

# LOCAL AND GLOBAL ASYMPTOTIC INFERENCE IN SMOOTHING SPLINE MODELS

BY ZUOFENG SHANG\* AND GUANG CHENG †

*University of Notre Dame and Purdue University*

AUGUST 26, 2013

This article studies local and global inference for smoothing spline estimation in a unified asymptotic framework. We first introduce a new technical tool called functional Bahadur representation, which significantly generalizes the traditional Bahadur representation in parametric models, i.e., Bahadur [2]. Equipped with this tool, we develop four interconnected procedures for inference: (i) pointwise confidence interval; (ii) local likelihood ratio testing; (iii) simultaneous confidence band; (iv) global likelihood ratio testing. In particular, our confidence intervals are proved to be asymptotically valid at any point in the support, and they are shorter on average than the Bayesian confidence intervals proposed by Wahba [51] and Nychka [36]. We also discuss a version of the Wilks phenomenon arising from local/global likelihood ratio testing. It is also worth noting that our simultaneous confidence bands are the first ones applicable to general quasi-likelihood models. Furthermore, issues relating to optimality and efficiency are carefully addressed. As a by-product, we discover a surprising relationship between periodic and nonperiodic smoothing splines in terms of inference.

**1. Introduction.** As a flexible modeling tool, smoothing splines provide a general framework for statistical analysis in a variety of fields; see [52, 53, 19]. The asymptotic studies on smoothing splines in the literature focus primarily on the estimation performance, and in particular the global convergence. However, in practice it is often of great interest to conduct *asymptotic inference* on the unknown functions. The procedures for inference developed in this article, together with

---

\*Postdoctoral Fellow.

†Corresponding Author. Associate Professor. Research Sponsored by NSF, DMS-0906497, and CAREER Award DMS-1151692

*AMS 2000 subject classifications:* Primary 62G20, 62F25; secondary 62F15, 62F12

*Keywords and phrases:* asymptotic normality, functional Bahadur representation, local/global likelihood ratio test, simultaneous confidence band, smoothing spline

their rigorously derived asymptotic properties, fill this long-standing gap in the smoothing spline literature.

As an illustration, consider two popular nonparametric regression models: (i) normal regression:  $Y|Z = z \sim N(g_0(z), \sigma^2)$  for some unknown  $\sigma^2 > 0$ ; (ii) logistic regression:  $P(Y = 1|Z = z) = \exp(g_0(z))/(1 + \exp(g_0(z)))$ . The function  $g_0$  is assumed to be smooth in both models. Our goal in this paper is to develop asymptotic theory for constructing pointwise confidence intervals and simultaneous confidence bands for  $g_0$ , testing on the value of  $g_0(z_0)$  at any given point  $z_0$ , and testing whether  $g_0$  satisfies certain global properties such as linearity. Pointwise confidence intervals and tests on a local value are known as local inference. Simultaneous confidence bands and tests on a global property are known as global inference. To the best of our knowledge, there has been little systematic and rigorous theoretical study of asymptotic inference. This is partly because of the technical restrictions of the widely used equivalent kernel method. The *functional Bahadur representation (FBR)* developed in this paper makes several important contributions to this area. Our main contribution is a set of procedures for local and global inference for a univariate smooth function in a general class of nonparametric regression models that cover both the aforementioned cases. Moreover, we investigate issues relating to optimality and efficiency that have not been treated in the existing smoothing spline literature.

The equivalent kernel has long been used as a standard tool for handling the asymptotic properties of smoothing spline estimators, but this method is restricted to least square regression; see [44, 34]. Furthermore, the equivalent kernel only “approximates” the reproducing kernel function, and the approximation formula becomes extremely complicated when the penalty order increases or the design points are nonuniform. To analyze the smoothing spline estimate in a more effective way, we employ empirical process theory to develop a new technical tool, the functional Bahadur representation, which directly handles the “exact” reproducing kernel, and makes it possible to study asymptotic inference in a broader range of nonparametric models. An immediate consequence of the FBR is the asymptotic normality of the smoothing spline estimate. This naturally leads to the construction of pointwise asymptotic confidence intervals (CIs). The classical Bayesian CIs in the literature ([51, 36]) are valid on average over the observed covariates. However, our CIs are proved to be theoretically valid at any point, and they even have shorter lengths than the Bayesian CIs. We next introduce a likelihood ratio method for testing the local value of a regression function. It is shown that the null limiting distribution is a scaled Chi-square with one degree of freedom, and that the scaling constant converges to one as the smoothness level of the regression function increases. Therefore, we have discovered an interesting Wilks phenomenon (meaning that the asymptotic null distribution is free of nuisance parameters) arising from the proposed nonparametric local testing.

Procedures for global inference are also useful. Simultaneous confidence bands (SCBs) accurately depict the global behavior of the regression function, and they have been extensively studied in the literature. However, most of the efforts were devoted to simple regression models with additive Gaussian errors based on kernel or local polynomial estimates; see [20, 46, 7, 14, 55]. By incorporating the reproducing kernel Hilbert space (RKHS) theory into [4], we obtain an SCB applicable to general nonparametric regression models, and we demonstrate its theoretical validity based on strong approximation techniques. To the best of our knowledge, this is the first SCB ever developed for a general nonparametric regression model in smoothing spline settings. We further demonstrate that our SCB is optimal in the sense that the minimum width of the SCB achieves the lower bound established by [17]. Model assessment is another important aspect of global inference. Fan et al. [15] used local polynomial estimates for testing nonparametric regression models, namely the generalized likelihood ratio test (GLRT). Based on smoothing spline estimates, we propose an alternative method called the penalized likelihood ratio test (PLRT), and we identify its null limiting distribution as nearly Chi-square with diverging degrees of freedom. Therefore, the Wilks phenomenon holds for the global test as well. More importantly, we show that the PLRT achieves the minimax rate of testing in the sense of [23]. In comparison, other smoothing-spline-based tests such as the locally most powerful (LMP) test, the generalized cross validation (GCV) test, and the generalized maximum likelihood ratio (GML) test (see [10, 52, 24, 6, 39, 30]) either lead to complicated null distributions with nuisance parameters or are not known to be optimal.

As a by-product, we derive the asymptotic equivalence of the proposed procedures based on periodic and nonperiodic smoothing splines under mild conditions; see Remark 5.2. In general, our findings reveal an intrinsic connection between the two rather different basis structures, which in turn facilitates the implementation of inference.

Our paper is mainly devoted to theoretical studies. However, a few practical issues are noteworthy. GCV is currently used for the empirical tuning of the smoothing parameter, and it is known to result in biased estimates if the true function is spatially inhomogeneous with peaks and troughs. Moreover, the penalty order is prespecified rather than data-driven. Future research is needed to develop an efficient method for choosing a suitable smoothing parameter for bias reduction and an empirical method for quantifying the penalty order through data. We also note that some of our asymptotic procedures are not fully automatic since certain quantities need to be estimated; see Example 6.3. A large sample size may be necessary for the benefits of our asymptotic methods to become apparent. Finally, we want to mention that extensions to more complicated models such as multivariate smoothing spline models and semiparametric models are conceptually feasible by applying similar FBR techniques and likelihood-based approaches.

The rest of this paper is organized as follows. Section 2 introduces the basic notation, the model assumptions, and some preliminary RKHS results. Section 3 presents the FBR and the local asymptotic results. In Sections 4 and 5, several procedures for local and global inference together with their theoretical properties are formally discussed. In Section 6, we give three concrete examples to illustrate our theory. Numerical studies are also provided for both periodic and nonperiodic splines. The proofs are included in an online supplementary document [42].

## 2. Preliminaries.

2.1. *Notation and Assumptions.* Suppose that the data  $T_i = (Y_i, Z_i)$ ,  $i = 1, \dots, n$ , are *i.i.d.* copies of  $T = (Y, Z)$ , where  $Y \in \mathcal{Y} \subseteq \mathbb{R}$  is the response variable,  $Z \in \mathbb{I}$  is the covariate variable, and  $\mathbb{I} = [0, 1]$ . Consider a general class of nonparametric regression models under the primary assumption

$$(2.1) \quad \mu_0(Z) \equiv E(Y|Z) = F(g_0(Z)),$$

where  $g_0(\cdot)$  is some unknown smooth function and  $F(\cdot)$  is a known link function. This framework covers two subclasses of statistical interest. The first subclass assumes that the data are modeled by  $y_i|z_i \sim p(y_i; \mu_0(z_i))$  for a conditional distribution  $p(y; \mu_0(z))$  unknown up to  $\mu_0$ . Instead of assuming known distributions, the second subclass specifies the relation between the conditional mean and variance as  $Var(Y|Z) = \mathcal{V}(\mu_0(Z))$ , where  $\mathcal{V}$  is a known positive-valued function. The nonparametric estimation of  $g$  in the second situation uses the quasi-likelihood  $Q(y; \mu) \equiv \int_y^\mu (y - s)/\mathcal{V}(s)ds$  as an objective function (see [54]), where  $\mu = F(g)$ . Despite distinct modeling principles, the two subclasses have a large overlap since  $Q(y; \mu)$  coincides with  $\log p(y; \mu)$  under many common combinations of  $(F, \mathcal{V})$ ; see Table 2.1 of [32].

From now on, we focus on a smooth criterion function  $\ell(y; a) : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}$  that covers the above two cases, i.e.,  $\ell(y; a) = Q(y; F(a))$  or  $\log p(y; F(a))$ . Throughout this paper we define the functional parameter space  $\mathcal{H}$  as the  $m$ th-order Sobolev space:

$$H^m(\mathbb{I}) \equiv \{g : \mathbb{I} \mapsto \mathbb{R} \mid g^{(j)} \text{ is absolutely continuous for } j = 0, 1, \dots, m-1, \text{ and } g^{(m)} \in L_2(\mathbb{I})\},$$

where  $m$  is assumed to be known and larger than  $1/2$ . With some abuse of notation,  $\mathcal{H}$  may also refer to the homogeneous subspace  $H_0^m(\mathbb{I})$  of  $H^m(\mathbb{I})$ . The space  $H_0^m(\mathbb{I})$  is also known as the class of periodic functions such that a function  $g \in H_0^m(\mathbb{I})$  has the additional restrictions  $g^{(j)}(0) = g^{(j)}(1)$  for  $j = 0, 1, \dots, m-1$ . Let  $J(g, \tilde{g}) = \int_{\mathbb{I}} g^{(m)}(z)\tilde{g}^{(m)}(z)dz$ . Consider the penalized nonparametric estimate  $\hat{g}_{n,\lambda}$ :

$$(2.2) \quad \hat{g}_{n,\lambda} = \arg \max_{g \in \mathcal{H}} \ell_{n,\lambda}(g) = \arg \max_{g \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell(Y_i; g(Z_i)) - (\lambda/2)J(g, g) \right\},$$

where  $J(g, g)$  is the roughness penalty and  $\lambda$  is the smoothing parameter, which converges to zero as  $n \rightarrow \infty$ . We use  $\lambda/2$  (rather than  $\lambda$ ) to simplify future expressions. The existence and uniqueness of  $\hat{g}_{n,\lambda}$  are guaranteed by Theorem 2.9 of [19] when the null space  $\mathcal{N}_m \equiv \{g \in \mathcal{H} : J(g, g) = 0\}$  is finite dimensional and  $\ell(y; a)$  is concave and continuous w.r.t.  $a$ .

We next assume a set of model conditions. Let  $\mathcal{I}_0$  be the range of  $g_0$ , which is obviously compact. Denote the first-, second-, and third-order derivatives of  $\ell(y; a)$  w.r.t.  $a$  by  $\dot{\ell}_a(y; a)$ ,  $\ddot{\ell}_a(y; a)$ , and  $\ell_a'''(y; a)$ , respectively. We assume the following smoothness and tail conditions on  $\ell$ :

ASSUMPTION A.1. (a)  $\ell(y; a)$  is three times continuously differentiable and concave w.r.t.  $a$ .

There exists a bounded open interval  $\mathcal{I} \supset \mathcal{I}_0$  and positive constants  $C_0$  and  $C_1$  s.t.

$$(2.3) \quad E \left\{ \exp \left( \sup_{a \in \mathcal{I}} |\ddot{\ell}_a(Y; a)| / C_0 \right) \middle| Z \right\} \leq C_0, \text{ a.s.},$$

and

$$(2.4) \quad E \left\{ \exp \left( \sup_{a \in \mathcal{I}} |\ell_a'''(Y; a)| / C_0 \right) \middle| Z \right\} \leq C_1, \text{ a.s.}$$

(b) There exists a positive constant  $C_2$  such that  $C_2^{-1} \leq I(Z) \equiv -E(\ddot{\ell}_a(Y; g_0(Z)) | Z) \leq C_2$  a.s.

(c)  $\epsilon \equiv \dot{\ell}_a(Y; g_0(Z))$  satisfies  $E(\epsilon | Z) = 0$  and  $E(\epsilon^2 | Z) = I(Z)$ , a.s.

Assumption A.1(a) implies the slow diverging rate  $O_P(\log n)$  of  $\max_{1 \leq i \leq n} \sup_{a \in \mathcal{I}} |\ddot{\ell}_a(Y_i; a)|$  and  $\max_{1 \leq i \leq n} \sup_{a \in \mathcal{I}} |\ell_a'''(Y_i; a)|$ . When  $\ell(y; a) = \log p(y; a)$ , Assumption A.1(b) imposes boundedness and positive definiteness of the Fisher information, and Assumption A.1(c) trivially holds if  $p$  satisfies certain regularity conditions. When  $\ell(y; a) = Q(y; F(a))$ , we have

$$(2.5) \quad \ddot{\ell}_a(Y; a) = F_1(a) + \varepsilon F_2(a) \quad \text{and} \quad \ell_a'''(Y; a) = \dot{F}_1(a) + \varepsilon \dot{F}_2(a),$$

where  $\varepsilon = Y - \mu_0(Z)$ ,  $F_1(a) = -|\dot{F}(a)|^2 / \mathcal{V}(F(a)) + (F(g_0(Z)) - F(a)) F_2(a)$ , and  $F_2(a) = (\ddot{F}(a) \mathcal{V}(F(a)) - \dot{\mathcal{V}}(F(a)) |\dot{F}(a)|^2) / \mathcal{V}^2(F(a))$ . Hence, Assumption A.1(a) holds if  $F_j(a)$ ,  $\dot{F}_j(a)$ ,  $j = 1, 2$ , are all bounded over  $a \in \mathcal{I}$ , and

$$(2.6) \quad E \{ \exp(|\varepsilon| / C_0) | Z \} \leq C_1, \text{ a.s.}$$

By (2.5), we have  $I(Z) = |\dot{F}(g_0(Z))|^2 / \mathcal{V}(F(g_0(Z)))$ . Thus, Assumption A.1(b) holds if

$$(2.7) \quad 1/C_2 \leq \frac{|\dot{F}(a)|^2}{\mathcal{V}(F(a))} \leq C_2 \quad \text{for all } a \in \mathcal{I}_0, \text{ a.s.}$$

Assumption A.1(c) follows from the definition of  $\mathcal{V}(\cdot)$ . The sub-exponential tail condition (2.6) and the boundedness condition (2.7) are very mild quasi-likelihood model assumptions (e.g., also assumed in [33]). The assumption that  $F_j$  and  $\dot{F}_j$  are both bounded over  $\mathcal{I}$  could be restrictive and can be removed in many cases, such as the binary logistic regression model, by applying empirical process arguments similar to those in Section 7 of [33].

2.2. *Reproducing Kernel Hilbert Space.* We now introduce a number of RKHS results as extensions of [9] and [37]. It is well known that, when  $m > 1/2$ ,  $\mathcal{H} = H^m(\mathbb{I})$  (or  $H_0^m(\mathbb{I})$ ) is an RKHS endowed with the inner product  $\langle g, \tilde{g} \rangle = E\{I(Z)g(Z)\tilde{g}(Z)\} + \lambda J(g, \tilde{g})$  and the norm

$$(2.8) \quad \|g\|^2 = \langle g, g \rangle.$$

The reproducing kernel  $K(z_1, z_2)$  defined on  $\mathbb{I} \times \mathbb{I}$  is known to have the following property:

$$K_z(\cdot) \equiv K(z, \cdot) \in \mathcal{H} \quad \text{and} \quad \langle K_z, g \rangle = g(z), \quad \text{for any } z \in \mathbb{I} \text{ and } g \in \mathcal{H}.$$

Obviously,  $K$  is symmetric with  $K(z_1, z_2) = K(z_2, z_1)$ . We further introduce a positive definite self-adjoint operator  $W_\lambda : \mathcal{H} \mapsto \mathcal{H}$  such that

$$(2.9) \quad \langle W_\lambda g, \tilde{g} \rangle = \lambda J(g, \tilde{g}),$$

for any  $g, \tilde{g} \in \mathcal{H}$ . Let  $V(g, \tilde{g}) = E\{I(Z)g(Z)\tilde{g}(Z)\}$ . Then  $\langle g, \tilde{g} \rangle = V(g, \tilde{g}) + \langle W_\lambda g, \tilde{g} \rangle$  and  $V(g, \tilde{g}) = \langle (id - W_\lambda)g, \tilde{g} \rangle$ , where  $id$  denotes the identity operator.

Next, we assume that there exists a sequence of basis functions in the space  $\mathcal{H}$  that simultaneously diagonalizes the bilinear forms  $V$  and  $J$ . Such eigenvalue/eigenfunction assumptions are typical in the smoothing spline literature, and they are critical to control the local behavior of the penalized estimates. Hereafter, we denote positive sequences  $a_\mu$  and  $b_\mu$  as  $a_\mu \asymp b_\mu$  if they satisfy  $\lim_{\mu \rightarrow \infty} (a_\mu/b_\mu) = c > 0$ . If  $c = 1$ , we write  $a_\mu \sim b_\mu$ . Let  $\sum_\nu$  denote the sum over  $\nu \in \mathbb{N} = \{0, 1, 2, \dots\}$  for convenience. Denote the sup-norm of  $g \in \mathcal{H}$  as  $\|g\|_{\text{sup}} = \sup_{z \in \mathbb{I}} |g(z)|$ .

**ASSUMPTION A.2.** *There exists a sequence of eigenfunctions  $h_\nu \in \mathcal{H}$  satisfying  $\sup_{\nu \in \mathbb{N}} \|h_\nu\|_{\text{sup}} < \infty$ , and a nondecreasing sequence of eigenvalues  $\gamma_\nu \asymp \nu^{2m}$  such that*

$$(2.10) \quad V(h_\mu, h_\nu) = \delta_{\mu\nu}, \quad J(h_\mu, h_\nu) = \gamma_\mu \delta_{\mu\nu}, \quad \mu, \nu \in \mathbb{N},$$

where  $\delta_{\mu\nu}$  is the Kronecker's delta. In particular, any  $g \in \mathcal{H}$  admits a Fourier expansion  $g = \sum_\nu V(g, h_\nu) h_\nu$  with convergence in the  $\|\cdot\|$ -norm.

Assumption A.2 enables us to derive explicit expressions for  $\|g\|$ ,  $K_z(\cdot)$ , and  $W_\lambda h_\nu(\cdot)$  for any  $g \in \mathcal{H}$  and  $z \in \mathbb{I}$ ; see Proposition 2.1 below.

**PROPOSITION 2.1.** *For any  $g \in \mathcal{H}$  and  $z \in \mathbb{I}$ , we have  $\|g\|^2 = \sum_\nu |V(g, h_\nu)|^2 (1 + \lambda \gamma_\nu)$ ,  $K_z(\cdot) = \sum_\nu \frac{h_\nu(z)}{1 + \lambda \gamma_\nu} h_\nu(\cdot)$ , and  $W_\lambda h_\nu(\cdot) = \frac{\lambda \gamma_\nu}{1 + \lambda \gamma_\nu} h_\nu(\cdot)$  under Assumption A.2.*

For future theoretical derivations, we need to figure out the underlying eigensystem that implies Assumption A.2. For example, when  $\ell(y; a) = -(y - a)^2/2$  and  $\mathcal{H} = H_0^m(\mathbb{I})$ , Assumption A.2 is

known to be satisfied if  $(\gamma_\nu, h_\nu)$  is chosen as the trigonometric polynomial basis specified in (6.2) of Example 6.1. For general  $\ell(y; a)$  with  $\mathcal{H} = H^m(\mathbb{I})$ , Proposition 2.2 below says that Assumption A.2 is still valid if  $(\gamma_\nu, h_\nu)$  is chosen as the (normalized) solution of the following equations:

$$(2.11) \quad (-1)^m h_\nu^{(2m)}(\cdot) = \gamma_\nu I(\cdot) \pi(\cdot) h_\nu(\cdot), \quad h_\nu^{(j)}(0) = h_\nu^{(j)}(1) = 0, \quad j = m, m+1, \dots, 2m-1,$$

where  $\pi(\cdot)$  is the marginal density of the covariate  $Z$ . Proposition 2.2 can be viewed as a nontrivial extension of [49], which assumes  $I = \pi = 1$ . The proof relies substantially on the ODE techniques developed in [5, 45]. Let  $C^m(\mathbb{I})$  be the class of the  $m$ th-order continuously differentiable functions over  $\mathbb{I}$ .

**PROPOSITION 2.2.** *If  $\pi(z), I(z) \in C^{2m-1}(\mathbb{I})$  are both bounded away from zero and infinity over  $\mathbb{I}$ , then the eigenvalues  $\gamma_\nu$  and the corresponding eigenfunctions  $h_\nu$ , found from (2.11) and normalized to  $V(h_\nu, h_\nu) = 1$ , satisfy Assumption A.2.*

Finally, for later use we summarize the notation for Fréchet derivatives. Let  $\Delta g, \Delta g_j \in \mathcal{H}$  for  $j = 1, 2, 3$ . The Fréchet derivative of  $\ell_{n,\lambda}$  can be identified as

$$\begin{aligned} D\ell_{n,\lambda}(g)\Delta g &= \frac{1}{n} \sum_{i=1}^n \dot{\ell}_a(Y_i; g(Z_i)) \langle K_{Z_i}, \Delta g \rangle - \langle W_\lambda g, \Delta g \rangle \\ &\equiv \langle S_n(g), \Delta g \rangle - \langle W_\lambda g, \Delta g \rangle \equiv \langle S_{n,\lambda}(g), \Delta g \rangle. \end{aligned}$$

Note that  $S_{n,\lambda}(\hat{g}_{n,\lambda}) = 0$ , and  $S_{n,\lambda}(g_0)$  can be expressed as

$$(2.12) \quad S_{n,\lambda}(g_0) = \frac{1}{n} \sum_{i=1}^n \epsilon_i K_{Z_i} - W_\lambda g_0.$$

The Fréchet derivative of  $S_{n,\lambda}$  ( $DS_{n,\lambda}$ ) is denoted  $DS_{n,\lambda}(g)\Delta g_1\Delta g_2$  ( $D^2S_{n,\lambda}(g)\Delta g_1\Delta g_2\Delta g_3$ ). These derivatives can be explicitly written as  $D^2\ell_{n,\lambda}(g)\Delta g_1\Delta g_2 = n^{-1} \sum_{i=1}^n \ddot{\ell}_a(Y_i; g(Z_i)) \langle K_{Z_i}, \Delta g_1 \rangle \langle K_{Z_i}, \Delta g_2 \rangle - \langle W_\lambda \Delta g_1, \Delta g_2 \rangle$  (or  $D^3\ell_{n,\lambda}(g)\Delta g_1\Delta g_2\Delta g_3 = n^{-1} \sum_{i=1}^n \ell_a'''(Y_i; g(Z_i)) \langle K_{Z_i}, \Delta g_1 \rangle \langle K_{Z_i}, \Delta g_2 \rangle \langle K_{Z_i}, \Delta g_3 \rangle$ ).

Define  $S(g) = E\{S_n(g)\}$ ,  $S_\lambda(g) = S(g) - W_\lambda g$ , and  $DS_\lambda(g) = DS(g) - W_\lambda$ , where  $DS(g)\Delta g_1\Delta g_2 = E\{\ddot{\ell}_a(Y; g(Z)) \langle K_Z, \Delta g_1 \rangle \langle K_Z, \Delta g_2 \rangle\}$ . Since  $\langle DS_\lambda(g_0)f, g \rangle = -\langle f, g \rangle$  for any  $f, g \in \mathcal{H}$ , we have the following result:

**PROPOSITION 2.3.**  *$DS_\lambda(g_0) = -id$ , where  $id$  is the identity operator on  $\mathcal{H}$ .*

**3. Functional Bahadur Representation.** In this section, we first develop the key technical tool of this paper: *functional Bahadur representation*, and we then present the local asymptotics of the smoothing spline estimate as a straightforward application. In fact, FBR provides a rigorous theoretical foundation for the procedures for inference developed in Sections 4 and 5.

3.1. *Functional Bahadur Representation.* We first present a relationship between the norms  $\|\cdot\|_{\text{sup}}$  and  $\|\cdot\|$  in Lemma 3.1 below, and we then derive a *concentration inequality* in Lemma 3.2 as the preliminary step for obtaining the FBR. For convenience, we denote  $\lambda^{1/(2m)}$  as  $h$ .

LEMMA 3.1. *There exists a constant  $c_m > 0$  s.t.  $|g(z)| \leq c_m h^{-1/2} \|g\|$  for any  $z \in \mathbb{I}$  and  $g \in \mathcal{H}$ . In particular,  $c_m$  is not dependent on the choice of  $z$  and  $g$ . Hence,  $\|g\|_{\text{sup}} \leq c_m h^{-1/2} \|g\|$ .*

Define  $\mathcal{G} = \{g \in \mathcal{H} : \|g\|_{\text{sup}} \leq 1, J(g, g) \leq c_m^{-2} h \lambda^{-1}\}$ , where  $c_m$  is specified in Lemma 3.1. Recall that  $\mathcal{T}$  denotes the domain of the full data variable  $T = (Y, Z)$ . We now prove a concentration inequality on the empirical process  $Z_n(g)$  defined, for any  $g \in \mathcal{G}$  and  $z \in \mathbb{I}$ , as

$$(3.1) \quad Z_n(g)(z) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [\psi_n(T_i; g) K_{Z_i}(z) - E(\psi_n(T; g) K_Z(z))],$$

where  $\psi_n(T; g)$  is a real-valued function (possibly depending on  $n$ ) defined on  $\mathcal{T} \times \mathcal{G}$ .

LEMMA 3.2. *Suppose that  $\psi_n$  satisfies the following Lipschitz continuity condition:*

$$(3.2) \quad |\psi_n(T; f) - \psi_n(T; g)| \leq c_m^{-1} h^{1/2} \|f - g\|_{\text{sup}} \text{ for any } f, g \in \mathcal{G},$$

where  $c_m$  is specified in Lemma 3.1. Then we have

$$\lim_{n \rightarrow \infty} P \left( \sup_{g \in \mathcal{G}} \frac{\|Z_n(g)\|}{h^{-(2m-1)/(4m)} \|g\|_{\text{sup}}^{1-1/(2m)} + n^{-1/2}} \leq (5 \log \log n)^{1/2} \right) = 1.$$

To obtain the FBR, we need to further assume a proper convergence rate for  $\widehat{g}_{n,\lambda}$ :

$$\text{ASSUMPTION A.3.} \quad \|\widehat{g}_{n,\lambda} - g_0\| = O_P((nh)^{-1/2} + h^m).$$

Simple (but not necessarily the weakest) sufficient conditions for Assumption A.3 are provided in Proposition 3.3 below. Before stating this result, we introduce another norm on the space  $\mathcal{H}$  that is more commonly used in functional analysis. For any  $g \in \mathcal{H}$ , define

$$(3.3) \quad \|g\|_{\mathcal{H}}^2 = E\{I(Z)g(Z)^2\} + J(g, g).$$

When  $\lambda \leq 1$ ,  $\|\cdot\|_{\mathcal{H}}$  is a type of Sobolev norm dominating  $\|\cdot\|$  defined in (2.8). Denote

$$(3.4) \quad \lambda^* \asymp n^{-2m/(2m+1)} \text{ or equivalently, } h^* \asymp n^{-1/(2m+1)}.$$

Note that  $\lambda^*$  is known as the optimal order when we estimate  $g_0 \in \mathcal{H}$ .

PROPOSITION 3.3. *Suppose that Assumption A.1 holds, and further  $\|\widehat{g}_{n,\lambda} - g_0\|_{\mathcal{H}} = o_P(1)$ . If  $h$  satisfies  $(n^{1/2}h)^{-1}(\log \log n)^{m/(2m-1)}(\log n)^{2m/(2m-1)} = o(1)$ , then Assumption A.3 is satisfied. In particular,  $\widehat{g}_{n,\lambda}$  achieves the optimal rate of convergence, i.e.,  $O_P(n^{-m/(2m+1)})$ , when  $\lambda = \lambda^*$ .*

Classical results on rates of convergence are obtained through either linearization techniques, e.g., [9], or quadratic approximation devices, e.g., [18, 19]. However, the proof of Proposition 3.3 relies on empirical process techniques. Hence, it is not surprising that Proposition 3.3 requires a different set of conditions than those used in [9, 18, 19], although the derived convergence rates are the same and in all approaches the optimal rate is achieved when  $\lambda = \lambda^*$ . For example, Cox and O’Sullivan [9] assumed a weaker smoothness condition on the likelihood function but a more restrictive condition on  $h$ , i.e.,  $(n^{1/2}h\lambda^\alpha)^{-1} = o(1)$  for some  $\alpha > 0$ .

Now we are ready to present the key technical tool: *functional Bahadur representation*, which is also of independent interest. Shang [41] developed a different form of Bahadur representation, which is of limited use in practice. This is due to the intractable form of the inverse operator  $DS_\lambda(g_0)^{-1}$ , constructed based on a different type of Sobolev norm. However, by incorporating  $\lambda$  into the norm (2.8), we can show  $DS_\lambda(g_0)^{-1} = -id$  based on Proposition 2.3, and thus obtain a more refined version of the representation of [41] that naturally applies to our general setting for inference purposes.

**THEOREM 3.4.** (*Functional Bahadur Representation*) *Suppose that Assumptions A.1–A.3 hold,  $h = o(1)$ , and  $nh^2 \rightarrow \infty$ . Recall that  $S_{n,\lambda}(g_0)$  is defined in (2.12). Then we have*

$$(3.5) \quad \|\widehat{g}_{n,\lambda} - g_0 - S_{n,\lambda}(g_0)\| = O_P(a_n \log n),$$

where  $a_n = n^{-1/2}((nh)^{-1/2} + h^m)h^{-(6m-1)/(4m)}(\log \log n)^{1/2} + C_\ell h^{-1/2}((nh)^{-1} + h^{2m})/\log n$  and  $C_\ell = \sup_{z \in \mathbb{I}} E\{\sup_{a \in \mathcal{I}} |\ell_a'''(Y; a)| | Z = z\}$ . When  $h = h^*$ , the RHS of (3.5) is  $o_P(n^{-m/(2m+1)})$ .

**3.2. Local Asymptotic Behavior.** In this section, we obtain the pointwise asymptotics of  $\widehat{g}_{n,\lambda}$  as a direct application of the FBR. The equivalent kernel method may be used for this purpose, but it is restricted to  $L_2$  regression, e.g., [44]. However, the FBR-based proof applies to more general regression. Notably, our results reveal that several well-known global convergence properties continue to hold locally.

**THEOREM 3.5.** (*General Regression*) *Assume Assumptions A.1–A.3, and suppose  $h = o(1)$ ,  $nh^2 \rightarrow \infty$ , and  $a_n \log n = o(n^{-1/2})$ , where  $a_n$  is defined in Theorem 3.4, as  $n \rightarrow \infty$ . Furthermore, for any  $z_0 \in \mathbb{I}$ ,*

$$(3.6) \quad hV(K_{z_0}, K_{z_0}) \rightarrow \sigma_{z_0}^2 \quad \text{as } n \rightarrow \infty.$$

Let  $g_0^* = (id - W_\lambda)g_0$  be the biased “true parameter.” Then we have

$$(3.7) \quad \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0^*(z_0)) \xrightarrow{d} N(0, \sigma_{z_0}^2),$$

where

$$(3.8) \quad \sigma_{z_0}^2 = \lim_{h \rightarrow 0} \sum_{\nu} \frac{h|h_{\nu}(z_0)|^2}{(1 + \lambda\gamma_{\nu})^2}.$$

From Theorem 3.5, we immediately obtain the following result:

COROLLARY 3.6. *Suppose that the conditions in Theorem 3.5 hold and*

$$(3.9) \quad \lim_{n \rightarrow \infty} (nh)^{1/2}(W_{\lambda}g_0)(z_0) = -b_{z_0}.$$

Then we have

$$(3.10) \quad \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \xrightarrow{d} N(b_{z_0}, \sigma_{z_0}^2),$$

where  $\sigma_{z_0}^2$  is defined as in (3.8).

To illustrate Corollary 3.6 in detail, we consider  $L_2$  regression in which  $W_{\lambda}g_0(z_0)$  (also  $b_{z_0}$ ) has an explicit expression under the additional boundary conditions:

$$(3.11) \quad g_0^{(j)}(0) = g_0^{(j)}(1) = 0, \text{ for } j = m, \dots, 2m - 1.$$

In fact, (3.11) is also the price we pay for obtaining the boundary results, i.e.,  $z_0 = 0, 1$ . However, (3.11) could be too strong in practice. Therefore, we provide an alternative set of conditions in (3.14) below, which can be implied by the so-called ‘‘exponential envelope condition’’ introduced in [37]. In Corollary 3.7 below, we consider two different cases:  $b_{z_0} \neq 0$  and  $b_{z_0} = 0$ .

COROLLARY 3.7. ( *$L_2$  Regression*) *Let  $m > (3 + \sqrt{5})/4 \approx 1.309$  and  $\ell(y; a) = -(y - a)^2/2$ . Suppose that Assumption A.3 and (3.6) hold, and the normalized eigenfunctions  $h_{\nu}$  satisfy (2.11). Assume that  $g_0 \in H^{2m}(\mathbb{I})$  satisfies  $\sum_{\nu} |V(g_0^{(2m)}, h_{\nu})h_{\nu}(z_0)| < \infty$ .*

(i) *Suppose  $g_0$  satisfies the boundary conditions (3.11). If  $h/n^{-1/(4m+1)} \rightarrow c > 0$ , then we have, for any  $z_0 \in [0, 1]$ ,*

$$(3.12) \quad \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \xrightarrow{d} N\left(\left((-1)^{m-1}c^{2m}g_0^{(2m)}(z_0)/\pi(z_0), \sigma_{z_0}^2\right)\right).$$

*If  $h \asymp n^{-d}$  for some  $\frac{1}{4m+1} < d \leq \frac{2m}{8m-1}$ , then we have, for any  $z_0 \in [0, 1]$ ,*

$$(3.13) \quad \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \xrightarrow{d} N(0, \sigma_{z_0}^2).$$

(ii) *If we replace the boundary conditions (3.11) by the following reproducing kernel conditions: for any  $z_0 \in (0, 1)$ , as  $h \rightarrow 0$*

$$(3.14) \quad \left. \frac{\partial^j}{\partial z^j} K_{z_0}(z) \right|_{z=0} = o(1), \quad \left. \frac{\partial^j}{\partial z^j} K_{z_0}(z) \right|_{z=1} = o(1), \text{ for } j = 0, \dots, m - 1,$$

*then (3.12) and (3.13) hold for any  $z_0 \in (0, 1)$ .*

We note that in (3.12) the asymptotic bias is proportional to  $g_0^{(2m)}(z_0)$ , and the asymptotic variance can be expressed as a weighted sum of squares of the (*infinitely many*) terms  $h_\nu(z_0)$ ; see (3.8). These observations are consistent with those in the polynomial spline setting insofar as the bias is proportional to  $g_0^{(2m)}(z_0)$ , and the variance is a weighted sum of squares of (*finitely many*) terms involving the normalized B-spline basis functions evaluated at  $z_0$ ; see [56]. Furthermore, (3.13) describes how to remove the asymptotic bias through undersmoothing, although the corresponding smoothing parameter yields suboptimal estimates in terms of the convergence rate.

The existing smoothing spline literature is mostly concerned with the global convergence properties of the estimates. For example, Nychka [37] and Rice and Rosenblatt [40] derived global convergence rates in terms of the (integrated) mean squared error. Instead, Theorem 3.5 and Corollaries 3.6 and 3.7 mainly focus on local asymptotics, and they conclude that the well-known global results on the rates of convergence also hold in the *local* sense.

**4. Local Asymptotic Inference.** We consider inferring  $g(\cdot)$  *locally* by constructing the pointwise asymptotic CI in Section 4.1 and testing the local hypothesis in Section 4.2.

4.1. *Pointwise Confidence Interval.* We consider the confidence interval for some real-valued smooth function of  $g_0(z_0)$  at any fixed  $z_0 \in \mathbb{I}$ , denoted  $\rho_0 = \rho(g_0(z_0))$ , e.g.,  $\rho_0 = \exp(g_0(z_0))/(1 + \exp(g_0(z_0)))$  in logistic regression. Corollary 3.6 together with the Delta method immediately implies Proposition 4.1 on the pointwise CI where the asymptotic estimation bias is assumed to be removed by undersmoothing.

PROPOSITION 4.1. (*Pointwise Confidence Interval*) Suppose that the Assumptions in Corollary 3.6 hold and that the estimation bias asymptotically vanishes, i.e.,  $\lim_{n \rightarrow \infty} (nh)^{1/2}(W_\lambda g_0)(z_0) = 0$ . Let  $\dot{\rho}(\cdot)$  be the first derivative of  $\rho(\cdot)$ . If  $\dot{\rho}(g_0(z_0)) \neq 0$ , we have

$$P \left( \rho_0 \in \left[ \rho(\widehat{g}_{n,\lambda}(z_0)) \pm \Phi(\alpha/2) \frac{\dot{\rho}(g_0(z_0))\sigma_{z_0}}{\sqrt{nh}} \right] \right) \longrightarrow 1 - \alpha,$$

where  $\Phi(\alpha)$  is the lower  $\alpha$ th quantile of  $N(0, 1)$ .

From now on, we focus on the pointwise CI for  $g_0(z_0)$  and compare it with the classical *Bayesian Confidence Intervals* proposed by Wahba [51] and Nychka [36]. For simplicity, we consider  $\ell(y; a) = -(y - a)^2/(2\sigma^2)$ ,  $Z \sim \text{Unif}[0, 1]$ , and  $\mathcal{H} = H_0^m(\mathbb{I})$  under which Proposition 4.1 implies the following asymptotic 95% CI for  $g_0(z_0)$ :

$$(4.1) \quad \widehat{g}_{n,\lambda}(z_0) \pm 1.96\sigma\sqrt{I_2/(n\pi h^\dagger)},$$

where  $h^\dagger = h\sigma^{1/m}$  and  $I_l = \int_0^1 (1+x^{2m})^{-l} dx$  for  $l = 1, 2$ ; see Case (I) of Example 6.1 for the derivations. When  $\sigma$  is unknown, we may replace it by any consistent estimate. As far as we are aware, (4.1) is the first rigorously proven pointwise CI for smoothing spline. It is well known that the Bayesian type CI only approximately achieves the 95% nominal level on average rather than pointwise. Specifically, its average coverage probability over the observed covariates is *not* exactly 95% even asymptotically. Furthermore, the Bayesian type CI ignores the important issue of coverage uniformity across the design space, and thus it may not be reliable if only evaluated at peaks or troughs, as pointed out in [36]. However, the asymptotic CI (4.1) is proved to be valid at any point, and is even shorter than the Bayesian CIs proposed in [51, 36].

As an illustration, we perform a detailed comparison of the three CIs for the special case  $m = 2$ . We first derive the asymptotically equivalent versions of the Bayesian CIs. Wahba [51] proposed the following heuristic CI under a Bayesian framework:

$$(4.2) \quad \widehat{g}_{n,\lambda}(z_0) \pm 1.96\sigma\sqrt{a(h^\dagger)},$$

where  $a(h^\dagger) = n^{-1} \left( 1 + (1 + (\pi nh^\dagger))^{-4} + 2 \sum_{\nu=1}^{n/2-1} (1 + (2\pi\nu h^\dagger))^{-4} \right)$ . Under the assumptions  $h^\dagger = o(1)$  and  $(nh^\dagger)^{-1} = o(1)$ , Lemma 6.1 in Example 6.1 implies  $2 \sum_{\nu=1}^{n/2-1} (1 + (2\pi\nu h^\dagger))^{-4} \sim I_1/(\pi h^\dagger) = 4I_2/(3\pi h^\dagger)$ , since  $I_2/I_1 = 3/4$  when  $m = 2$ . Hence, we obtain an asymptotically equivalent version of Wahba's Bayesian CI as

$$(4.3) \quad \widehat{g}_{n,\lambda}(z_0) \pm 1.96\sigma\sqrt{(4/3) \cdot I_2/(n\pi h^\dagger)}.$$

Nychka [36] further shortened (4.2) by proposing

$$(4.4) \quad \widehat{g}_{n,\lambda}(z_0) \pm 1.96\sqrt{\text{Var}(b(z_0)) + \text{Var}(v(z_0))},$$

where  $b(z_0) = E\{\widehat{g}_{n,\lambda}(z_0)\} - g_0(z_0)$  and  $v(z_0) = \widehat{g}_{n,\lambda}(z_0) - E\{\widehat{g}_{n,\lambda}(z_0)\}$ , and showed that

$$(4.5) \quad \sigma^2 a(h^\dagger)/(\text{Var}(b(z_0)) + \text{Var}(v(z_0))) \rightarrow 32/27 \text{ as } n \rightarrow \infty \text{ and } \text{Var}(v(z_0)) = 8\text{Var}(b(z_0));$$

see his equation (2.3) and the Appendix. Hence, we have

$$(4.6) \quad \text{Var}(v(z_0)) \sim \sigma^2 \cdot (I_2/(n\pi h^\dagger)) \quad \text{and} \quad \text{Var}(b(z_0)) \sim (\sigma^2/8) \cdot (I_2/(n\pi h^\dagger)).$$

Therefore, Nychka's Bayesian CI (4.4) is asymptotically equivalent to

$$(4.7) \quad \widehat{g}_{n,\lambda}(z_0) \pm 1.96\sigma\sqrt{(9/8) \cdot I_2/(n\pi h^\dagger)}.$$

In view of (4.3) and (4.7), we find that the Bayesian CIs of Wahba and Nychka are asymptotically 15.4% and 6.1%, respectively, wider than (4.1). Meanwhile, by (4.6) we find that (4.1) turns out to

be a corrected version of Nychka's CI (4.4) by removing the random bias term  $b(z_0)$ . The inclusion of  $b(z_0)$  in (4.4) might be problematic in that (i) it makes the pointwise limit distribution non-normal and thus leads to biased coverage probability; and (ii) it introduces additional variance, which unnecessarily increases the length of the interval. By removing  $b(z_0)$ , we can achieve both pointwise consistency and a shorter length. Similar considerations apply when  $m > 2$ . Furthermore, the simulation results in Example 6.1 demonstrate the superior performance of our CI in both periodic and nonperiodic splines.

*4.2. Local Likelihood Ratio Test.* In this section, we propose a likelihood ratio method for testing the value of  $g_0(z_0)$  at any  $z_0 \in \mathbb{I}$ . First, we show that the null limiting distribution is a scaled noncentral Chi-square with one degree of freedom. Second, by removing the estimation bias, we obtain a more useful central Chi-square limit distribution. We also note that as the smoothness order  $m$  approaches infinity, the scaling constant eventually converges to one. Therefore, we have unveiled an interesting Wilks phenomenon arising from the proposed nonparametric local testing. A relevant study was conducted by Banerjee [3], who considered a likelihood ratio test for *monotone* functions, but his estimation method and null limiting distribution are fundamentally different from ours.

For some prespecified point  $(z_0, w_0)$ , we consider the following hypothesis:

$$(4.8) \quad H_0 : g(z_0) = w_0 \text{ versus } H_1 : g(z_0) \neq w_0.$$

The ‘‘constrained’’ penalized log-likelihood is defined as  $L_{n,\lambda}(g) = n^{-1} \sum_{i=1}^n \ell(Y_i; w_0 + g(Z_i)) - (\lambda/2)J(g, g)$ , where  $g \in \mathcal{H}_0 = \{g \in \mathcal{H} : g(z_0) = 0\}$ . We consider the likelihood ratio test (LRT) statistic defined as

$$(4.9) \quad LRT_{n,\lambda} = \ell_{n,\lambda}(w_0 + \hat{g}_{n,\lambda}^0) - \ell_{n,\lambda}(\hat{g}_{n,\lambda}),$$

where  $\hat{g}_{n,\lambda}^0$  is the MLE of  $g$  under the local restriction, i.e.,  $\hat{g}_{n,\lambda}^0 = \arg \max_{g \in \mathcal{H}_0} L_{n,\lambda}(g)$ .

Endowed with the norm  $\|\cdot\|$ ,  $\mathcal{H}_0$  is a closed subset in  $\mathcal{H}$ , and thus a Hilbert space. Proposition 4.2 below says that  $\mathcal{H}_0$  also inherits the reproducing kernel and penalty operator from  $\mathcal{H}$ . The proof is trivial and thus omitted.

**PROPOSITION 4.2.** (a) Recall that  $K(z_1, z_2)$  is the reproducing kernel for  $\mathcal{H}$  under  $\langle \cdot, \cdot \rangle$ . The bivariate function  $K^*(z_1, z_2) = K(z_1, z_2) - (K(z_1, z_0)K(z_0, z_2))/K(z_0, z_0)$  is a reproducing kernel for  $(\mathcal{H}_0, \langle \cdot, \cdot \rangle)$ . That is, for any  $z' \in \mathbb{I}$  and  $g \in \mathcal{H}_0$ , we have  $K_{z'}^* \equiv K^*(z', \cdot) \in \mathcal{H}_0$  and  $\langle K_{z'}^*, g \rangle = g(z')$ . (b) The operator  $W_\lambda^*$  defined by  $W_\lambda^*g = W_\lambda g - [(W_\lambda g)(z_0)/K(z_0, z_0)] \cdot K_{z_0}$  is bounded linear from  $\mathcal{H}_0$  to  $\mathcal{H}_0$  and satisfies  $\langle W_\lambda^*g, \tilde{g} \rangle = \lambda J(g, \tilde{g})$ , for any  $g, \tilde{g} \in \mathcal{H}_0$ .

On the basis of Proposition 4.2, we derive the *restricted* FBR for  $\widehat{g}_{n,\lambda}^0$ , which will be used to obtain the null limiting distribution. By straightforward calculation we can find the Fréchet derivatives of  $L_{n,\lambda}$  (under  $\mathcal{H}_0$ ). Let  $\Delta g, \Delta g_j \in \mathcal{H}_0$  for  $j = 1, 2, 3$ . The first-order Fréchet derivative of  $L_{n,\lambda}$  is

$$\begin{aligned} DL_{n,\lambda}(g)\Delta g &= n^{-1} \sum_{i=1}^n \dot{\ell}_a(Y_i; w_0 + g(Z_i)) \langle K_{Z_i}^*, \Delta g \rangle - \langle W_\lambda^* g, \Delta g \rangle \\ &\equiv \langle S_n^0(g), \Delta g \rangle - \langle W_\lambda^* g, \Delta g \rangle \equiv \langle S_{n,\lambda}^0(g), \Delta g \rangle. \end{aligned}$$

Clearly, we have  $S_{n,\lambda}^0(\widehat{g}_{n,\lambda}^0) = 0$ . Define  $S^0(g)\Delta g = E\{\langle S_n^0(g), \Delta g \rangle\}$  and  $S_\lambda^0(g)\Delta g = S^0(g)\Delta g - \langle W_\lambda^* g, \Delta g \rangle$ . The second-order derivatives are  $DS_{n,\lambda}^0(g)\Delta g_1\Delta g_2 = D^2L_{n,\lambda}(g)\Delta g_1\Delta g_2$  and  $DS_\lambda^0(g)\Delta g_1\Delta g_2 = DS^0(g)\Delta g_1\Delta g_2 - \langle W_\lambda^* \Delta g_1, g_2 \rangle$ , where

$$DS^0(g)\Delta g_1\Delta g_2 = E\{\ddot{\ell}_a(Y; w_0 + g(Z)) \langle K_Z^*, \Delta g_1 \rangle \langle K_Z^*, \Delta g_2 \rangle\}.$$

The third-order Fréchet derivative of  $L_{n,\lambda}$  is

$$D^3L_{n,\lambda}(g)\Delta g_1\Delta g_2\Delta g_3 = n^{-1} \sum_{i=1}^n \ell_a'''(Y_i; w_0 + g(Z_i)) \langle K_{Z_i}^*, \Delta g_1 \rangle \langle K_{Z_i}^*, \Delta g_2 \rangle \langle K_{Z_i}^*, \Delta g_3 \rangle.$$

Similarly to Theorem 3.4, we need an additional assumption on the convergence rate of  $\widehat{g}_{n,\lambda}^0$ :

ASSUMPTION A.4. Under  $H_0$ ,  $\|\widehat{g}_{n,\lambda}^0 - g_0^0\| = O_P((nh)^{-1/2} + h^m)$ , where  $g_0^0(\cdot) = (g_0(\cdot) - w_0) \in \mathcal{H}_0$ .

Assumption A.4 is easy to verify by assuming (2.3), (2.4), and  $\|\widehat{g}_{n,\lambda}^0 - g_0^0\|_{\mathcal{H}} = o_P(1)$ . The proof is similar to that of Proposition 3.3 by replacing  $\mathcal{H}$  with the subspace  $\mathcal{H}_0$ .

THEOREM 4.3. (*Restricted FBR*) Suppose that Assumptions A.1, A.2, A.4, and  $H_0$  are satisfied. If  $h = o(1)$  and  $nh^2 \rightarrow \infty$ , then  $\|\widehat{g}_{n,\lambda}^0 - g_0^0 - S_{n,\lambda}^0(g_0^0)\| = O_P(a_n \log n)$ .

Our main result on the local LRT is presented below. Define  $r_n = (nh)^{-1/2} + h^m$ .

THEOREM 4.4. (*Local Likelihood Ratio Test*) Suppose that Assumptions A.1 through A.4 are satisfied. Also assume  $h = o(1)$ ,  $nh^2 \rightarrow \infty$ ,  $a_n = o(\min\{r_n, n^{-1}r_n^{-1}(\log n)^{-1}, n^{-1/2}(\log n)^{-1}\})$ , and  $r_n^2 h^{-1/2} = o(a_n)$ . Furthermore, for any  $z_0 \in [0, 1]$ ,  $n^{1/2}(W_\lambda g_0)(z_0)/\sqrt{K(z_0, z_0)} \rightarrow -c_{z_0}$ ,

$$(4.10) \quad \lim_{h \rightarrow 0} hV(K_{z_0}, K_{z_0}) \rightarrow \sigma_{z_0}^2 > 0, \quad \text{and} \quad \lim_{\lambda \rightarrow 0} E\{I(Z)|K_{z_0}(Z)|^2\}/K(z_0, z_0) \equiv c_0 \in (0, 1].$$

Under  $H_0$ , we show: (i)  $\|\widehat{g}_{n,\lambda} - \widehat{g}_{n,\lambda}^0 - w_0\| = O_P(n^{-1/2})$ ; (ii)  $-2n \cdot LRT_{n,\lambda} = n\|\widehat{g}_{n,\lambda} - \widehat{g}_{n,\lambda}^0 - w_0\|^2 + o_P(1)$ ; and

$$(4.11) \quad (iii) \quad -2n \cdot LRT_{n,\lambda} \xrightarrow{d} c_0 \chi_1^2(c_{z_0}^2/c_0),$$

with noncentrality parameter  $c_{z_0}^2/c_0$ .

Note that the parametric convergence rate stated in (i) of Theorem 4.4 is reasonable since the restriction is local. By Proposition 2.1, it can be explicitly shown that

$$(4.12) \quad c_0 = \lim_{\lambda \rightarrow 0} \frac{Q_2(\lambda, z_0)}{Q_1(\lambda, z_0)}, \quad \text{where } Q_l(\lambda, z) \equiv \sum_{\nu \in \mathbb{N}} \frac{|h_\nu(z)|^2}{(1 + \lambda \gamma_\nu)^l} \quad \text{for } l = 1, 2.$$

The reproducing kernel  $K$ , if it exists, is uniquely determined by the corresponding RKHS; see [11]. Therefore,  $c_0$  defined in (4.10) depends only on the parameter space. Hence, different choices of  $(\gamma_\nu, h_\nu)$  in (4.12) will give exactly the same value of  $c_0$ , although certain choices can facilitate the calculations. For example, when  $\mathcal{H} = H_0^m(\mathbb{I})$ , we can explicitly calculate the value of  $c_0$  as 0.75 (0.83) when  $m = 2$  (3) by choosing the trigonometric polynomial basis (6.2). Interestingly, when  $\mathcal{H} = H^2(\mathbb{I})$ , we can obtain the same value of  $c_0$  even without specifying its (rather different) eigensystem; see Remark 5.2 for more details. In contrast, the value of  $c_{z_0}$  in (4.11) depends on the asymptotic bias specified in (3.9), whose estimation is notoriously difficult. Fortunately, under various undersmoothing conditions, we can show  $c_{z_0} = 0$  and thus establish a central Chi-square limit distribution. For example, we can assume higher order smoothness on the true function, as in Corollary 4.5 below.

**COROLLARY 4.5.** *Suppose that Assumptions A.1 through A.4 are satisfied and  $H_0$  holds. Let  $m > 1 + \sqrt{3}/2 \approx 1.866$ . Also assume that the Fourier coefficients  $\{V(g_0, h_\nu)\}_{\nu \in \mathbb{N}}$  of  $g_0$  satisfy  $\sum_\nu |V(g_0, h_\nu)|^2 \gamma_\nu^d$  for some  $d > 1 + 1/(2m)$ , which holds if  $g_0 \in H^{md}(\mathbb{I})$ . Furthermore, if (4.10) is satisfied for any  $z_0 \in [0, 1]$ , then (4.11) holds with the limiting distribution  $c_0 \chi_1^2$  under  $\lambda = \lambda^*$ .*

Corollary 4.5 demonstrates a nonparametric type of the Wilks phenomenon, which approaches the parametric type as  $m \rightarrow \infty$  since  $\lim_{m \rightarrow \infty} c_0 = 1$ . This result provides a theoretical insight for nonparametric local hypothesis testing; see its *global* counterpart in Section 5.2.

**5. Global Asymptotic Inference.** Depicting the global behavior of a smooth function is crucial in practice. In Sections 5.1 and 5.2, we develop the *global* counterparts of Section 4 by constructing simultaneous confidence bands and testing a set of global hypotheses.

**5.1. Simultaneous Confidence Band.** In this section, we construct the SCBs for  $g$  using the approach of [4]. We demonstrate the theoretical validity of the proposed SCB based on the FBR and strong approximation techniques. The approach of [4] was originally developed in the context of density estimation, and it was then extended to M-estimation by [20] and local polynomial estimation by [14, 7, 55]. The volume-of-tube method ([46]) is another approach, but it requires the error distribution to be symmetric; see [56, 28]. Sun et al. [47] relaxed the restrictive error

assumption in generalized linear models, but they had to translate the nonparametric estimation into parametric estimation. Our SCBs work for a general class of nonparametric models including normal regression and logistic regression. Additionally, the minimum width of the proposed SCB is shown to achieve the lower bound established by [17]; see Remark 5.3. An interesting by-product is that, under the equivalent kernel conditions given in this section, the local asymptotic inference procedures developed from cubic splines and periodic splines are essentially the same despite the intrinsic difference in their eigensystems; see Remark 5.2 for technical details.

The key conditions assumed in this section are the equivalent kernel conditions (5.1)–(5.3). Specifically, we assume that there exists a real-valued function  $\omega(\cdot)$  defined on  $\mathbb{R}$  satisfying, for any fixed  $0 < \varphi < 1$ ,  $h^\varphi \leq z \leq 1 - h^\varphi$ , and  $t \in \mathbb{I}$ ,

$$(5.1) \quad \left| \frac{d^j}{dt^j} (h^{-1}\omega((z-t)/h) - K(z,t)) \right| \leq C_K h^{-(j+1)} \exp(-C_2 h^{-1+\varphi}) \quad \text{for } j = 0, 1,$$

where  $C_2, C_K$  are positive constants. Condition (5.1) implies that  $\omega$  is an equivalent kernel of the reproducing kernel function  $K$  with a certain degree of approximation accuracy. We also require two regularity conditions on  $\omega$ :

$$(5.2) \quad |\omega(u)| \leq C_\omega \exp(-|u|/C_3), \quad |\omega'(u)| \leq C_\omega \exp(-|u|/C_3), \quad \text{for any } u \in \mathbb{R},$$

and there exists a constant  $0 < \rho \leq 2$  s.t.

$$(5.3) \quad \int_{-\infty}^{\infty} \omega(t)\omega(t+z)dt = \sigma_\omega^2 - C_\rho |z|^\rho + o(|z|^\rho), \quad \text{as } |z| \rightarrow \infty,$$

where  $C_3, C_\omega, C_\rho$  are positive constants and  $\sigma_\omega^2 = \int_{\mathbb{R}} \omega(t)^2 dt$ . An example of  $\omega$  satisfying (5.1)–(5.3) will be given in Proposition 5.2. The following exponential envelope condition is also needed:

$$(5.4) \quad \sup_{z,t \in \mathbb{I}} \left| \frac{\partial}{\partial z} K(z,t) \right| = O(h^{-2}).$$

**THEOREM 5.1.** (*Simultaneous Confidence Band*) *Suppose Assumptions A.1 through A.3 are satisfied, and  $Z$  is uniform on  $\mathbb{I}$ . Let  $m > (3+\sqrt{5})/4 \approx 1.3091$  and  $h = n^{-\delta}$  for any  $\delta \in (0, 2m/(8m-1))$ . Furthermore,  $E\{\exp(|\epsilon|/C_0)|Z\} \leq C_1$ , a.s., and (5.1)–(5.4) hold. The conditional density of  $\epsilon$  given  $Z = z$ , denoted  $\pi(\epsilon|z)$ , satisfies the following: for some positive constants  $\rho_1$  and  $\rho_2$ ,*

$$(5.5) \quad \left| \frac{d}{dz} \log \pi(\epsilon|z) \right| \leq \rho_1 (1 + |\epsilon|^{\rho_2}), \quad \text{for any } \epsilon \in \mathbb{R} \text{ and } z \in \mathbb{I}.$$

Then, for any  $0 < \varphi < 1$  and  $u \in \mathbb{R}$ ,

$$(5.6) \quad P \left( (2\delta \log n)^{1/2} \left\{ \sup_{h^\varphi \leq z \leq 1-h^\varphi} (nh)^{1/2} \sigma_\omega^{-1} I(z)^{-1/2} |\hat{g}_{n,\lambda}(z) - g_0(z) + (W_\lambda g_0)(z)| - d_n \right\} \leq u \right) \rightarrow \exp(-2 \exp(-u)),$$

where  $d_n$  relies only on  $h$ ,  $\rho$ ,  $\varphi$ , and  $C_\rho$ .

The FBR developed in Section 3.1 and the strong approximation techniques developed by [4] are crucial to the proof of Theorem 5.1. The uniform distribution on  $Z$  is assumed only for simplicity, and this condition can be relaxed by requiring that the density is bounded away from zero and infinity. Condition (5.5) holds in various situations such as the conditional normal model  $\epsilon|Z = z \sim N(0, \sigma^2(z))$ , where  $\sigma^2(z)$  satisfies  $\inf_z \sigma^2(z) > 0$ , and  $\sigma(z)$  and  $\sigma'(z)$  both have finite upper bounds. The existence of the bias term  $W_\lambda g_0(z)$  in (5.6) may result in poor finite-sample performance. We address this issue by assuming undersmoothing, which is advocated by [35, 21, 22]; they showed that undersmoothing is more efficient than explicit bias correction when the goal is to minimize the coverage error. Specifically, the bias term will asymptotically vanish if we assume that

$$(5.7) \quad \lim_{n \rightarrow \infty} \left\{ \sup_{h^\varphi \leq z \leq 1-h^\varphi} \sqrt{nh \log n} |W_\lambda g_0(z)| \right\} = 0.$$

Condition (5.7) is slightly stronger than the undersmoothing condition  $\sqrt{nh}(W_\lambda g_0)(z_0) = o(1)$  assumed in Proposition 4.1. By the generalized Fourier expansion of  $W_\lambda g_0$  and the uniform boundedness of  $h_\nu$  (see Assumption A.2), we can show that (5.7) holds if we properly increase the amount of smoothness on  $g_0$  or choose a suboptimal  $\lambda$ , as in Corollary 3.7 and Corollary 4.5.

Proposition 5.2 below demonstrates the validity of Conditions (5.1)–(5.3) in  $L_2$  regression. The proof relies on an explicit construction of the equivalent kernel function obtained by [34]. We consider only  $m = 2$  for simplicity.

**PROPOSITION 5.2.** ( *$L_2$  regression*) *Let  $\ell(y; a) = -(y - a)^2/(2\sigma^2)$ ,  $Z \sim \text{Unif}[0, 1]$ , and  $\mathcal{H} = H^2(\mathbb{I})$ , i.e.,  $m = 2$ . Then, (5.1)–(5.3) hold with  $\omega(t) = \sigma^{2-1/m}\omega_0(\sigma^{-1/m}t)$  for  $t \in \mathbb{R}$ , where  $\omega_0(t) = \frac{1}{2\sqrt{2}} \exp(-|t|/\sqrt{2}) (\cos(t/\sqrt{2}) + \sin(|t|/\sqrt{2}))$ . In particular, (5.3) holds for arbitrary  $\rho \in (0, 2]$  and  $C_\rho = 0$ .*

**REMARK 5.1.** *In the setting of Proposition 5.2, we are able to explicitly find the constants  $\sigma_\omega^2$  and  $d_n$  in Theorem 5.1. Specifically, by direct calculation it can be found that  $\sigma_\omega^2 = 0.265165\sigma^{7/2}$  since  $\sigma_{\omega_0}^2 = \int_{-\infty}^{\infty} |\omega_0(t)|^2 dt = 0.265165$  when  $m = 2$ . Choose  $C_\rho = 0$  for arbitrary  $\rho \in (0, 2]$ . By the formula of  $B(t)$  given in Theorem A1 of [4], we know that*

$$(5.8) \quad d_n = (2 \log(h^{-1} - 2h^{\varphi-1}))^{1/2} + \frac{(1/\rho - 1/2) \log \log(h^{-1} - 2h^{\varphi-1})}{(2 \log(h^{-1} - 2h^{\varphi-1}))^{1/2}}.$$

*When  $\rho = 2$ , the above  $d_n$  is simplified as  $(2 \log(h^{-1} - 2h^{\varphi-1}))^{1/2}$ . In general, we observe that  $d_n \sim (-2 \log h)^{1/2} \asymp \sqrt{\log n}$  for sufficiently large  $n$  since  $h = n^{-\delta}$ . Given that the estimation bias is removed, e.g., under (5.7), we obtain the following  $100 \times (1 - \alpha)\%$  SCB:*

$$(5.9) \quad \left\{ \left[ \widehat{g}_{n,\lambda}(z) \pm 0.5149418(nh)^{-1/2} \widehat{\sigma}^{3/4} \left( c_\alpha^* / \sqrt{-2 \log h} + d_n \right) \right] : h^\varphi \leq z \leq 1 - h^\varphi \right\},$$

where  $d_n = (-2 \log h)^{1/2}$ ,  $c_\alpha^* = -\log(-\log(1-\alpha)/2)$ , and  $\hat{\sigma}$  is a consistent estimate of  $\sigma$ . Therefore, to obtain uniform coverage, we have to increase the bandwidth up to an order of  $\sqrt{\log n}$  over the length of the pointwise CI given in (4.1). Note that we have excluded the boundary points in (5.9).

REMARK 5.2. An interesting by-product we discover in the setting of Proposition 5.2 is that the pointwise asymptotic CIs for  $g_0(z_0)$  based on cubic splines and periodic splines share the same length at any  $z_0 \in (0, 1)$ . This result is surprising since the two splines have intrinsically different structures. Under (5.1), it can be shown that

$$\begin{aligned} \sigma_{z_0}^2 &\sim \sigma^{-2} h \int_0^1 |K(z_0, z)|^2 dz \\ &\sim \sigma^{-2} h^{-1} \int_0^1 \left| \omega \left( \frac{z - z_0}{h} \right) \right|^2 dz \\ &= \sigma^{-2} \int_{-z_0/h}^{(1-z_0)/h} |\omega(s)|^2 ds \sim \sigma^{-2} \int_{\mathbb{R}} |\omega(s)|^2 ds = \sigma^{3/2} \sigma_{\omega_0}^2, \end{aligned}$$

given the choice of  $\omega$  in Proposition 5.2. Thus, Corollary 3.6 implies the following 95% CI

$$(5.10) \quad \hat{g}_{n,\lambda}(z_0) \pm 1.96(nh)^{-1/2} \sigma^{3/4} \sigma_{\omega_0} = \hat{g}_{n,\lambda}(z_0) \pm 1.96(nh^\dagger)^{-1/2} \sigma \sigma_{\omega_0}.$$

Since  $\sigma_{\omega_0}^2 = I_2/\pi$ , the lengths of the CIs (4.1) (periodic spline) and (5.10) (cubic spline) coincide with each other. The above calculation of  $\sigma_{z_0}^2$  relies on  $L_2$  regression. For general models such as logistic regression, one can instead use a weighted version of (2.2) with the weights  $B(Z_i)^{-1}$  to obtain the exact formula. Another application of Proposition 5.2 is to find the value of  $c_0$  in Theorem 4.3 for the construction of the local LRT test when  $\mathcal{H} = H^2(\mathbb{I})$ . According to the definition of  $c_0$ , i.e., (4.10), we have  $c_0 \sim \sigma_{z_0}^2 / (hK(z_0, z_0))$ . Under (5.1), we have  $K(z_0, z_0) \sim h^{-1} \omega(0) = h^{-1} \sigma^{3/2} \omega_0(0) = 0.3535534h^{-1} \sigma^{3/2}$ . Since  $\sigma_{z_0}^2 \sim \sigma^{3/2} \sigma_{\omega_0}^2$  and  $\sigma_{\omega_0}^2 = I_2/\pi$ , we have  $c_0 = 0.75$ . This value coincides with that found in periodic splines, i.e.,  $\mathcal{H} = H_0^2(\mathbb{I})$ . These intriguing phenomena have never been observed in the literature and may be useful for simplifying the construction of CIs and local LRT.

REMARK 5.3. Genovese and Wasserman [17] showed that when  $g_0$  belongs to an  $m$ th-order Sobolev ball, the lower bound for the average width of an SCB is proportional to  $b_n n^{-m/(2m+1)}$ , where  $b_n$  depends only on  $\log n$ . We next show that the (minimum) bandwidth of the proposed SCB can achieve this lower bound with  $b_n = (\log n)^{(m+1)/(2m+1)}$ . Based on Theorem 5.1, the width of the SCB is of order  $d_n(nh)^{-1/2}$ , where  $d_n \asymp \sqrt{\log n}$ ; see Remark 5.1. Meanwhile, Condition (5.7) is crucial for our band to maintain the desired coverage probability. Suppose that the Fourier coefficients of  $g_0$  satisfy  $\sum_\nu |V(g_0, h_\nu)| \gamma_\nu^{1/2} < \infty$ . It can be verified that (5.7) holds when  $nh^{2m+1} \log n = O(1)$ , which

sets an upper bound for  $h$ , i.e.,  $O(n \log n)^{-1/(2m+1)}$ . When  $h$  is chosen as the above upper bound and  $d_n \asymp \sqrt{\log n}$ , our SCB achieves the minimum order of bandwidth  $n^{-m/(2m+1)}(\log n)^{(m+1)/(2m+1)}$ , which turns out to be optimal according to [17].

In practice, the construction of our SCB requires a delicate choice of  $(h, \varphi)$ . Otherwise, over-coverage or undercoverage of the true function may occur near the boundary points. There is no practical or theoretical guideline on how to find the optimal  $(h, \varphi)$ , although, as noted by [4], one can choose a proper  $h$  to make the band as thin as possible. Hence, in the next section, we propose a more straightforward likelihood-ratio-based approach for testing the global behavior, which requires only tuning  $h$ .

**5.2. Global Likelihood Ratio Test.** There is a vast literature dealing with nonparametric hypothesis testing, among which the GLRT proposed by Fan et al. [15] stands out. Because of the technical complexity, they focused on the local polynomial fitting; see [16] for a sieve version. Based on smoothing spline estimation, we propose the PLRT, which is applicable to both simple and composite hypotheses. The null limiting distribution is identified to be nearly Chi-square with diverging degrees of freedom. The degrees of freedom depend only on the functional parameter space, while the null limiting distribution of the GLRT depends on the choice of kernel functions; see Table 2 in [15]. Furthermore, the PLRT is shown to achieve the minimax rate of testing in the sense of [23]. As demonstrated in our simulations, the PLRT performs better than the GLRT in terms of power, especially in small-sample situations. Other smoothing-spline-based testing such as LMP, GCV, and GML (see [10, 52, 24, 6, 39, 30]) use ad-hoc discrepancy measures leading to complicated null distributions involving nuisance parameters; see a thorough review in [30].

Consider the following “global” hypothesis:

$$(5.11) \quad H_0^{global} : g = g_0 \text{ versus } H_1^{global} : g \in \mathcal{H} - \{g_0\},$$

where  $g_0 \in \mathcal{H}$  can be either known or unknown. The PLRT statistic is defined to be

$$(5.12) \quad PLRT_{n,\lambda} = \ell_{n,\lambda}(g_0) - \ell_{n,\lambda}(\hat{g}_{n,\lambda}).$$

Theorem 5.3 below derives the null limiting distribution of  $PLRT_{n,\lambda}$ . We remark that the null limiting distribution remains the same even when the hypothesized value  $g_0$  is unknown (whether its dimension is finite or infinite). This nice property can be used to test the composite hypothesis; see Remark 5.4.

**THEOREM 5.3.** (*Penalized Likelihood Ratio Test*) *Let Assumptions A.1 through A.3 be satisfied. Also assume  $nh^{2m+1} = O(1)$ ,  $nh^2 \rightarrow \infty$ ,  $a_n = o(\min\{r_n, n^{-1}r_n^{-1}h^{-1/2}(\log n)^{-1}, n^{-1/2}(\log n)^{-1}\})$ ,*

and  $r_n^2 h^{-1/2} = o(a_n)$ . Furthermore, under  $H_0^{\text{global}}$ ,  $E\{\epsilon^4|Z\} \leq C$ , a.s., for some constant  $C > 0$ , where  $\epsilon = \dot{\ell}_a(Y; g_0(Z))$  represents the ‘‘model error.’’ Under  $H_0^{\text{global}}$ , we have

$$(5.13) \quad (2u_n)^{-1/2} (-2nr_K \cdot PLRT_{n,\lambda} - nr_K \|W_\lambda g_0\|^2 - u_n) \xrightarrow{d} N(0, 1),$$

where  $u_n = h^{-1} \sigma_K^4 / \rho_K^2$ ,  $r_K = \sigma_K^2 / \rho_K^2$ ,

$$(5.14) \quad \sigma_K^2 = hE\{\epsilon^2 K(Z, Z)\} = \sum_\nu \frac{h}{(1 + \lambda\gamma_\nu)}, \rho_K^2 = hE\{\epsilon_1^2 \epsilon_2^2 K(Z_1, Z_2)^2\} = \sum_\nu \frac{h}{(1 + \lambda\gamma_\nu)^2},$$

and  $(\epsilon_i, Z_i)$ ,  $i = 1, 2$  are i.i.d. copies of  $(\epsilon, Z)$ .

A direct examination reveals that  $h \asymp n^{-d}$  with  $\frac{1}{2m+1} \leq d < \frac{2m}{8m-1}$  satisfies the rate conditions required by Theorem 5.3 when  $m > (3 + \sqrt{5})/4 \approx 1.309$ . By the proof of Theorem 5.3, it can be shown that  $n \|W_\lambda g_0\|^2 = o(h^{-1}) = o(u_n)$ . Therefore,  $-2nr_K \cdot PLRT_{n,\lambda}$  is asymptotically  $N(u_n, 2u_n)$ . As  $n$  approaches infinity,  $N(u_n, 2u_n)$  is nearly  $\chi_{u_n}^2$ . Hence,  $-2nr_K \cdot PLRT_{n,\lambda}$  is approximately distributed as  $\chi_{u_n}^2$ , denoted

$$(5.15) \quad -2nr_K \cdot PLRT_{n,\lambda} \overset{a}{\sim} \chi_{u_n}^2.$$

That is, the Wilks phenomenon holds for the PLRT. The specifications of (5.15), i.e.,  $\sigma_K^2$  and  $\rho_K^2$ , are determined only by the parameter space and model setup. We also note that undersmoothing is not required for our global test.

We next discuss the calculation of  $(r_K, u_n)$ . In the setting of Proposition 5.2, it can be shown by the equivalent kernel conditions that  $\sigma_K^2 = h\sigma^{-2} \int_0^1 K(z, z) dz \sim h\sigma^{-2}(h^{-1}\omega(0)) = \sigma^{-1/2}\omega_0(0) = 0.3535534\sigma^{-1/2}$ , and  $\rho_K^2 \sim \sigma^{-1/2}\sigma_{\omega_0}^2 = 0.265165\sigma^{-1/2}$ . So  $r_K = 1.3333$  and  $u_n = 0.4714h^{-1}\sigma^{-1/2}$ . If we replace  $H^2(\mathbb{I})$  by  $H_0^2(\mathbb{I})$ , direct calculation in Case (I) of Example 6.1 reveals that  $(r_K, u_n)$  have exactly the same values. When  $\mathcal{H} = H_0^m(\mathbb{I})$ , we have  $2r_K \rightarrow 2$  as  $m$  tends to infinity. This limit is consistent with the scaling constant two in the parametric likelihood ratio theory. In  $L_2$  regression, the possibly unknown parameter  $\sigma$  in  $u_n$  can be profiled out without changing the null limiting distribution. In practice, by the wild bootstrap we can directly simulate the null limiting distribution by fixing the nuisance parameters at some reasonable values or estimates without finding the values of  $(r_K, u_n)$ . This is a major advantage of the Wilks type of results.

REMARK 5.4. *We discuss composite hypothesis testing via the PLRT. Specifically, we test whether  $g$  belongs to some finite-dimensional class of functions, which is much larger than the null space  $\mathcal{N}_m$  considered in the literature. For instance, for any integer  $q \geq 0$ , consider the null hypothesis*

$$(5.16) \quad H_0^{\text{global}} : g \in \mathcal{L}_q(\mathbb{I}),$$

where  $\mathcal{L}_q(\mathbb{I}) \equiv \{g(z) = \sum_{l=0}^q a_l z^l : a = (a_0, a_1, \dots, a_q)^T \in \mathbb{R}^{q+1}\}$  is the class of the  $q$ th-order polynomials. Let  $\hat{a}_* = \arg \max_{a \in \mathbb{R}^{q+1}} \{(1/n) \sum_{i=1}^n \ell(Y_i; \sum_{l=0}^q a_l Z_i^l) - (\lambda/2) a^T D a\}$ , where

$$D = \int_0^1 (0, 0, 2, 6z, \dots, q(q-1)z^{q-2})^T (0, 0, 2, 6z, \dots, q(q-1)z^{q-2}) dz$$

is a  $(q+1) \times (q+1)$  matrix. Hence, under  $H_0^{global}$ , the penalized MLE is  $\hat{g}_*(z) = \sum_{l=0}^q \hat{a}_{*l} z^l$ . Let  $g_{0q}$  be an unknown “true parameter” in  $\mathcal{L}_q(\mathbb{I})$  corresponding to a vector of polynomial coefficients  $a^0 = (a_0^0, a_1^0, \dots, a_q^0)^T$ . To test (5.16), we decompose the PLRT statistic as  $PLRT_{n,\lambda}^{com} = L_{n1} - L_{n2}$ , where  $L_{n1} = \ell_{n,\lambda}(g_{0q}) - \ell_{n,\lambda}(\hat{g}_{n,\lambda})$  and  $L_{n2} = \ell_{n,\lambda}(g_{0q}) - \ell_{n,\lambda}(\hat{g}_*)$ . When we formulate

$$H'_0 : a = a^0 \text{ versus } H'_1 : a \neq a^0,$$

$L_{n2}$  appears to be the PLRT statistic in the parametric setup. It can be shown that  $L_{n2} = O_P(n^{-1})$  whether  $q < m$  (by applying the parametric theory in [43]) or  $q \geq m$  (by slightly modifying the proof of Theorem 4.4). On the other hand,  $L_{n1}$  is exactly the PLRT for testing

$$H_0' : g = g_{0q} \text{ versus } H_1^{global} : g \neq g_{0q}.$$

By Theorem 5.3,  $L_{n1}$  follows the limit distribution specified in (5.15). In summary, under (5.16),  $PLRT_{n,\lambda}^{com}$  has the same limit distribution since  $L_{n2} = O_P(n^{-1})$  is negligible.

To conclude this section, we show that the PLRT achieves the optimal minimax rate of testing specified in Ingster [23] based on a uniform version of the FBR. For convenience, we consider only  $\ell(Y; a) = -(Y - a)^2/2$ . Extensions to a more general setup can be found in the supplementary document [42] under stronger assumptions, e.g., a more restrictive alternative set.

Write the local alternative as  $H_{1n} : g = g_{n0}$ , where  $g_{n0} = g_0 + g_n$ ,  $g_0 \in \mathcal{H}$ , and  $g_n$  belongs to the alternative value set  $\mathcal{G}_a \equiv \{g \in \mathcal{H} | \text{Var}(g(Z)^2) \leq \zeta E^2\{g(Z)^2\}, J(g, g) \leq \zeta\}$  for some constant  $\zeta > 0$ .

**THEOREM 5.4.** *Let  $m > (3 + \sqrt{5})/4 \approx 1.309$ , and  $h \asymp n^{-d}$  for  $\frac{1}{2m+1} \leq d < \frac{2m}{8m-1}$ . Suppose that Assumption A.2 is satisfied, and uniformly over  $g_n \in \mathcal{G}_a$ ,  $\|\hat{g}_{n,\lambda} - g_{n0}\| = O_P(r_n)$  holds under  $H_{1n} : g = g_{n0}$ . Then for any  $\delta \in (0, 1)$ , there exist positive constants  $C$  and  $N$  such that*

$$(5.17) \quad \inf_{n \geq N} \inf_{\substack{g_n \in \mathcal{G}_a \\ \|g_n\| \geq C\eta_n}} P(\text{reject } H_0^{global} | H_{1n} \text{ is true}) \geq 1 - \delta,$$

where  $\eta_n \geq \sqrt{h^{2m} + (nh^{1/2})^{-1}}$ . The minimal lower bound of  $\eta_n$ , i.e.,  $n^{-2m/(4m+1)}$ , is achieved when  $h = h^{**} \equiv n^{-2/(4m+1)}$ .

The condition “uniformly over  $g_n \in \mathcal{G}_a$ ,  $\|\hat{g}_{n,\lambda} - g_{n0}\| = O_P(r_n)$  holds under  $H_{1n} : g = g_{n0}$ ” means that for any  $\tilde{\delta} > 0$ , there exist constants  $\tilde{C}$  and  $\tilde{N}$ , both unrelated to  $g_n \in \mathcal{G}_a$ , such that  $\inf_{n \geq \tilde{N}} \inf_{g_n \in \mathcal{G}_a} P_{g_{n0}}(\|\hat{g}_{n,\lambda} - g_{n0}\| \leq \tilde{C}r_n) \geq 1 - \tilde{\delta}$ .

Theorem 5.4 proves that, when  $h = h^{**}$ , the PLRT can detect any local alternatives with separation rates no faster than  $n^{-2m/(4m+1)}$ , which turns out to be the minimax rate of testing in the sense of Ingster [23]; see Remark 5.5 below.

REMARK 5.5. *The minimax rate of testing established in Ingster [23] is under the usual  $\|\cdot\|_{L_2}$ -norm (w.r.t. the Lebesgue measure). However, the separation rate derived under the  $\|\cdot\|$ -norm is still optimal because of the trivial domination of  $\|\cdot\|$  over  $\|\cdot\|_{L_2}$  (under the conditions of Theorem 5.4). Next we heuristically explain why the minimax rates of testing associated with  $\|\cdot\|$ , denoted  $b'_n$ , and with  $\|\cdot\|_{L_2}$ , denoted  $b_n$ , are the same. By definition, whenever  $\|g_n\| \geq b'_n$  or  $\|g_n\|_{L_2} \geq b_n$ ,  $H_0^{global}$  can be rejected with a large probability, or equivalently, the local alternatives can be detected.  $b'_n$  and  $b_n$  are the minimum rates that satisfy this property. Ingster [23] has shown that  $b_n \asymp n^{-2m/(4m+1)}$ . Since  $\|g_n\|_{L_2} \geq b'_n$  implies  $\|g_n\| \geq b'_n$ ,  $H_0^{global}$  is rejected. This means  $b'_n$  is an upper bound for detecting the local alternatives in terms of  $\|\cdot\|_{L_2}$ , and so  $b_n \leq b'_n$ . On the other hand, suppose  $h = h^{**} \asymp n^{-2/(4m+1)}$  and  $\|g_n\| \geq Cn^{-2m/(4m+1)} \asymp b_n$  for some large  $C > \zeta^{1/2}$ . Since  $\lambda J(g_n, g_n) \leq \zeta \lambda \asymp \zeta n^{-4m/(4m+1)}$ , it follows that  $\|g_n\|_{L_2} \geq (C^2 - \zeta)^{1/2} n^{-2m/(4m+1)} \asymp b_n$ . This means  $b_n$  is a upper bound for detecting the local alternatives in terms of  $\|\cdot\|$ , and so  $b'_n \leq b_n$ . Therefore,  $b'_n$  and  $b_n$  are of the same order.*

**6. Examples.** In this section, we provide three concrete examples together with simulations.

EXAMPLE 6.1. (*L<sub>2</sub> Regression*) We consider the regression model with an additive error

$$(6.1) \quad Y = g_0(Z) + \epsilon,$$

where  $\epsilon \sim N(0, \sigma^2)$  with an unknown variance  $\sigma^2$ . Hence,  $I(Z) = \sigma^{-2}$  and  $V(g, \tilde{g}) = \sigma^{-2} E\{g(Z)\tilde{g}(Z)\}$ . For simplicity,  $Z$  was generated uniformly over  $\mathbb{I}$ . The function `ssr()` in the R package `assist` was used to select the smoothing parameter  $\lambda$  based on CV or GCV; see [29]. We first consider  $\mathcal{H} = H_0^m(\mathbb{I})$  in Case (I) and then  $\mathcal{H} = H^m(\mathbb{I})$  in Case (II).

**Case (I).**  $\mathcal{H} = H_0^m(\mathbb{I})$ : In this case, we choose the basis functions as

$$(6.2) \quad h_\mu(z) = \begin{cases} \sigma, & \mu = 0, \\ \sqrt{2}\sigma \cos(2\pi kz), & \mu = 2k, k = 1, 2, \dots, \\ \sqrt{2}\sigma \sin(2\pi kz), & \mu = 2k - 1, k = 1, 2, \dots, \end{cases}$$

with the corresponding eigenvalues  $\gamma_{2k-1} = \gamma_{2k} = \sigma^2(2\pi k)^{2m}$  for  $k \geq 1$  and  $\gamma_0 = 0$ . Assumption A.2 trivially holds for this choice of  $(h_\mu, \gamma_\mu)$ . The Lemma below is useful for identifying the critical quantities for inference.

LEMMA 6.1. Let  $I_l = \int_0^\infty (1+x^{2m})^{-l} dx$  for  $l = 1, 2$  and  $h^\dagger = h\sigma^{1/m}$ . Then

$$(6.3) \quad \sum_{k=1}^{\infty} \frac{1}{(1+(2\pi h^\dagger k)^{2m})^l} \sim \frac{I_l}{2\pi h^\dagger}.$$

By Proposition 4.1, the asymptotic 95% pointwise CI for  $g(z_0)$  is  $\widehat{g}_{n,\lambda}(z_0) \pm 1.96\sigma_{z_0}/\sqrt{nh}$  when ignoring the bias. By the definition of  $\sigma_{z_0}^2$  and Lemma 6.1, we have

$$\sigma_{z_0}^2 \sim hV(K_{z_0}, K_{z_0}) = \sigma^2 h \left( 1 + 2 \sum_{k=1}^{\infty} (1 + (2\pi h^\dagger k)^{2m})^{-2} \right) \sim (I_2 \sigma^{2-1/m})/\pi.$$

Hence, the CI becomes

$$(6.4) \quad \widehat{g}_{n,\lambda}(z_0) \pm 1.96\widehat{\sigma}^{1-1/(2m)} \sqrt{I_2/(\pi nh)},$$

where  $\widehat{\sigma}^2 = \sum_i (Y_i - \widehat{g}_{n,\lambda}(Z_i))^2 / (n - \text{trace}(A(\lambda)))$  is a consistent estimate of  $\sigma^2$  and  $A(\lambda)$  denotes the smoothing matrix; see [52]. By (4.12) and (6.2), for  $l = 1, 2$ ,

$$\begin{aligned} Q_l(\lambda, z_0) &= \sigma^2 + \sum_{k \geq 1} \left\{ \frac{|h_{2k}(z_0)|^2}{(1 + \lambda\sigma^2(2\pi k)^{2m})^l} + \frac{|h_{2k-1}(z_0)|^2}{(1 + \lambda\sigma^2(2\pi k)^{2m})^l} \right\} \\ &= \sigma^2 + 2\sigma^2 \sum_{k \geq 1} \frac{1}{(1 + \lambda\sigma^2(2\pi k)^{2m})^l} = \sigma^2 + 2\sigma^2 \sum_{k \geq 1} \frac{1}{(1 + (2\pi h^\dagger k)^{2m})^l}. \end{aligned}$$

By Lemma 6.1, we have  $c_0 = I_2/I_1$ . In particular,  $c_0 = 0.75$  (0.83) when  $m = 2$  (3).

To examine the pointwise asymptotic CI, we considered the true function  $g_0(z) = 0.6\beta_{30,17}(z) + 0.4\beta_{3,11}(z)$ , where  $\beta_{a,b}$  is the density function for  $Beta(a, b)$ , and estimated it using periodic splines with  $m = 2$ ;  $\sigma$  was chosen as 0.05. In Figure 1, we compare the coverage probability (CP) of our asymptotic CI (6.4), denoted ACI, Wahba's Bayesian CI (4.3), denoted WCI, and Nychka's Bayesian CI (4.7), denoted NCI, at thirty equally spaced grid points of  $\mathbb{I}$ . The CP was computed as the proportion of the CIs that cover  $g_0$  at each point based on 1,000 replications. We observe that, in general, all CIs exhibit similar patterns, e.g., undercoverage near peaks or troughs. However, when the sample size is sufficiently large, e.g.,  $n = 2000$ , the CP of ACI is uniformly closer to 95% than that of WCI and NCI in smooth regions such as  $[0.1, 0.4]$  and  $[0.8, 0.9]$ . We also report the average lengths of the three CIs in the titles of the plots. The ACI is the shortest, as indicated in Figure 1.

In Figure 3 of the supplementary document [42], we construct the SCB for  $g$  based on formula (5.9) by taking  $d_n = (-2 \log h)^{1/2}$ . We compare it with the *pointwise* confidence bands constructed by linking the endpoints of the ACI, WCI, and NCI at each observed covariate, denoted ACB, BCB1, and BCB2, respectively. The data were generated under the same setup as above. We

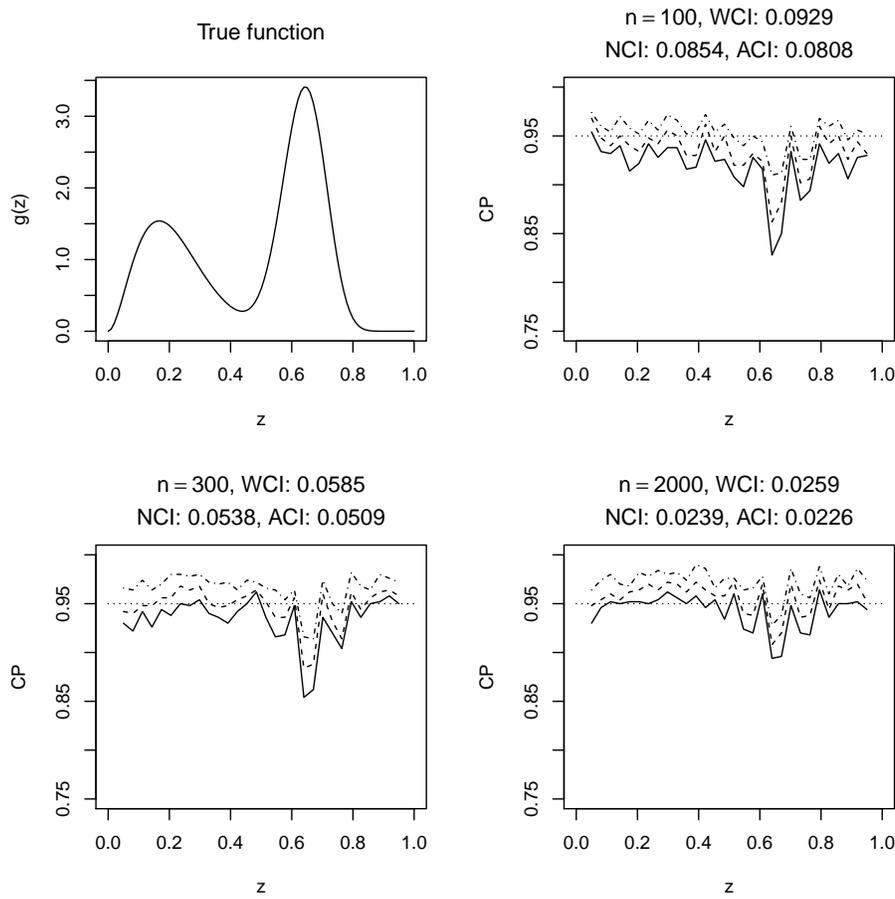


FIG 1. The first panel displays the true function  $g_0(z) = 0.6\beta_{30,17}(z) + 0.4\beta_{3,11}(z)$  used in Case (I) of Example 6.1. The other panels contain the coverage probabilities (CPs) of ACI (solid), NCI (dashed), and WCI (dotted dashed), and the average lengths of the three CIs (numbers in the plot titles). The CIs were built upon thirty equally spaced covariates.

observe that the coverage properties of all bands are reasonably good, and they become better as  $n$  grows. Meanwhile, the band areas, i.e., the areas covered by the bands, shrink to zero as  $n$  grows. We also note that the ACB has the smallest band area, while the SCB has the largest. This is because of the  $d_n$  factor in the construction of SCB; see Remark 5.1 for more details.

To conclude Case (I), we tested  $H_0 : g$  is linear at the 95% significance level by the PLRT and GLRT. By Lemma 6.1 and (6.2), direct calculation leads to  $r_K = 1.3333$  and  $u_n = 0.4714(h\sigma^{1/2})^{-1}$  when  $m = 2$ . The data were generated under the same setup except that different test functions  $g(z) = -0.5 + z + c(\sin(\pi z) - 0.5)$ ,  $c = 0, 0.5, 1.5, 2$ , were used for the purpose of the power comparison. For the GLRT method, the R function `glkerns()` provided in the `lokern` package (see [13]) was used for the local polynomial fitting based on the Epanechnikov kernel. For the PLRT method, GCV was used to select the smoothing parameter. Table 1 compares the power (the

proportion of rejections based on 1,000 replications) for  $n = 20, 30, 70, 200$ . When  $c \geq 1.5$  ( $c = 0$ ) and  $n = 70$  or larger, both testing methods achieve 100% power (5% correct level). We also observe that (i) the power increases as  $c$  increases, i.e., the test function becomes more nonlinear; and (ii) the PLRT shows moderate advantages over the GLRT, especially in small samples such as  $n = 20$ . An intuitive reason for (ii) is that the smoothing spline estimate in the PLRT uses the full data information, while the local polynomial estimate employed in the GLRT uses only local data information. Of course, as  $n$  grows, this difference rapidly vanishes because of the increasing data information.

$n$	100× Power%							
	$c = 0$		$c = 0.5$		$c = 1.5$		$c = 2$	
	PLRT	GLRT	PLRT	GLRT	PLRT	GLRT	PLRT	GLRT
20	18.60	20.10	28.40	30.10	89.60	86.30	97.30	96.10
30	13.60	14.40	33.00	30.60	98.10	96.80	99.60	99.60
70	8.30	9.40	54.40	48.40	100	100	100	100
200	5.20	5.50	95.10	92.70	100	100	100	100

TABLE 1

Power comparison of the PLRT and the GLRT in Case (I) of Example 6.1 where the test function is  $g_0(z) = -0.5 + z + c(\sin(\pi z) - 0.5)$  with various  $c$  values. The significance level is 95%.

**Case (II).**  $\mathcal{H} = H^m(\mathbb{I})$ : We used cubic splines and repeated most of the procedures in Case (I). A different true function  $g_0(z) = \sin(2.8\pi z)$  was chosen to examine the CIs. Figure 4 in the supplementary document [42] summarizes the SCB and the pointwise bands constructed by ACB, BCB1, and BCB2. In particular, BCB1 was computed by (4.2) and BCB2 was constructed by scaling the length of BCB1 by the factor  $\sqrt{27/32} \approx 0.919$ . We also tested the linearity of  $g_0$  at the 95% significance level, using the test functions  $g_0(z) = -0.5 + z + c(\sin(2.8\pi z) - 0.5)$ , for  $c = 0, 0.5, 1.5, 2$ . Table 2 compares the power of the PLRT and GLRT. From Figure 4 and Table 2, we conclude that all findings in Case (I) are also true in Case (II).

$n$	100× Power%							
	$c = 0$		$c = 0.5$		$c = 1.5$		$c = 2$	
	PLRT	GLRT	PLRT	GLRT	PLRT	GLRT	PLRT	GLRT
20	16.00	17.40	71.10	67.60	100	100	100	100
30	12.70	14.00	83.20	81.20	100	100	100	100
70	6.50	7.40	99.80	99.70	100	100	100	100
200	5.10	5.30	100	100	100	100	100	100

TABLE 2

Power comparison of the PLRT and the GLRT in Case (II) of Example 6.1 where the test function is  $g_0(z) = -0.5 + z + c(\sin(2.8\pi z) - 0.5)$  with various  $c$  values. The significance level is 95%.

EXAMPLE 6.2. (*Nonparametric Gamma Model*) Consider a two-parameter exponential model

$$Y|Z \sim \text{Gamma}(\alpha, \exp(g_0(Z))),$$

where  $\alpha > 0$ ,  $g_0 \in H_0^m(\mathbb{I})$ , and  $Z$  is uniform over  $[0, 1]$ . This framework leads to  $\ell(y; g(z)) = \alpha g(z) + (\alpha - 1) \log y - y \exp(g(z))$ . Thus,  $I(z) = \alpha$ , leading us to choose the trigonometric polynomial basis defined as in (6.2) with  $\sigma$  replaced with  $\alpha^{-1/2}$ , and the eigenvalues  $\gamma_0 = 0$  and  $\gamma_{2k} = \gamma_{2k-1} = \alpha^{-1}(2\pi k)^{2m}$  for  $k \geq 1$ . Local and global inference can be conducted similarly to Example 6.1.

EXAMPLE 6.3. (*Nonparametric Logistic Regression*) In this example, we consider the binary response  $Y \in \{0, 1\}$  modeled by the logistic relationship

$$(6.5) \quad P(Y = 1|Z = z) = \frac{\exp(g_0(z))}{1 + \exp(g_0(z))},$$

where  $g_0 \in H^m(\mathbb{I})$ . Given the length of this paper, we conducted simulations only for the ACI and PLRT. A straightforward calculation gives  $I(z) = \frac{\exp(g_0(z))}{(1 + \exp(g_0(z)))^2}$ , which can be estimated by  $\hat{I}(z) = \frac{\exp(\hat{g}_{n,\lambda}(z))}{(1 + \exp(\hat{g}_{n,\lambda}(z)))^2}$ . Given the estimate  $\hat{I}(z)$  and the marginal density estimate  $\hat{\pi}(z)$ , we find the approximate eigenvalues and eigenfunctions via (2.11).

The results are based on 1,000 replicated data sets drawn from (6.5), with  $n = 70, 100, 300, 500$ . To test whether  $g$  is linear, we considered two test functions,  $g_0(z) = -0.5 + z + c(\sin(\pi z) - 0.5)$  and  $g_0(z) = -0.5 + z + c(\sin(2.8\pi z) - 0.5)$ , for  $c = 0, 1, 1.5, 2$ . We use  $m = 2$ . Numerical calculations reveal that the eigenvalues are  $\gamma_\nu \approx (\alpha\nu)^{2m}$ , where  $\alpha = 4.40, 4.41, 4.47, 4.52$  and  $\alpha = 4.40, 4.44, 4.71, 4.91$  corresponding to the two test functions and the four values of  $c$ . This simplifies the calculations of  $\sigma_K^2$  and  $\rho_K^2$  defined in Theorem 5.3. For instance, when  $\gamma_\nu \approx (4.40\nu)^{2m}$ , using a result analogous to Lemma 6.1 we have  $\sigma_K^2 \approx 0.25$  and  $\rho_K^2 \approx 0.19$ . Then the quantities  $r_K$  and  $u_n$  are found for the PLRT method. To evaluate ACI, we considered the true function  $g_0(z) = (0.15)10^6 z^{11}(1 - z)^6 + (0.5)10^4 z^3(1 - z)^{10} - 1$ . The CP and the average lengths of the ACI are calculated at thirty evenly spaced points in  $\mathbb{I}$  under three sample sizes,  $n = 200, 500, 2000$ .

The results on the power of the PLRT are summarized in Tables 3 and 4, which demonstrate the validity of the proposed testing method. Specifically, when  $c = 0$ , the power reduces to the desired size 0.05; when  $c \geq 1.5$  and  $n \geq 300$ , the power approaches one. The results for the CPs and average lengths of ACIs are summarized in Figure 2. The CP uniformly approaches the desired 95% confidence level as  $n$  grows, showing the validity of the intervals.

$n$	100× Power%			
	$c = 0$	$c = 1$	$c = 1.5$	$c = 2$
70	4.10	16.90	30.20	50.80
100	4.50	17.30	38.90	63.40
300	5.00	52.50	92.00	99.30
500	5.00	79.70	99.30	100

TABLE 3

Power of PLRT in Example 6.3 where the test function is  $g_0(z) = -0.5 + z + c(\sin(\pi z) - 0.5)$  with various  $c$  values. The significance level is 95%.

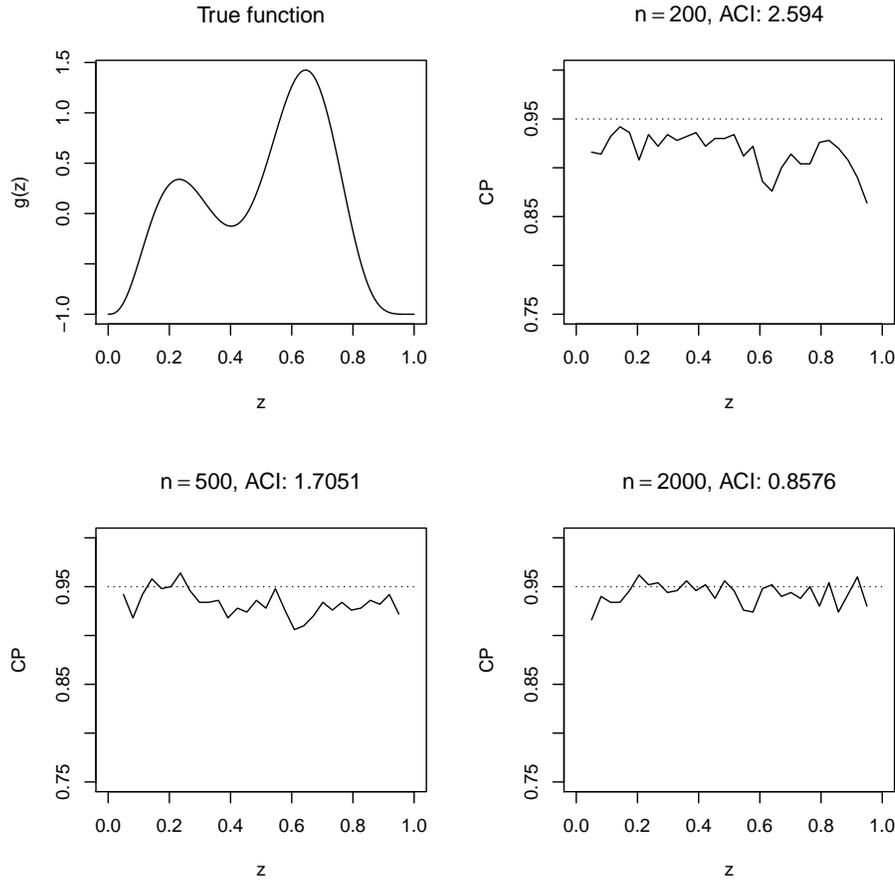


FIG 2. The first panel displays the true function  $g_0(z) = (0.15)10^6 z^{11}(1-z)^6 + (0.5)10^4 z^3(1-z)^{10} - 1$  used in Example 6.3. The other panels contain the CP and average length (number in the plot title) of each ACI. The ACIs were built upon thirty equally spaced covariates.

**Acknowledgement:** We are grateful for helpful discussions with Professor Chong Gu. The authors also thank the Co-editor Peter Hall, the Associate Editor, and two referees for insightful comments that led to important improvements in the paper.

## REFERENCES

- [1] Adams, R. A. (1975). *Sobolev Spaces*. Academic Press, New York-London. Pure and Applied Mathematics, Vol. 65.
- [2] Bahadur, R.R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics*, **37**, 577–580.
- [3] Banerjee, M. (2007). Likelihood based inference for monotone response models. *Annals of Statistics*, **35**, 931–956.
- [4] Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics*, **1**, 1071–1095.
- [5] Birkhoff, D. (1908). Boundary value and expansion problems of ordinary linear differential equations. *Transactions of the American Mathematical Society*, **9**, 373–395.

$n$	100× Power%			
	$c = 0$	$c = 1$	$c = 1.5$	$c = 2$
70	4.10	56.20	90.10	99.00
100	5.00	71.90	96.90	100
300	5.00	99.80	100	100
500	5.00	100	100	100

TABLE 4

Power of PLRT in Example 6.3 where the test function is  $g_0(z) = -0.5 + z + c(\sin(2.8\pi z) - 0.5)$  with various  $c$  values. The significance level is 95%.

- [6] Chen, J.C. (1994). Testing goodness of fit of polynomial models via spline smoothing techniques. *Statistics & Probability Letters*, **19**, 65–76.
- [7] Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, **6**, 1852–1884.
- [8] Coddington, E. A. and Levinson, N. (1987). *Theory of Ordinary Differential Equations*. Tata McGraw-Hill Publishing CO. LTD, New Delhi.
- [9] Cox, D. and O’Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics*, **18**, 1676–1695.
- [10] Cox, D., Koh, E., Wahba, G. and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semi-parametric) partial and generalized spline models. *Annals of Statistics*, **16**, 113–119.
- [11] Davis, P. J. (1963). *Interpolation and Approximation*. Blaisdell, New York.
- [12] de Jong, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory & Related Fields*, **75**, 261–277.
- [13] Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Graphical and Computational Statistics*, **6**, 35C–654.
- [14] Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715–731.
- [15] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, **1**, 153–193.
- [16] Fan, J. and Zhang, J. (2004). Sieve empirical likelihood ratio tests for nonparametric functions. *Annals of Statistics*, **32**, 1858–1907.
- [17] Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *Annals of Statistics*, **36**, 875–905.
- [18] Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Annals of Statistics*, **21**, 217–234.
- [19] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag.
- [20] Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *Journal of Multivariate Analysis*, **29**, 163–179.
- [21] Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics*, **22**, 215–232.
- [22] Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics*, **20**, 675–694.
- [23] Ingster, Yu I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I–III. *Mathematical Methods of Statistics*, **2**, 85–114; **3**, 171–189; **4**, 249–268.

- [24] Jayasuriya, B. R. (1996). Testing for polynomial regression using nonparametric regression techniques. *Journal of the American Statistical Association*, **91**, 1626–1630.
- [25] Johnson, W. P. (2002). The curious history of Faá di Bruno’s formula. *American Mathematical Monthly*, **109**, 217–234.
- [26] Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd eds. Springer.
- [27] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York.
- [28] Krivobokova, T., Kneib, T. and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, **105**, 852–863.
- [29] Ke, C. and Wang, Y. (2002) ASSIST: A Suite of S-plus functions Implementing Spline smoothing Techniques. Preprint.
- [30] Liu, A. and Wang, Y. (2002). Hypothesis testing in smoothing spline models. *Journal of Statistical Computation and Simulation*, **74**, 581–597.
- [31] Lorentz, G. G. (1966). *Approximation of Functions*. Holt, Rinehart and Winston, Inc.
- [32] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition. London: Chapman and Hall.
- [33] Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics*, **25**, 1014–1035.
- [34] Messer, K. and Goldstein, L. (1993). A new class of kernels for nonparametric curve estimation. *Annals of Statistics*, **21**, 179–195.
- [35] Neumann, H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, **9**, 307–333.
- [36] Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, **83**, 1134–1143.
- [37] Nychka, D. (1995). Splines as local smoothers. *Annals of Statistics*, **23**, 1175–1197.
- [38] Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Annals of Probability*, **22**, 1679–1706.
- [39] Ramil-Novo, L.A. and GonzKalez-Manteiga, W. (2000). F-tests and regression ANOVA based on smoothing spline estimators. *Statistica Sinica*, **10**, 819 – 837.
- [40] Rice, J. and Rosenblatt, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Annals of Statistics*, **11**, 141–156.
- [41] Shang, Z. (2010). Convergence rate and Bahadur type representation of general smoothing spline M-Estimates. *Electronic Journal of Statistics*, **4**, 1411–1442.
- [42] Shang, Z. and Cheng, G. (2012). Online supplementary to “local and global asymptotic inference in smoothing spline models”.
- [43] Shao, J. (2003). *Mathematical Statistics*, 2nd Ed. Springer Texts in Statistics. Springer, New York.
- [44] Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, **12**, 898–916.
- [45] Stone, M. H. (1926). A comparison of the series of Fourier and Birkhoff. *Transactions of the American Mathematical Society*, **28**, 695–761.
- [46] Sun, J. and Loader, C. (1994). Simultaneous confidence bands for linear regression and smoothing. *Annals of Statistics*, **3**, 1328–1345.

- [47] Sun, J., Loader, C. and McCormick, W. P. (2000). Confidence bands in generalized linear models. *Annals of Statistics*, **28**, 429–460.
- [48] Tusnady, G. (1977). A remark on the approximation of the sample distribution function in the multidimensional case. *Periodica Mathematica Hungarica*, **8**, 53–55.
- [49] Utreras, F. I. (1988). Boundary effects on convergence rates for Tikhonov regularization. *Journal of Approximation Theory*, **54**, 235–249.
- [50] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- [51] Wahba, G. (1983). Bayesian “confidence intervals” for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, **45**, 133–150.
- [52] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.
- [53] Wang, Y. (2011). *Smoothing Splines: Methods and Applications Monographs on Statistics & Applied Probability*. Chapman & Hall/CRC.
- [54] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.
- [55] Zhang, W. and Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalized varying-coefficient models. *Journal of Multivariate Analysis*, **101**, 1656–1680.
- [56] Zhou, S., Shen, X. and D. A. Wolfe. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, **5**, 1760–1782.