(V>n\sqrt{1+\epsilon}\right)+\Pr\left(V<n\sqrt{1-\epsilon}\right) \\
&\leq \exp\left(-n(\sqrt{1+\epsilon})t_1\right)\mathrm{E}\left(\exp(t_1V)\right)+\exp\left(-n(\sqrt{1-\epsilon})t_2\right)\mathrm{E}\left(\exp(t_2V)\right) \\
&= \exp\left(-n(\sqrt{1+\epsilon})t_1\right)(1-t_1)^{-n}+\exp\left(-n(\sqrt{1-\epsilon})t_2\right)(1-t_2)^{-n}.
\end{aligned}
$$

Take

$$
t_1=1-\frac{1}{\sqrt{1+\epsilon}}\quad\text{and}\quad t_2=1-\frac{1}{\sqrt{1-\epsilon}}.
$$

Then,

$$
\begin{aligned}
\mathrm{E}_{P_0}(\Psi_{2n}) &\leq \exp\left(-n(\sqrt{1+\epsilon}-1)+n\log\left[1+\sqrt{1+\epsilon}-1\right]\right) \\
&\quad +\exp\left(n(1-\sqrt{1-\epsilon})+n\log\left[1-(1-\sqrt{1-\epsilon}\right]\right) \\
&\leq \exp\left(-n\left[\frac{(\sqrt{1+\epsilon}-1)^2}{2}-\frac{(\sqrt{1+\epsilon}-1)^3}{3}\right]\right)+\exp\left(-n\frac{(1-\sqrt{1-\epsilon})^2}{2}\right),
\end{aligned}
$$

where the last line follows from the fact that $\log(1+x)\leq x-x^2/2+x^3/3,\quad x>0$ and $\log(1-x)\leq -x-x^2/2,\ x>0$.

Therefore, $\mathrm{E}_{P_0}(\Psi_n)\leq C_3\exp(-2c_n^2)$ for sufficiently large n.

(2) Type II error:

Again, there are three types of alternatives to deal with : (i) $\|\eta-\eta_1\|_\infty<r/4$, $\sigma=\sigma_0$, (ii) $\eta=\eta_0$, $|\sigma/\sigma_0-1|>\epsilon$ and (iii) $\|\eta-\eta_1\|_\infty<r/4$, $|\sigma/\sigma_0-1|>\epsilon$.

As in the previous Type II error calculation for normal case, first, assume that $\sigma \le (1+\epsilon)\sigma_0$, and n is large enough so that $c_n/\sqrt{n} < r/(4\sigma_0)$. This handles alternative (i) and part of alternative (iii). Let $\eta_{*j} = \eta(x_j)$ for $j = 1, \dots, n$. In this case,

$$
\begin{aligned}
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) &\le \mathrm{E}_P(1-\Psi_{1n}) \\
&= P\left\{\sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{0j}}{\sigma_0}\right) \le 2c_n\sqrt{n}\right\} \\
&= P\left\{\sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{*j}}{\sigma_1}\right) + \sum_{j=1}^n b_j \left(\frac{\eta_{*j}-\eta_{1j}}{\sigma}\right)\right. \\
&\quad \left. + \sum_{j=1}^n \left|\frac{\eta_{1j}-\eta_{0j}}{\sigma}\right| \le 2c_n\sqrt{n}\frac{\sigma_0}{\sigma}\right\} \\
&\le P\left\{\sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{*j}}{\sigma}\right) \le \frac{rn}{4\sigma} - \frac{rn}{\sigma} + 2c_n\sqrt{n}\frac{\sigma_0}{\sigma}\right\} \\
&\le P\left\{\sum_{j=1}^n b_j \left(\frac{Y_j-\eta_{*j}}{\sigma}\right) \le \frac{-rn}{4\sigma_0(1+\epsilon)}\right\} \\
&= \exp\left(\left[-t\frac{r}{\sigma_0(1+\epsilon)} - \log(1-t^2)\right]\cdot n\right), \quad \text{for some } t,\ 0<t<1 \\
&\le \exp(-\xi\cdot n), \quad \exists\, \xi(=C_9\epsilon) > 0
\end{aligned}
$$

The last inequality is established by the following argument.
Let $c > 0$ and $f(t) = -t\cdot c - \log(1-t^2)$, $0 < t < 1$. Then,

$$
\begin{aligned}
\Rightarrow\ f'(t) &= -c + \frac{2t}{1-t^2} \\
\mathit{Set}\ f'(t) &= 0 \\
\Rightarrow\ t^* &= \frac{-1\pm\sqrt{1+c^2}}{c}, \quad 0 < t^* < 1 \\
\Rightarrow\ f(t^*) &= 1-\sqrt{1+c^2} + \log\frac{c^2}{2(-1+\sqrt{1+c^2})} \\
&= 1-\sqrt{1+c^2} + \log(\sqrt{1+c^2}+1) - \log 2
\end{aligned}
$$

Let $g(x) = 1 - x + \log(x+1) - \log 2$, $x > 1$. Then,

$$
g'(x) = -1 + \frac{1}{x+1} = \frac{-x}{1+x} < 0
$$

$$
\Rightarrow\ g(1) = 0 \ \Rightarrow\ g(x) < 0, \quad x > 1
$$

Therefore, $f(t)$ can have negative values.

For the next case, assume that $\sigma > (1+\epsilon)\sigma_0$. This handles the rest of alternative (iii) and half of alternative (ii). Let $V \sim \mathit{Gamma}(n,1)$

$$
\begin{aligned}
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) &\le \mathrm{E}_P(1-\Psi_{2n}) \\
&= P\left\{ n[1-\epsilon] \le \sum_{j=1}^{n} \left| \frac{Y_j-\eta_{0j}}{\sigma_0} \right| \le n\sqrt{1+\epsilon} \right\} \\
&\le P\left\{ \sum_{j=1}^{n} \left| \frac{Y_j-\eta_{0j}}{\sigma} \right| \frac{\sigma}{\sigma_0} \le n\sqrt{1+\epsilon} \right\} \\
&= P\left\{ \sum_{j=1}^{n} \left| \left(\frac{Y_j-\eta_{*j}}{\sigma}\right) + \left(\frac{\eta_{*j}-\eta_{0j}}{\sigma}\right) \right| \le n\frac{\sigma_0}{\sigma}\sqrt{1+\epsilon} \right\} \\
&\le P\left\{ \sum_{j=1}^{n} \left| \frac{Y_j-\eta_{*j}}{\sigma} \right| \le n\frac{\sigma_0}{\sigma}\sqrt{1+\epsilon} \right\} \qquad (14) \\
&= \Pr\left\{ V \le n\frac{\sigma_0}{\sigma}\sqrt{1+\epsilon} \right\} \\
&\le \Pr\left\{ \exp(Vt) \ge \exp\left(\frac{nt}{\sqrt{1+\epsilon}}\right) \right\}, \quad t<0,\ \sigma>(1+\epsilon)\sigma_0 \\
&\le \exp\left(-\frac{nt}{\sqrt{1+\epsilon}}\right)(1-t)^{-n}
\end{aligned}
$$

The inequality (14) follows from Proposition 1.
Finally, let $t = 1-\sqrt{1+\epsilon}$ to get

$$
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) \le \exp\left(n\left[1 - \frac{1}{\sqrt{1+\epsilon}} - \log\left(1+\sqrt{1+\epsilon}-1\right) \right] \right) = \exp\left(-nC_9\epsilon\right),
$$

where $-C_9\epsilon = \left[1 - \frac{1}{\sqrt{1+\epsilon}} - \log\left(1+\sqrt{1+\epsilon}-1\right) \right] < 0,\ \epsilon > 0$
Finally, assume that $\sigma < (1-\epsilon)\sigma_0$ to handle the rest of alternative (ii). For all $t > 0$,

$$
\begin{aligned}
\mathrm{E}_P(1-\Psi_n[\eta_1,\epsilon]) &\le \mathrm{E}_P(1-\Psi_{2n}) \\
&= P\left\{ n[\sqrt{1-\epsilon}] \le \sum_{j=1}^{n} \left| \frac{Y_j-\eta_{0j}}{\sigma_0} \right| \le n[\sqrt{1+\epsilon}] \right\} \\
&\le P\left\{ n[\sqrt{1-\epsilon}] \le \sum_{j=1}^{n} \left| \frac{Y_j-\eta_{0j}}{\sigma} \right| \frac{\sigma}{\sigma_0} \right\} \\
&\le P_1\left\{ n[\sqrt{1-\epsilon}]\frac{\sigma_0}{\sigma} \le \sum_{j=1}^{n} \left| \frac{Y_j-\eta_{0j}}{\sigma} \right| \right\} \\
&\le \Pr\left\{ \frac{n}{\sqrt{1-\epsilon}} \le V \right\} \\
&= \Pr\left\{ \exp(Vt) \ge \exp\left(\frac{nt}{\sqrt{1-\epsilon}}\right) \right\} \\
&\le \exp\left(-\frac{nt}{\sqrt{1-\epsilon}}\right)(1-t)^{-n}
\end{aligned}
$$

Let $t = 1 - \sqrt{1-\epsilon}$ to get

$$\mathrm{E}_P(1 - \Psi_n[\eta_1, \epsilon]) \le \exp\left(n \left[1 - \frac{1}{\sqrt{1-\epsilon}} - \log(\sqrt{1-\epsilon}) \right] \right) = \exp\left(-nC_9\epsilon\right),$$

where $-C_9\epsilon = \left[1 - \frac{1}{\sqrt{1-\epsilon}} - \log(\sqrt{1-\epsilon}) \right] < 0, \ \epsilon > 0.$

□

To create a test that doesn't depend on a specific choice of η_1 in Lemma 7, we make use of the covering number of the sieve. Let r be the same number that appears in Lemma 7. Let $t = \min\{\epsilon/2, r/4\}$. Let N_t be the t covering number of Θ_{1n} in the supremum (L^∞) norm. With $M_n = O(n^{1/2})$ and $c_n = n^{3/7}$, we have $\log(N_t) = o(c_n^2)$ from Lemma 4. Let $\eta^1, \ldots, \eta^{N_t} \in \Theta_{1n}$ be such that for each $\eta \in \Theta_{1n}$ there exists j such that $\|\eta - \eta^j\|_\infty < t$. If $\|\eta - \eta_0\|_1 > \epsilon$, then $\|\eta^j - \eta_0\|_1 > \epsilon/2$. Similarly, if $d_Q(\eta, \eta_0) > \epsilon$, then $d_Q(\eta^j, \eta_0) > \epsilon/2$. Define

$$\Psi_n = \max_{1 \le j \le N_\epsilon} \Psi_n[\eta^j, \epsilon/2].$$

If we can verify that there exists r such that (13) holds for every such η^j, then

$$\begin{aligned}
\mathrm{E}_{P_0}\Psi_n &\le \sum_{j=1}^{N_t} \mathrm{E}_{P_0}\Psi_n[\eta^j, \epsilon/2] \\
&\le C_3 N_t \exp(-2c_n^2) \\
&= \exp C_3(\log[N_t] - 2c_n^2) \\
&\le C_3 \exp(-c_n^2).
\end{aligned}$$

For $\theta = (\eta, \sigma) \in U_\epsilon^C \cap \Theta_n$ or in $W_\epsilon \cap \Theta_n$, the type II error probability of Ψ_n is no larger than the minimum of the individual type II error probabilities of the $\Psi_n[\eta^j, \epsilon/2]$ tests. Hence, we have a uniformly consistent test Ψ_n, which has exponentially small type II error evaluated at θ.

Verifying (13) is done differently for the random and nonrandom design cases.

For the random design case, Lemma 8 tells us that (13) occurs all but finitely often with probability 1. Since there are only finitely many η^j to consider for each n, this suffices to complete the proof of Theorem 3.

Lemma 8. *Assume Assumption RD. Let η be a function such that $d_Q(\eta, \eta_0) > \epsilon$. Let $0 < r < \epsilon^2$, and define*

$$A_n = \left\{ \sum_{i=1}^n |\eta(X_i) - \eta_0(X_i)| \ge rn \right\}.$$

Then there exists $C_{11} > 0$ such that $\Pr(A_n^C) \le \exp(-C_{11}n)$ for all n and A_n occurs all but finitely often with probability 1. The same C_{11} works for all η such that $d_Q(\eta, \eta_0) > \epsilon$.

Proof. Let $B = \{x | \eta(x) - \eta_0(x)| > \epsilon\}$, so that $Q(B) > \epsilon$. Let $Z = n - \sum_{i=1}^n I_B(X_i)$, and notice that Z has a binomial distribution with parameters n and $1 - Q(B)$. Let $q = r/\epsilon < \epsilon$, and let Z' have a binomial distribution with parameters n and $1-\epsilon$ so that Z' stochastically dominates Z. Then

$$\Pr(A_n^C) \le \Pr(Z > n[1-q]) \le \Pr(Z' > n[1-q]).$$

Write

$$\begin{aligned}\Pr(Z' > n[1-q]) &= \Pr(\exp(tZ') > \exp(tn[1-q])), \text{ for all } t > 0, \\ &\leq [\epsilon + [1-\epsilon]\exp(t)]^n \exp(-tn[1-q]).\end{aligned}$$

Let

$$\begin{aligned}t &= \log\left(\frac{\epsilon(1-q)}{q(1-\epsilon)}\right) > 0, \\ C_{11} &= q\log\left(\frac{q}{\epsilon}\right) + (1-q)\log\left(\frac{1-q}{1-\epsilon}\right) > 0.\end{aligned}$$

Then $\Pr(A_n^C) \leq \exp(-C_{11}n)$, and C_{11} doesn't depend on the particular η. The probability one claim follows from the first Borel-Cantelli lemma. $\square$

For the nonrandom design case, we verify (13) for all η_1 that are far from η_0 in L^1 distance.

LEMMA 9. *Assume Assumption NRD. Let λ be Lebesgue measure. Let K_1 be the constant mentioned in Assumption NRD. Let $V > 0$ be a constant. For each integer n, let A_n be the set of all continuously differentiable functions γ such that $\|\gamma'\|_\infty < M_n + V$. For each function γ and $\epsilon > 0$, define $B_{\epsilon,\gamma} = \{x : |\gamma(x)| > \epsilon\}$. Then for each $\epsilon > 0$ there exist an integer N such that, for all $n \geq N$ and all $\gamma \in A_n$,*

$$\sum_{i=1}^{n} |\gamma(x_i)| \geq (\lambda(B_{\epsilon,\gamma})K_1 n - 1)\frac{\epsilon}{2}. \tag{15}$$

PROOF. Let N be large enough so that $(M_n + V)/(K_1 n) < \epsilon/2$ for all $n \geq N$. Because γ is continuous, $B_{\epsilon,\gamma}$ is an open set and it is the union of a countable collection of disjoint open intervals, i.e. $B_{\epsilon,\gamma} = \cup_{i=1}^{\infty} B_i$, where $B_i = (x_{L,i}, x_{R,i})$ is an open interval whose length is $\lambda(B_i) = x_{R,i} - x_{L,i} \geq 0$. Some of the B_i intervals might lie entirely between successive design points. Let $x_0 = 0$ and $x_{n+1} = 1$. Define, for $j = 0, \ldots, n$,

$$\begin{aligned}a_j &= \{i : x_j < x_{R,i} < x_{j+1}\}, \\ b_j &= \{i : x_j < x_{L,i} < x_{j+1}\}, \\ \ell_j &= \inf\{x_{L,i} : i \in a_j\}, \\ u_j &= \sup\{x_{R,i} : i \in b_j\}.\end{aligned}$$

Then the open interval $F_j = (\ell_j, u_j)$ contains the same design points as

$$D_j = \bigcup_{i \in a_j \cup b_j} B_i.$$

If $x_j \in F_j$ (for $j \in \{1, \ldots, n\}$), then $\ell_j < u_{j-1}$ and (ℓ_{j-1}, u_j) contains the same design points as $D_{j-1} \cup D_j$. By combining all of the overlapping F_j intervals, we obtain finitely many disjoint intervals $E_1, \ldots, E_m$ whose total length is at least $\lambda(B_{\gamma,\epsilon})$ (because their union contains $B_{\epsilon,\gamma}$) and that contain the same design points as $B_{\epsilon,\gamma}$. Write each $E_j = (f_j, g_j)$ and $L_j = g_j - f_j$, and assume that the intervals are ordered so that $g_j < f_{j+1}$ for all j. Let $E = \cup_{j=1}^{m} E_j$. Let $\lfloor a \rfloor$ denote the integer part of a. Each E_j contains at least $\lfloor L_j K_1 n \rfloor$ design points because the maximum spacing is assumed to be less than or equal to $1/(K_1 n)$. For each j such that $f_j > x_1$, let x_j^* be the largest $x_i \leq f_j$. Then $x_j^* \notin E$, the the derivative of γ is at most $M_n + V$, and $|x_j^* - f_j| < 1/(K_1 n)$. Hence,

$$|\gamma(x^*)| > \epsilon - \frac{M_n + V}{K_1 n} > \frac{\epsilon}{2}.$$

If we include x_j^* with the design points already in E_j, we have, associated with each j such that $f_j > x_1$, at least $\lceil L_j K_1 n \rceil$ design points x with $|\gamma(x)| > \epsilon/2$, where $\lceil a \rceil$ is the smallest integer greater than or equal to a. There is at most one j such that $f_j \leq x_1$ for which we have at least $\lceil L_j K_1 n \rceil - 1$ design points with $|\gamma(x)| > \epsilon/2$. Since we have not counted any design points more than once, we can add over all j to see that there are at least $\lambda(B_{\gamma,\epsilon})K_1 n - 1$ design points x in $B_{\epsilon/2,\gamma}$ so long as $n \geq N$. Hence we satisfy (15). $\square$

LEMMA 10. *Assume Assumption NRD. For each integer n, let A_n be the set of all continuously differentiable functions η such that $\|\eta\| < M_n$ and $\|\eta'\|_\infty < M_n$. Then for each $\epsilon > 0$ there exist an integer N and $r > 0$ such that, for all $n \geq N$ and all $\eta \in A_n$ such that $\|\eta - \eta_0\|_1 > \epsilon$, $\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| \geq rn$.*

PROOF. Let V be an upper bound on the derivative of η_0. Let $0 < \delta < \epsilon$ and let N be large enough so that $(M_n + V)/n < 2K_1(\epsilon - \delta)$ for all $n \geq N$. Let $r = K_1(\epsilon - \delta)/2$ and $D_i = \{x : (i-1)\delta < |\eta(x) - \eta_0(x)| < i\delta\}$.

Let λ be Lebesgue measure. Then, $\|\eta - \eta_0\|_1$ can be bounded as follows.

$$\sum_i i\delta\lambda(D_i) \geq \|\eta - \eta_0\|_1 > \epsilon. \tag{16}$$

Let $\zeta(x) = |\eta(x) - \eta_0(x)|$ and $\zeta_m(x) = \min\{m\delta, \zeta(x)\}$, for $m = 0, \ldots, n$. Note that $\zeta_{\lceil (M_n+V)/\delta \rceil}(x)$ is the same as $\zeta(x)$.

For $m = 1, \ldots, n$, define $B_m \equiv \{x : \zeta_m(x) > (2m-1)\delta/2\}$. Then, for all $x \in B_m$, $\zeta_m(x) - \zeta_{m-1}(x) > \delta/2$. Thus, Lemma 9 (with $\gamma = \zeta_m - \zeta_{m-1}$ and $\epsilon = \delta/2$) implies

$$\sum_{i=1}^n (\zeta_m(x_i) - \zeta_{m-1}(x_i)) \geq (\lambda(B_m)K_1 n - 1)\frac{\delta}{4}.$$

Now, write

$$\begin{aligned}
\sum_{i=1}^n |\eta(x_i) - \eta_0(x_i)| &= \sum_{i=1}^n \zeta_{\lceil (M_n+V)/\delta \rceil}(x_i) \\
&= \sum_{i=1}^n \sum_{m=1}^{\lceil (M_n+V)/\delta \rceil} \{\zeta_m(x_i) - \zeta_{m-1}(x_i)\} \\
&= \sum_{m=1}^{\lceil (M_n+V)/\delta \rceil} \sum_{i=1}^n \{\zeta_m(x_i) - \zeta_{m-1}(x_i)\}, \\
&\geq \sum_{m=1}^{\lceil (M_n+V)/\delta \rceil} [\lambda(B_m)K_1 n - 1]\frac{\delta}{4}. \qquad (17)
\end{aligned}$$

Also, for $m = 1, \ldots, n$,

$$B_m \supset \bigcup_{i=m+1}^{\lceil (M_n+V)/\delta \rceil} D_i.$$

It follows that

$$\sum_{m=1}^{\lceil (M_n+V)/\delta \rceil} \delta\lambda(B_m) \geq \sum_{i=2}^{\lceil (M_n+V)/\delta \rceil} (i-1)\delta\lambda(D_i)$$

$$
\begin{aligned}
&\geq \sum_{i=2}^{\lceil (M_n+V)/\delta \rceil} i\delta\lambda(D_i) - \delta \\
&\geq \epsilon - \delta,
\end{aligned}
$$

where the last inequality follows from (16). Combining this with (17) gives

$$
\sum_{i=1}^{n} |\eta(x_i) - \eta_0(x_i)| \geq K_1 n(\epsilon - \delta) - \frac{M_n + V}{4} \geq n\frac{K_1(\epsilon - \delta)}{2},
$$

for all $n \geq N$. □

A.7. Proof of Theorem 4. First, we calculate the Hellinger distance between two density functions, $d_H(f, f_0)$, where f is the joint density of (X, Y) when η and σ are arbitrary, and f_0 is the density when $\eta = \eta_0$ and $\sigma = \sigma_0$.

Let ν be the product of Q and Lebesgue measure λ. Then the joint densities of X and Y defined above with respect to ν are given by

$$
f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{[y - \eta(x)]^2}{2\sigma^2} \right\} \quad \textit{and} \quad f_0(y|x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{[y - \eta_0(x)]^2}{2\sigma_0^2} \right\}
$$

or

$$
f(y|x) = \frac{1}{2\sigma} \exp\left\{ -\frac{|y - \eta(x)|}{\sigma} \right\} \quad \textit{and} \quad f_0(y|x) = \frac{1}{2\sigma_0} \exp\left\{ -\frac{|y - \eta_0(x)|}{\sigma_0} \right\}
$$

To simplify the calculation, we consider the quantity $h(f, f_0)$ defined as

$$
h(f, f_0) = \frac{1}{2} d_H^2(f, f_0) = 1 - \int \sqrt{f f_0}\, d\mu
$$

and $h(f, f_0)$ is calculated as follows.

1. $Y_i | X_i \overset{\text{ind}}{\sim} N(\eta(X_i), \sigma^2)$

$$
\begin{aligned}
h(f, f_0) &= 1 - \frac{1}{\sqrt{2\pi\sigma\sigma_0}} \int\int \exp\left\{ -\frac{1}{4\sigma^2}[y - \eta(x)]^2 - \frac{1}{4\sigma_0^2}[y - \eta_0(x)]^2 \right\} dy dQ \\
&= 1 - \int\int \exp\left\{ -\left(\frac{1}{4\sigma^2} + \frac{1}{4\sigma_0^2}\right) \left[y - \left(\frac{\eta(x)}{4\sigma_1^2} + \frac{\eta_0(x)}{4\sigma_0^2}\right) \Big/ \left(\frac{1}{4\sigma^2} + \frac{1}{4\sigma_0^2}\right) \right]^2 \right\} \\
&\qquad \frac{1}{\sqrt{2\pi\sigma\sigma_0}} \times \exp\left\{ -\frac{\eta(x)^2}{4\sigma^2} - \frac{\eta_0(x)^2}{4\sigma_0^2} + \left(\frac{\eta(x)}{4\sigma^2} + \frac{\eta_0(x)}{4\sigma_0^2}\right)^2 \Big/ \left(\frac{1}{4\sigma^2} + \frac{1}{4\sigma_0^2}\right) \right\} dy dQ \\
&= 1 - \int \sqrt{\frac{2\sigma\sigma_0}{\sigma^2 + \sigma_0^2}} \exp\left\{ -\frac{1}{16\sigma^2\sigma_0^2}[\eta(x) - \eta_0(x)]^2 \Big/ \left(\frac{1}{4\sigma^2} + \frac{1}{4\sigma_0^2}\right) \right\} dQ \qquad (18)
\end{aligned}
$$

The integral in (18) is of the form $\int c_1 \exp(-c_2[\eta(x) - \eta_0(x)]^2) dQ(x)$, where c_1 can be made arbitrarily close to 1 by choosing $|\sigma/\sigma_0 - 1|$ small enough and c_2 is bounded when σ is close to σ_0. It follows that for each ϵ there exists a δ such that (18) will be less than $\epsilon^2/2$ whenever $|\sigma/\sigma_0 - 1| < \delta$ and $d_Q(\eta, \eta_0) < \delta$.

2. $Y_i|X_i \stackrel{\text{ind}}{\sim} DE(\eta(X_i), \sigma)$

$$
\begin{aligned}
h(f, f_0) &= 1 - \frac{1}{\sqrt{4\sigma\sigma_0}} \int\int \exp\left\{-\frac{1}{2\sigma}|y-\eta(x)| - \frac{1}{2\sigma_0}|y-\eta_0(x)|\right\} dydQ \\
&\le 1 - \int\int \exp\left\{-\left(\frac{1}{2\sigma}+\frac{1}{2\sigma_0}\right)\left|y - \left(\frac{\eta(x)+\eta_0(x)}{2}\right)\right|\right\} \\
&\quad \frac{1}{\sqrt{4\sigma\sigma_0}} \times \exp\left\{-\left(\frac{1}{4\sigma}+\frac{1}{4\sigma_0}\right)|\eta(x)-\eta_0(x)|\right\} dydQ \\
&\le 1 - \int \left[\frac{1}{\sqrt{4\sigma\sigma_0}} \times \exp\left\{-\left(\frac{1}{4\sigma}+\frac{1}{4\sigma_0}\right)|\eta(x)-\eta_0(x)|\right\} \times \left(\frac{1}{4\sigma}+\frac{1}{4\sigma_0}\right)^{-1}\right] dQ \qquad (19)
\end{aligned}
$$

The integral in (19) is of the form $\int c_1 \exp(-c_2|\eta(x)-\eta_0(x)|)dQ(x)$, where c_1 can be made arbitrarily close to 1 by choosing $|\sigma/\sigma_0 - 1|$ small enough and c_2 is bounded when σ is close to σ_0. It follows that for each ϵ there exists a δ such that (19) will be less than $\epsilon^2/2$ whenever $|\sigma/\sigma_0 - 1| < \delta$ and $d_Q(\eta, \eta_0) < \delta$.

A.8. Proof of Theorem 5. For bounded functions, convergence in probability is equivalent to L^p convergence for all finite p. In particular, for every $\epsilon > 0$ and every finite p, there exists an ϵ' such that $U_{\epsilon'} \subseteq W_\epsilon$. Hence, Theorem 3 implies the conclusion to Theorem 5 as long as the GP prior defined in Assumption B also satisfy all the conditions required in Theorem 3.

If a GP satisfies the smoothness conditions that follow from Assumption P, then the conditional process given a set of bounded functions with positive probability also satisfies the smoothness conditions. We have already verified the prior positivity condition (A1). For subpart (iii) of (A2), we note that, if A and d are the constants guaranteed by Lemma 5 for Π_1', then

$$\Pi_1\left\{\sup_{0\le s\le 1}|\eta'(s)| > M\right\} \le A\exp(-dM^2)/\Pi_1'(\Omega).$$

REFERENCES

Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center.

Adler, R. J. (1981). *The Geometry of Random Fields*, Wiley, New York.

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for Gaussian processes*, IMS Lecture notes-monograph series vol. 12, Hwyward, CA.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli* **9** 291–312.

Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561.

Brown, L. D. and Low, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

Choudhuri, N., Ghosal, S. and Roy, A. (2003). Bayesian methods for function estimation. In *Handbook of Statistics* (D. Dey, ed.). Marcel Dekker, New York.

Choudhuri, N., Ghosal, S. and Roy, A. (2004a). Bayesian estimation of the spectral density of time series. unpublished manuscript. Preprint.

Choudhuri, N., Ghosal, S. and Roy, A. (2004b). Nonparametric binary regression using a Gaussian process prior. unpublished manuscript. Preprint.

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

Cramer, H. and Leadbetter, M. R. (1967). *Stationary and related stochastic processes*, Wiley, New York.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *JRSS B.* **60** 333–350.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika.* **88** 183–201.

Doob, J. L. (1949). Application of the theory of martingales. *Coll. Int. du C. N. R. S. Paris.* 23–27.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist* **27** 1119–1141.

Fuentes, M. and Smith, R. (2001). A new class of nonstationary spatial models. unpublished manuscript. North Carolina State University, Department of Statistics.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Consistency issues in bayesian nonparametrics. In *Asymptotics, Nonparametrics and Time Series : A Tribute to Madan Lal Puri* (S. Ghosh, ed.). Marcel Dekker, New York.

Ghosal, S. and van der Vaart, A. W. (2004). Convergence rates of posterior distributions for noniid observations. unpublished manuscript. Preprint.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*, Sprinber-Verlag, New York.

Higdon, D., Swall, J. and Kern, J. (1998). Non-stationary spatial modeling. In *Bayesian Statistics 6* (J. M. Bernardo, J. Berger, A. Dawid and A. Smith, eds.). Oxford Univ. Press, Oxford, U.K.

Hjort, N. L. (2002). Topics in nonparametric bayesian statistics. In *Highly Structured Stochastic Systems* (P. J. Green, N. L. Hjort and S. Richardson, eds.). Oxford Univ. Press, Oxford.

Huang, T. M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32** accepted.

Kleijn, B. and Van der Vaart, A. W. (2002). Misspecification in infite dimensional Bayesian statistics. unpublished manuscript. Preprint.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*, Sprinber-Verlag, New York.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *JRSS B.* **40** 1–42.

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modeling.* Ph.D. dissertation, Department of Statistics, Carnegie Mellon University.

Paciorek, C. J. and Schervish, M. J. (2004). Nonstationary covariance functions for gaussian process regression. In *Advances in Neural Information Processing Systems 16* (D. S. Touretzky, M. C. Mozer and M. E. Hasselmo, eds.). MIT Press, MA.

Rasmussen, C. (1996). *Evaluation of Gaussian processes and other methods for non-linear regression.* Ph.D. dissertation, Department of Computer Science, University of Toronto.

Schwartz, L. (1965). On Bayes procedures. *Z. Wahr. Verw. Gebiete* **4** 10–26.

Seeger, M. (2004). Bayesian Gaussian processes for machine learning. unpublished manuscript. Preprint.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 666–686.

van de Geer, S. (2000). *Empirical Processes in M-Estimation.* CAMBRIDGE Univ. Press, Cambridge, UK.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Wahba, G. (1978). Imroper priors, spline smoothing and the problems of guarding agains model errors in regression. *JRSS B.* **40** 364–372.

Walker, S. G. (2003). Bayesian consistency for a class of regression problems. *South African Stat. Journal* **37** 149–167.

Wasserman, L. (1998). Asymptotic properties of nonparametric bayesian procedures. In *Practical nonparametric and semiparametric Bayesian statistics, Lecture Notes in Statistics 133* (D. Dey, P. Müller and D. Sinha, eds.). Springer-Verlag, New York.

Zhao, L. H. (2000). Bayesian aspects of some nonparametric problems . *Ann. Statist.* **28** 532–552.