Submitted to the Annals of Statistics

JOINT ASYMPTOTICS FOR SEMI-NONPARAMETRIC MODELS UNDER PENALIZATION

By Guang Cheng[‡] and Zuofeng Shang[§]

Department of Statistics, Purdue University

November 11, 2013

We consider a joint asymptotic framework for studying semi-nonparametric models where (finite dimensional) Euclidean parameters and (infinite dimensional) functional parameters are both of interest. A class of generalized partially linear models is used as a prototypical example (under the penalized quasi-likelihood estimation). We first show that the Euclidean estimator and (pointwise) functional estimator, which are re-scaled at different rates, jointly converge to a Gaussian vector. This weak convergence result reveals a surprising *joint asymptotics phenomenon*: these two estimators become asymptotically independent while the Euclidean estimator achieves the semiparametric efficiency bound. A major goal of this paper is to provide theoretical insights into the above phenomenon. We next consider likelihood ratio testing for a set of joint local hypotheses. A semi-nonparametric version of the Wilks phenomenon is unveiled. Our joint asymptotic results are challenging to establish because of the rather different nature of the two model components: parametric vs. nonparametric.

1. Introduction. In the literature, a statistical model is called *semi-nonparametric* if it contains both finite-dimensional and infinite-dimensional unknown parameters of interest. An example is the semi-nonparametric copula models that can be applied to address tail dependence among shocks to different financial series and also to recover the shape of the "news impact curve" for individual financial series. Another example is the semi-nonparametric binary regression models proposed by Banerjee et al. [3] to define the conditional probability of attending primary school in Indian villages through an appropriate link function influenced by a set of covariates such as gender and household income. As a first step in exploring the joint asymptotics, we mainly focus

[‡]Corresponding Author. Associate Professor, Research Sponsored by NSF, DMS-0906497, CAREER Award DMS-1151692

[§]Visiting Assistant Professor

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

Keywords and phrases: joint asymptotics, joint Bahadur representation, local likelihood ratio test, semi-nonparametric models, smoothing spline

on semi-nonparametric regression models with a partially linear structure.

The existing semiparametric literature is concerned with asymptotic theories and inference procedures for the Euclidean parameter only. The functional parameter is profiled out as an infinite-dimensional nuisance parameter; see [5, 22, 7, 8, 9, 26]. In the special case where both parameters are estimable at the same root-n convergence rate, e.g., [18, 19], we can combine them as an infinite-dimensional parameter and then apply the functional Z-estimation theorem, e.g., Theorem 3.3.1 in [28], to study their joint asymptotic distribution. However, it is common for the two parameters to be estimated at different parametric and nonparametric rates. Their radically different parameter dimensionality also poses technical challenges for the construction of valid procedures for joint inference. In this paper, we employ modern empirical processes theory for dealing with this nontrivial issue in a general setup. As far as we are aware, our general joint asymptotic theories and inference are new. The only relevant reference of which we are aware, i.e., [24], is concerned with a fully parametric setting.

Within the penalized estimation framework, we derive a joint limit distribution for the Euclidean estimator and the functional estimator, which are re-scaled according to their different convergence rates, as a zero-mean Gaussian vector. One surprising result is that these two estimators become asymptotically independent while the Euclidean estimator achieves the semiparametric efficiency bound. This asymptotic independence will prove to be useful in making joint inference, e.g., joint confidence intervals (CIs). Under similar (essentially weaker) conditions, the marginal limit distribution for the Euclidean estimator coincides with that derived in [20]. On the other hand, we observe that the (pointwise) marginal asymptotic results for the nonparametric component are generally different from those derived in the purely nonparametric setup (without the Euclidean parameter), i.e., [27], even though the Euclidean parameter is estimated at a faster rate; see Remark 5.1. This conclusion is counterintuitive.

We next consider likelihood ratio testing for a variety of joint local hypotheses such as $H_0 : \theta = \theta_0$ & $g(z_0) = w_0$ and $H_0 : x^T\theta + g(z_0) = \alpha$, where θ and g denote the parametric and nonparametric components, respectively. Conventional semiparametric testing focuses on only the parametric components; see [22, 9]. However, in practice, it is of great interest to evaluate the nonparametric components at the same time. For example, we may test the joint effect of child gender and household income, which are respectively modeled by θ and g, on the probability of attending primary school in the Indian schooling model. Testing the joint hypothesis also provides a method for constructing the joint confidence intervals without estimating the asymptotic variances. We show that the null limit distribution is a mixture of two independent Chi-square distributions that are contributed by the parametric and nonparametric components, respectively. Therefore, we have unveiled a semi-nonparametric version of the Wilks phenomenon (meaning that the asymptotic null distribution is free of nuisance parameters) arising from the proposed joint local

testing. As far as we are aware, this result is new. The only relevant paper of which we are aware is [3], which considers two separate null hypotheses, i.e., $H_{01} : \theta = \theta_0$ and $H_{02} : g(z_0) = w_0$, under the monotonicity constraint of $g(\cdot)$.

Our paper is mainly concerned with penalized estimation. One possible extension is to sieve estimation because of the known duality between these two regularization methods. However, it is more interesting to consider the estimation without any regularization, e.g., under shape constraints. In this case, the joint asymptotic theory is expected to be more intriguing in the sense that the nonparametric estimate may exhibit nonstandard asymptotic behavior, e.g., [16, 6]. Another research direction is the development of joint *global* inference, e.g., joint global testing, which may require very different techniques. Extensions to more complicated models such as partially linear additive models and partially linear Cox models, e.g., [10], are conceptually feasible by modifications to our JBR techniques.

The rest of this paper is organized as follows. Section 2 introduces the model assumptions and builds the theoretical foundation. Sections 3 and 4 formally discuss the joint limit distribution and joint local hypothesis testing, respectively. In Section 5, we give three concrete examples with extensive simulations to illustrate our theory. The proofs are postponed to the Appendix or the online supplementary document.

2. Preliminaries. This section introduces the model assumptions and establishes the theoretical foundation of our results in two layers: (i) the partially linear extension of reproducing kernel Hilbert space (RKHS) theory; (ii) the joint Bahadur representation. Both technical results are of independent interest.

2.1. *Notation and Model Assumptions.* Suppose that the data $T_i = (Y_i, X_i, Z_i)$, $i = 1, \ldots, n$, are *i.i.d.* copies of $T = (Y, X, Z)$, where $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the response variable, $U = (X, Z) \in \mathcal{U} \equiv \mathbb{I}^p \times \mathbb{I}$ is the covariate variable, and $\mathbb{I} = [0, 1]$. Consider a general class of semi-nonparametric regression models under the following partially linear structure:

$$\mu_0(U) \equiv E(Y|U) = F(X^T\theta_0 + g_0(Z)), \tag{2.1}$$

where $F(\cdot)$ is some known link function and $g_0(\cdot)$ is some unknown smooth function. This primary assumption covers two classes of statistical models. The first class is called *generalized partially linear models* ([4]); here the data are modeled by $y|u \sim p(y; \mu_0(u))$ for a conditional distribution p. Instead of assuming the underlying distribution, the second class specifies only the relationship between the conditional mean and the conditional variance: $Var(Y|U) = \mathcal{V}(\mu_0(U))$ for some known positive-valued function $\mathcal{V}$. The nonparametric estimation of g in the second situation uses the quasi-likelihood $Q(y; \mu) \equiv \int_y^\mu (y-s)/\mathcal{V}(s)ds$ with $\mu = F(x^T\theta+g(z))$ ([30]). Despite the distinct modeling principles, these two classes have a large overlap under many common combinations of

$(F, \mathcal{V})$, as summarized in Table 2.1 of [21]. From now on, we work with a general criterion function $\ell(y; a) : \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}$, which can represent either $\log p(y; F(a))$ or $Q(y; F(a))$.

Let the full parameter space for $f \equiv (\theta, g)$ be $\mathcal{H} \equiv \mathbb{R}^p \times H^m(\mathbb{I})$, where $H^m(\mathbb{I})$ is the mth order Sobolev space defined as

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

With some abuse of notation, $\mathcal{H}$ may also refer to $\mathbb{R}^p \times H_0^m(\mathbb{I})$, where $H_0^m(\mathbb{I})$ is a homogeneous subspace 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 additional restrictions $g^{(j)}(0) = g^{(j)}(1)$ for $j = 0, 1, \ldots, m-1$. Throughout this paper we assume $m > 1/2$ to be known. Consider the penalized semi-nonparametric estimator

$$(\widehat{\theta}_{n,\lambda}, \widehat{g}_{n,\lambda}) = \arg \max_{(\theta,g)\in\mathcal{H}} \ell_{n,\lambda}(f) = \arg \max_{(\theta,g)\in\mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i; X_i^T\theta + g(Z_i)) - (\lambda/2) J(g, g) \right\}, \tag{2.2}$$

where $J(g, \widetilde{g}) = \int_{\mathbb{I}} g^{(m)}(z) \widetilde{g}^{(m)}(z) dz$ and $\lambda \to 0$ as $n \to \infty$. Here, we use $\lambda/2$ (rather than λ) to simplify future expressions. Write $\widehat{f}_{n,\lambda} = (\widehat{\theta}_{n,\lambda}, \widehat{g}_{n,\lambda})$. The existence of $\widehat{g}_{n,\lambda}$ is guaranteed by Theorem 2.9 of [14] when the null space $\mathcal{N}_m \equiv \{g \in H^m(\mathbb{I}) : J(g, g) = 0\}$ is finite dimensional and $\ell(y; a)$ is concave and continuous w.r.t. a.

We next assume some basic model conditions. For simplicity, throughout the paper we do not distinguish $f = (\theta, g) \in \mathcal{H}$ from its associated function $f \in \mathcal{F} \equiv \{f(x, z) = x^T\theta + g(z) : (\theta, g) \in \mathcal{H}, (x, z) \in \mathcal{U}\}$. Let $\mathcal{I}_0$ be the range for the true function $f_0(x, z) \in \mathcal{F}$, i.e., a compact interval. Denote the first-, second-, and third-order derivatives of $\ell(y; a)$ (w.r.t. a) by $\dot{\ell}_a$, $\ddot{\ell}_a$, and ℓ'''_a.

ASSUMPTION A1. *(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.*

$$E\left\{ \exp(\sup_{a\in\mathcal{I}} |\ddot{\ell}_a(Y; a)|/C_0) \middle| U \right\} \le C_1, \ \ a.s. \tag{2.3}$$

and

$$E\left\{ \exp(\sup_{a\in\mathcal{I}} |\ell'''_a(Y; a)|/C_0) \middle| U \right\} \le C_1, \ \ a.s.. \tag{2.4}$$

(b) There exists a positive constant C_2 s.t. $C_2^{-1} \le I(U) \equiv -E(\ddot{\ell}_a(Y; X^T\theta_0 + g_0(Z))|U) \le C_2$, a.s.
(c) $\epsilon \equiv \dot{\ell}_a(Y; X^T\theta_0 + g_0(Z))$ satisfies $E(\epsilon|U) = 0$ and $E(\epsilon^2|U) = I(U)$, a.s.

A detailed discussion of the above model assumptions can be found in [27]. In particular, Assumption A1 (a) is typically used in semiparametric quasi-likelihood models; see [20].

Hereinafter, if for positive sequences a_μ and b_μ we have that a_μ/b_μ tends to a strictly positive constant, we write $a_\mu \asymp b_\mu$. If that constant is one, we write $a_\mu \sim b_\mu$. Let $\sum_\nu$ denote the sum over

$\nu \in \mathbb{N} = \{0, 1, 2, \ldots\}$ for convenience. Let the sup-norm of $g \in H^m(\mathbb{I})$ be $\|g\|_{\sup} = \sup_{z\in\mathbb{I}} |g(z)|$. Let λ^* be the optimal smoothing parameter; $\lambda^* \asymp n^{-2m/(2m+1)}$. For simplicity, we write $\lambda^{1/(2m)}$ as h, and thus $h^* \asymp n^{-1/(2m+1)}$.

2.2. *A Partially Linear Extension of RKHS Theory.* In this section, we adapt the nonparametric RKHS framework to our semi-nonparametric setup.

We define the inner product for $\mathcal{H}$ to be, for any $(\theta, g), (\widetilde{\theta}, \widetilde{g}) \in \mathcal{H}$,

$$\langle (\theta, g), (\widetilde{\theta}, \widetilde{g}) \rangle = E_U\{I(U)(X^T\theta + g(Z))(X^T\widetilde{\theta} + \widetilde{g}(Z))\} + \lambda J(g, \widetilde{g}), \tag{2.5}$$

and we define the norm to be $\|(\theta, g)\|^2 = \langle (\theta, g), (\theta, g) \rangle$. Under this Sobolev norm, we will construct two linear operators, $R_u : \mathcal{U} \mapsto \mathcal{H}$ and $P_\lambda : \mathcal{H} \mapsto \mathcal{H}$ satisfying

$$\langle R_u, f \rangle = x^T\theta + g(z) \text{ for any } u \in \mathcal{U} \text{ and } f \in \mathcal{H} \tag{2.6}$$

and

$$\langle P_\lambda f, \widetilde{f} \rangle = \lambda J(g, \widetilde{g}) \text{ for any } f = (\theta, g), \widetilde{f} = (\widetilde{\theta}, \widetilde{g}) \in \mathcal{H}. \tag{2.7}$$

As will be seen, R_u and P_λ are major building blocks of this generalized RKHS framework. In particular, Propositions 2.1 and 2.2 show that these two operators are actually built upon their nonparametric counterparts K_z and W_λ defined below.

Let $K(z_1, z_2)$ be a (symmetric) reproducing kernel of $H^m(\mathbb{I})$ endowed with the inner product $\langle g, \widetilde{g} \rangle_1 = E_Z\{B(Z)g(Z)\widetilde{g}(Z)\} + \lambda J(g, \widetilde{g})$ and norm $\|g\|_1^2 = \langle g, g \rangle_1$, where $B(Z) = E\{I(U)|Z\}$. Hence, $K_z(\cdot) \equiv K(z, \cdot)$ satisfies $\langle K_z, g \rangle_1 = g(z)$. We next define a positive definite self-adjoint operator $W_\lambda : H^m(\mathbb{I}) \mapsto H^m(\mathbb{I})$ satisfying $\langle W_\lambda g, \widetilde{g} \rangle_1 = \lambda J(g, \widetilde{g})$ for any $g, \widetilde{g} \in H^m(\mathbb{I})$. Write $V(g, \widetilde{g}) = E_Z\{B(Z)g(Z)\widetilde{g}(Z)\}$. Hence, $\langle g, \widetilde{g} \rangle_1 = V(g, \widetilde{g}) + \langle W_\lambda g, \widetilde{g} \rangle_1$, which implies

$$V(g, \widetilde{g}) = \langle (id - W_\lambda)g, \widetilde{g} \rangle_1, \tag{2.8}$$

where id denotes the identity operator. We next assume that there exists a sequence of basis functions in the space $H^m(\mathbb{I})$ that simultaneously diagonalizes the bilinear forms V and J. Such an eigensystem assumption is typical in the smoothing spline literature; see [14].

Assumption A2. *There exists a sequence of eigenfunctions $h_\nu \in H^m(\mathbb{I})$, $\nu \in \mathbb{N}$ satisfying $\sup_{\nu\in\mathbb{N}} \|h_\nu\|_{\sup} < \infty$ and a nondecreasing sequence of eigenvalues $\gamma_\nu \asymp \nu^{2m}$ such that $V(h_\mu, h_\nu) = \delta_{\mu\nu}$ and $J(h_\mu, h_\nu) = \gamma_\mu\delta_{\mu\nu}$ for any $\mu, \nu \in \mathbb{N}$, where $\delta_{\mu\nu}$ is the Kronecker's delta. Furthermore, any $g \in H^m(\mathbb{I})$ admits the Fourier expansion $g = \sum_\nu V(g, h_\nu)h_\nu$ under the $\|\cdot\|_1$-norm.*

Under Assumption A2, we can easily derive explicit expressions for $\|g\|_1$, $W_\lambda h_\nu(\cdot)$, and $K_z(\cdot)$ in terms of the eigenvalues and eigenfunctions as follows:

$$\|g\|_1^2 = \sum_\nu |V(g, h_\nu)|^2(1 + \lambda\gamma_\nu), \; W_\lambda h_\nu(\cdot) = \frac{\lambda\gamma_\nu}{1 + \lambda\gamma_\nu} h_\nu(\cdot), \text{ and } K_z(\cdot) = \sum_\nu \frac{h_\nu(z)}{1 + \lambda\gamma_\nu} h_\nu(\cdot). \tag{2.9}$$

Using similar arguments to those in Proposition 2.2 of [27], we know that Assumption A2 holds when the h_νs are chosen as the (normalized) solutions of the following ODE problem:

$$(-1)^m h_\nu^{(2m)}(\cdot) = \gamma_\nu B(\cdot)\pi(\cdot)h_\nu(\cdot),\ h_\nu^{(j)}(0) = h_\nu^{(j)}(1) = 0,\ j = m, m+1, \ldots, 2m-1, \tag{2.10}$$

where $\pi(\cdot)$ is the marginal density function of the covariate Z. For example, the h_νs are constructed as an explicit trigonometric basis in Case (I) of Example 5.1.

We next state a regularity assumption, A3, guaranteeing that R_u and P_λ are both well defined. Define $A_0(Z) = E\{I(U)X|Z\}$ and $G(Z) = A_0(Z)/B(Z)$. Note that $G = (G_1, \ldots, G_p)^T$ is a p-dimensional vector-valued function, e.g., $G(Z) = E(X|Z)$ in the L_2 regression.

ASSUMPTION A3. *$G_1, \ldots, G_p \in L_2(P_Z)$, i.e., G_k has a finite second moment, and the $p \times p$ matrix $\Omega \equiv E\{I(U)(X - G(Z))(X - G(Z))^T\}$ is positive definite.*

Under the assumption that $G_k \in L_2(P_Z)$, the linear functional $\mathcal{A}_k$ defined by $\mathcal{A}_k g = V(G_k, g)$ is bounded (or equivalently, continuous) for any $g \in H^m(\mathbb{I})$ because of the following inequality: $|\mathcal{A}_k g| \le V^{1/2}(G_k, G_k)V^{1/2}(g, g) \le V^{1/2}(G_k, G_k)\|g\|_1 < \infty$. Thus, by Riesz's representation theorem, there exists an $A_k \in H^m(\mathbb{I})$ such that $\mathcal{A}_k g = \langle A_k, g\rangle_1$ for any $g \in H^m(\mathbb{I})$. Thus, if we write $A = (A_1, \ldots, A_p)^T$, then

$$V(G, g) = \langle A, g\rangle_1. \tag{2.11}$$

We also note that $A = (id - W_\lambda)G$ when $G_1, \ldots, G_p \in H^m(\mathbb{I})$ based on (2.8). Taking $g = K_z$ in (2.11) and applying (2.9), we find that

$$A(z) = \sum_\nu \frac{V(G, h_\nu)}{1 + \lambda\gamma_\nu} h_\nu(z) \ \text{ and } \ (W_\lambda A)(z) = \sum_\nu \frac{V(G, h_\nu)\lambda\gamma_\nu}{(1 + \lambda\gamma_\nu)^2} h_\nu(z). \tag{2.12}$$

Now, we are ready to construct R_u and P_λ in Propositions 2.1 and 2.2, respectively. Define $\Sigma_\lambda = E_Z\{B(Z)G(Z)(G(Z) - A(Z))^T\}$ as a $p \times p$ matrix.

PROPOSITION 2.1. *R_u defined in (2.6) can be expressed as $R_u : u \mapsto (H_u, T_u) \in \mathcal{H}$, where*

$$H_u = (\Omega + \Sigma_\lambda)^{-1}(x - A(z)) \ \text{ and } \ T_u = K_z - A^T(\Omega + \Sigma_\lambda)^{-1}(x - A(z)). \tag{2.13}$$

PROPOSITION 2.2. *P_λ defined in (2.7) can be expressed as $P_\lambda : (\theta, g) \mapsto (H_g^*, T_g^*) \in \mathcal{H}$, where*

$$\begin{cases} H_g^* = -(\Omega + \Sigma_\lambda)^{-1}E\{B(Z)G(Z)(W_\lambda g)(Z)\} \\ T_g^* = E\{B(Z)G(Z)^T(W_\lambda g)(Z)\}(\Omega + \Sigma_\lambda)^{-1}A + W_\lambda g. \end{cases}$$

Note that $\lim_{\lambda\to 0}\Sigma_\lambda = 0$ according to (A.2) in the Appendix. Therefore, $(\Omega+\Sigma_\lambda)^{-1}$ above is well defined under Assumption A3. In addition, we note that P_λ is self-adjoint and bounded because of the following inequality:

$$\|P_\lambda f\| = \sup_{\|\widetilde{f}\|=1} |\langle P_\lambda f, \widetilde{f}\rangle| = \sup_{\|\widetilde{f}\|=1} |\lambda J(g,\widetilde{g})| \le \sqrt{\lambda J(g,g)} \sup_{\|\widetilde{f}\|=1} \sqrt{\lambda J(\widetilde{g},\widetilde{g})} \le \|f\|. \tag{2.14}$$

Finally, we derive the Fréchet derivatives of $\ell_{n,\lambda}(f)$ defined in (2.2). Let $\Delta f, \Delta f_j \in \mathcal{H}$ for $j = 1, 2, 3$. The Fréchet derivative of $\ell_{n,\lambda}(f)$ is given by

$$\begin{aligned} D\ell_{n,\lambda}(f)\Delta f &= \frac{1}{n}\sum_{i=1}^n \dot{\ell}_a(Y_i; X_i^T\theta + g(Z_i))\langle R_{U_i}, \Delta f\rangle - \langle P_\lambda f, \Delta f\rangle \\ &\equiv \langle S_n(f), \Delta f\rangle - \langle P_\lambda f, \Delta f\rangle \equiv \langle S_{n,\lambda}(f), \Delta f\rangle. \end{aligned}$$

Note that $S_{n,\lambda}(\widehat{f}_{n,\lambda}) = 0$. In particular, $S_{n,\lambda}(f_0)$ is of interest, and it can be expressed as

$$S_{n,\lambda}(f_0) = \frac{1}{n}\sum_{i=1}^n \epsilon_i R_{U_i} - P_\lambda f_0. \tag{2.15}$$

The Frechét derivative of $S_{n,\lambda}$ ($DS_{n,\lambda}$) is denoted $DS_{n,\lambda}(f)\Delta f_1\Delta f_2$ ($D^2S_{n,\lambda}(f)\Delta f_1\Delta f_2\Delta f_3$) and can be explicitly calculated as $(1/n)\sum_{i=1}^n \ddot{\ell}_a(Y_i; X_i^T\theta+g(Z_i))\langle R_{U_i}, \Delta f_1\rangle\langle R_{U_i}, \Delta f_2\rangle - \langle P_\lambda \Delta f_1, \Delta f_2\rangle$ ($(1/n)\sum_{i=1}^n \ell'''_a(Y_i; X_i^T\theta+g(Z_i))\langle R_{U_i}, \Delta f_1\rangle\langle R_{U_i}, \Delta f_2\rangle\langle R_{U_i}, \Delta f_3\rangle$). Define $S(f) = E\{S_n(f)\}$, $S_\lambda(f) = S(f) - P_\lambda f$, and $DS_\lambda(f) = DS(f) - P_\lambda$, where $DS(f)\Delta f_1\Delta f_2 = E\{\ddot{\ell}_a(Y; X^T\theta+g(Z))\langle R_U, \Delta f_1\rangle\langle R_U, \Delta f_2\rangle\}$. Since $\langle DS_\lambda(f_0)f, \widetilde{f}\rangle = -\langle f, \widetilde{f}\rangle$ for any $f, \widetilde{f} \in \mathcal{H}$, we have the following result:

PROPOSITION 2.3. *$DS_\lambda(f_0) = -id$, where id is the identity operator on $\mathcal{H}$.*

2.3. *Joint Bahadur Representation.* This section presents the second layer of our theoretical foundation: the *joint Bahadur representation (JBR)*. The JBR is developed based on empirical processes theory and will prove to be a powerful tool in the study of the joint asymptotics.

We start with a useful lemma stating the relationship between $\|f\|$ and $\|f\|_{\sup}$, where the former $f = (\theta, g)$ and the latter $f = x^T\theta + g(z)$.

LEMMA 2.4. *There exists a constant $c_m > 0$ such that $\|R_u\| \le c_m h^{-1/2}$ and $|f(u)| \le c_m h^{-1/2}\|f\|$ for any $u \in \mathcal{U}$ and $(\theta, g) \in \mathcal{H}$. In particular, c_m does not depend on the choice of u and (θ, g). Hence, $\|f\|_{\sup} \le c_m h^{-1/2}\|f\|$.*

An additional convergence-rate condition is needed to obtain JBR. Assumption A4 implies that $\widehat{f}_{n,\lambda}$ achieves the optimal rate of convergence, i.e., $O_P(n^{-m/(2m+1)})$, when $\lambda = \lambda^*$.

ASSUMPTION A4. $\|\widehat{f}_{n,\lambda} - f_0\| = O_P((nh)^{-1/2} + h^m)$.

When Assumption A1 holds and $\lambda = \lambda^*$, we can show that the above rate condition can be reduced to the consistency of $\widehat{f}_{n,\lambda}$ in terms of another Sobolev norm:

$$\|f\|_{\mathcal{H}} = E_U\{I(U)(X^T\theta + g(Z))^2\} + J(g,g), \tag{2.16}$$

which is commonly used in functional analysis.

PROPOSITION 2.5. *Suppose that Assumption A1 holds, and further that* $\|\widehat{f}_{n,\lambda} - f_0\|_{\mathcal{H}} = o_P(1)$. *If* h *satisfies* $(n^{1/2}h)^{-1}(\log\log n)^{m/(2m-1)}(\log n)^{2m/(2m-1)} = o(1)$, *then Assumption A4 is valid.*

The following *joint Bahadur representation* can be viewed as a nontrivial extension that adds a functional parameter to the traditional Bahadur representation for parametric models ([2]).

THEOREM 2.6. *(Joint Bahadur Representation) Suppose that Assumptions A1 through A4 hold,* $h = o(1)$, *and* $nh^2 \to \infty$. *Recall that* $S_{n,\lambda}(f_0)$ *is defined in (2.15). Then we have*

$$\|\widehat{f}_{n,\lambda} - f_0 - S_{n,\lambda}(f_0)\| = O_P(a_n \log n), \tag{2.17}$$

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* $C_\ell = \sup_{u\in\mathcal{U}} E\{\sup_{a\in\mathcal{I}} |\ell'''_a(Y;a)|\big|U = u\}$.

The proof of Theorem 2.6 relies heavily on modern empirical process theory, and in particular a *concentration inequality* given in the supplementary material.

3. Joint Limit Distribution. We prove our joint asymptotic theory in two steps. We start from a preliminary result that for any given $z_0 \in \mathbb{I}$, $(\sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0^*), \sqrt{nh}(\widehat{g}_{n,\lambda} - g_0^*)(z_0))$ weakly converges to a zero-mean Gaussian vector, i.e., Theorem 3.1. Unfortunately, the center $(\theta_0^*, g_0^*) \equiv f_0 - P_\lambda f_0$ is not unbiased. While removing the estimation bias for θ_0 (so that $\widehat{\theta}_{n,\lambda}$ achieves the semiparametric efficiency bound), we discover a surprising phenomenon: $\widehat{\theta}_{n,\lambda}$ *and* $\widehat{g}_{n,\lambda}(z_0)$ *become asymptotically independent at the same time.* This leads to a new form of the joint limit distribution given in Theorem 3.3. It is worth noting that all the above results are obtained simply by imposing the regularity conditions in the semiparametric literature; see [22, 11]. We compare our new results with the classical results in the semiparametric ([20]) and nonparametric ([27]) literature; see Remarks 3.1 and 3.2. Finally, we illustrate explicit forms of the asymptotic bias and covariance matrix for the partial smoothing spline model.

THEOREM 3.1. *(Joint Asymptotic Theory I) Let Assumptions A1 through A4 be satisfied. Suppose that as* $n \to \infty$, $h = o(1)$, $nh^2 \to \infty$, *and* $a_n \log n = o(n^{-1/2}h^{1/2})$, *where* a_n *is defined as in (2.17). Furthermore, assume that, as* $n \to \infty$,

$$hV(K_{z_0}, K_{z_0}) \to \sigma^2_{z_0},\ h^{1/2}(W_\lambda A)(z_0) \to \alpha_{z_0} \in \mathbb{R}^p,\ \text{and } h^{1/2}A(z_0) \to -\beta_{z_0} \in \mathbb{R}^p, \tag{3.1}$$

where A is the Riesz representer defined in (2.11). Then we have, for any $z_0 \in \mathbb{I}$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0^*) \\ \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0^*(z_0)) \end{pmatrix} \xrightarrow{d} N(0, \Psi^*), \tag{3.2}$$

where

$$\Psi^* = \begin{pmatrix} \Omega^{-1} & \Omega^{-1}(\alpha_{z_0} + \beta_{z_0}) \\ (\alpha_{z_0} + \beta_{z_0})^T \Omega^{-1} & \sigma_{z_0}^2 + 2\beta_{z_0}^T \Omega^{-1} \alpha_{z_0} + \beta_{z_0}^T \Omega^{-1} \beta_{z_0} \end{pmatrix}. \tag{3.3}$$

Note that Ω^{-1} is well defined under Assumption A3. It follows from (2.9) and (2.12) that

$$\begin{aligned} \sigma_{z_0}^2 &= \lim_{h \to 0} \sum_{\nu} \frac{h |h_\nu(z_0)|^2}{(1 + \lambda \gamma_\nu)^2}, \\ \alpha_{z_0} &= \lim_{h \to 0} h^{1/2} \sum_{\nu} \frac{V(G, h_\nu) \lambda \gamma_\nu}{(1 + \lambda \gamma_\nu)^2} h_\nu(z_0), \\ \beta_{z_0} &= -\lim_{h \to 0} h^{1/2} \sum_{\nu} \frac{V(G, h_\nu)}{1 + \lambda \gamma_\nu} h_\nu(z_0). \end{aligned} \tag{3.4}$$

Theorem 3.1 is not directly useful since $\widehat{f}_{n,\lambda}$ centers on a biased "target parameter" $f_0^* \equiv f_0 - P_\lambda f_0$. In fact, even $\widehat{\theta}_{n,\lambda}$ is inconsistent. However, this observation does not conflict with the semiparametric result obtained by Mammen and van de Geer [20] that $\widehat{\theta}_{n,\lambda}$ is semiparametric efficient, since their assumption, i.e., $G_k \in H^m(\mathbb{I})$, is much stronger than ours, i.e., $G_k \in L_2(P_Z)$. Basically, they assume some smoothness on the underlying least favorable curve; see the discussion in [22, 11]. This leads to a natural question: is one able to remove the undesirable estimation bias of the parametric component by strengthening the condition on G_k? Our lemma below gives an affirmative answer.

LEMMA 3.2. *Suppose that there exists $b > 1/(2m)$ such that G_k satisfies*

$$\sum_{\nu} |V(G_k, h_\nu)|^2 \gamma_\nu^b < \infty \quad \text{for any } k = 1, \ldots, p. \tag{3.5}$$

Then we have, for any $z_0 \in \mathbb{I}$, $h^{1/2} A(z_0) = o(1)$, $h^{1/2}(W_\lambda A)(z_0) = o(1)$. Furthermore, if $n^{1/2} h^{m(1+b)} = o(1)$, then as $n \to \infty$,

$$\begin{pmatrix} \sqrt{n}(\theta_0^* - \theta_0) \\ \sqrt{nh}(g_0^*(z_0) - g_0(z_0) + (W_\lambda g_0)(z_0)) \end{pmatrix} \xrightarrow{d} 0. \tag{3.6}$$

An immediate implication of Lemma 3.2 is that we can completely remove the asymptotic estimation bias for θ (but only partly for g). When $b = 0$, Condition (3.5) reduces to Assumption A3 that $G_k \in L_2(P_Z)$. However, we require $b > 1/(2m)$ such that the Fourier coefficients $V(G_k, h_\nu)$ in (3.5) converge to zero at a faster rate than ν^{-mb} because $\gamma_\nu \asymp \nu^{2m}$; see Assumption A2. It is

well known that a faster decaying rate of the Fourier coefficients $V(G_k, h_\nu)$ implies a smoother G_k; see [12]. Therefore, Condition (3.5) amounts to requiring more smoothness of G_k. In fact, (3.5) is equivalent to $G_k \in H^{mb}(\mathbb{I})$ with $b > 1/(2m)$. Hence, the condition $G_k \in H^m(\mathbb{I})$ assumed in the classical semiparametric work by Mammen and van de Geer [20] may actually be weakened.

A more surprising fact revealed by Theorem 3.3 is that $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$ become asymptotically independent while $\widehat{\theta}_{n,\lambda}$ achieves the semiparametric efficiency bound (after the parametric estimation bias is removed). We call this discovery the *joint asymptotic phenomenon.*

THEOREM 3.3. *(Joint Asymptotic Theory II) Suppose that the conditions of Theorem 3.1 and Lemma 3.2 hold. Then we have, for any $z_0 \in \mathbb{I}$,*

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}\,\{\widehat{g}_{n,\lambda}(z_0) - g_0(z_0) + (W_\lambda g_0)(z_0)\} \end{pmatrix} \xrightarrow{d} N\left(0, \Psi\right), \tag{3.7}$$

where

$$\Psi = \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & \sigma_{z_0}^2 \end{pmatrix}. \tag{3.8}$$

In particular, h can be chosen as h^* in (3.7). Furthermore, if

$$\lim_{n\to\infty} (nh)^{1/2}(W_\lambda g_0)(z_0) = -b_{z_0}, \tag{3.9}$$

then we have

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ b_{z_0} \end{pmatrix}, \Psi\right). \tag{3.10}$$

We next make some remarks on Theorem 3.3. First, the asymptotic independence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$ greatly facilitates the construction of the joint CI for $(\theta_0, g_0(z_0))$ by directly building on the marginal CIs. Second, based on Theorem 3.3 and the Delta method, we can easily establish the prediction interval for a new response Y_{new} given future data $u_0 = (x_0, z_0)$ and the CI for some real-valued smooth function of $(\theta_0, g_0(z_0))$; see Section 5. Third, the nonparametric estimation bias, i.e., b_{z_0}, can be further removed through undersmoothing; see [27]. Lastly, the asymptotic variance for $\widehat{\theta}_{n,\lambda}$ remains the same after the above bias removal procedure, while that for $\widehat{g}_{n,\lambda}(z_0)$ has a significant change; see Ψ and Ψ^*.

In Remarks 3.1 and 3.2 below, we compare the marginal limit distributions implied by Theorem 3.3 with those derived in the semiparametric ([20]) and nonparametric ([27]) literature.

REMARK 3.1. *Our parametric limit distribution is $\sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \xrightarrow{d} N(0, \Omega^{-1})$, where $\Omega = E\{I(U)(X - G(Z))(X - G(Z))^T\}$. We find that it is exactly the same as that obtained in [20].*

However, the profile approach of [20] treats g as a nuisance parameter, and thus it cannot be adapted to obtain our joint limiting distribution. Some efficiency calculations further reveal that $\widehat{\theta}_{n,\lambda}$ is essentially semiparametric efficient if the conditional distribution of Y belongs to an exponential family. For example, in the partially linear models under Gaussian errors, Ω reduces to the well-known semiparametric efficiency bound $E(X - E(X|Z))^{\otimes 2}$; see [17]. □

REMARK 3.2. *Our (pointwise) nonparametric limit distribution, i.e., $\sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \xrightarrow{d} N(b_{z_0}, \sigma^2_{z_0})$, is in general different from that obtained in the nonparametric smoothing spline setup (without θ) in terms of different values of b_{z_0} and $\sigma^2_{z_0}$; see [27]. This is mainly due to the eigensystem difference in the two setups; see Remark 5.1 for more illustrations. An exception is the L_2 regression in which the two eigensystems happen to coincide. Our general finding corrects a common intuition in the literature that the nonparametric limit distribution is not affected by the involvement of a parametric component that is estimated at a faster convergence rate.* □

To further illustrate Theorem 3.3, we consider partial smoothing spline models where b_{z_0} and Ψ have explicit expressions. In addition, we explicitly specify a smoothing parameter condition under which both the parametric and nonparametric estimation biases asymptotically vanish.

COROLLARY 3.4. *(Joint Asymptotics for Partial Smoothing Spline Models) Let $m > 1 + \sqrt{3}/2 \approx 1.866$ and $\ell(y; a) = -(y-a)^2/2$. Suppose that Assumptions A2, A4, and (3.1) hold, and also that (3.5) holds for $b > 1 + 1/(2m)$ and $E(X - E(X|Z))^{\otimes 2}$ is positive definite. Furthermore, we assume that $g_0 \in H^{2m}(\mathbb{I})$ and satisfies $\sum_\nu |V(g_0^{(2m)}, h_\nu)h_\nu(z_0)| < \infty$.*

(i) Suppose g_0 satisfies the boundary condition

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

If $h/n^{-1/(4m+1)} \to c > 0$, then we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ b_{z_0} \end{pmatrix}, \Psi \right). \tag{3.12}$$

If $h \asymp n^{-d}$ for some $\frac{1}{4m+1} < d \le \frac{2m}{10m-1}$, then we have, for any $z_0 \in [0,1]$,

$$\begin{pmatrix} \sqrt{n}(\widehat{\theta}_{n,\lambda} - \theta_0) \\ \sqrt{nh}(\widehat{g}_{n,\lambda}(z_0) - g_0(z_0)) \end{pmatrix} \xrightarrow{d} N\left(0, \Psi\right). \tag{3.13}$$

In the above, $b_{z_0} = (-1)^{m-1} c^{2m} g_0^{(2m)}(z_0)/\pi(z_0)$ and Ψ takes the following explicit form:

$$\Psi = \begin{pmatrix} \{E[X - E(X|Z)]^{\otimes 2}\}^{-1} & 0 \\ 0 & \frac{\int_0^\infty (1+x^{2m})^{-2} dx}{\pi} \end{pmatrix}.$$

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

$$\frac{\partial^j}{\partial z^j} K_{z_0}(z)\Big|_{z=0} = o(1), \quad \frac{\partial^j}{\partial z^j} K_{z_0}(z)\Big|_{z=1} = o(1), \text{ for } j = 0, \ldots, m-1, \tag{3.14}$$

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

Corollary 3.4, i.e., (3.13), describes how to remove the nonparametric estimation bias through undersmoothing, although the corresponding smoothing parameter yields suboptimal nonparametric estimators in terms of their convergence rate. The reproducing kernel condition (3.14) can be implied by the so-called "exponential envelope condition" introduced in [23].

4. Joint Hypothesis Testing. In this section, we consider likelihood ratio testing for a set of joint local hypotheses in the general form (4.1). Under very general conditions, the null limit distribution is proved to be a mixture of a Chi-square distribution with p degrees of freedom and a scaled non-central Chi-square distribution with one degree of freedom. Obviously, these two Chi-square distributions are contributed by the parametric and nonparametric components, respectively. Given mild undersmoothing conditions, we further give more explicit null limit distributions for three commonly used joint hypotheses. We thus reveal a semi-nonparametric version of the Wilks phenomenon arising from the proposed joint local testing. In addition, as the smoothness order m approaches infinity, the scaling constant is found to converge to one eventually. A key technical tool in this section is a *restricted* version of JBR.

Consider the following joint hypothesis:

$$H_0 : M\theta + Qg(z_0) = \alpha \text{ vs. } H_1 : M\theta + Qg(z_0) \neq \alpha, \tag{4.1}$$

where $M = (M_1^T, \ldots, M_k^T)^T$ is a $k \times p$ matrix with $k \leq p+1$, $Q = (q_1, \ldots, q_k)^T$, and the α are k-vectors. Without loss of generality, we assume $N \equiv (M, Q)$ to have elements in $\mathbb{I} = [0,1]$. We further assume that the matrix N has full rank. M, Q, and α are all prespecified according to the testing needs. For example, when N is the identity matrix I_{p+1} and $\alpha = (\theta_0^T, w_0)^T$, H_0 reduces to $(\theta^T, g(z_0))^T = (\theta_0^T, w_0)^T$. See Corollary 4.6 for more examples. This provides another way to construct the joint CIs for $(\theta_0^T, g_0(z_0))^T$ without estimating Ω^{-1} or σ_{z_0}. The simultaneous testing of two marginal hypotheses, i.e., $H_0^P : \theta = \theta_0$ and $H_0^N : g(z_0) = w_0$, can also be used for this purpose, but it requires the very conservative Bonferroni correction. Moreover, our joint hypothesis is more general, and the testing approach is more straightforward to implement.

To define the likelihood ratio statistic, we establish the constrained estimate under (4.1) in three steps: (i) Arbitrarily choose $(\theta^\dagger, w^\dagger) \in \mathbb{R}^p \times \mathbb{R}$ satisfying $M\theta^\dagger + Qw^\dagger = \alpha$; (ii) Define

$\widehat{f}^0_{n,\lambda} \equiv (\widehat{\theta}^0_{n,\lambda}, \widehat{g}^0_{n,\lambda}) = \arg\max_{f\in\mathcal{H}_0} L_{n,\lambda}(f)$, where $\mathcal{H}_0 \equiv \{(\theta, g) \in \mathcal{H} | M\theta + Qg(z_0) = 0\}$ and

$$L_{n,\lambda}(f) = n^{-1}\sum_{i=1}^{n} \ell(Y_i; X_i^T\theta + g(Z_i) + X_i^T\theta^\dagger + w^\dagger) - (1/2)\lambda J(g, g); \tag{4.2}$$

(iii) Define the constrained estimate as $\widehat{f}^{H_0}_{n,\lambda} = (\widehat{\theta}^0_{n,\lambda} + \theta^\dagger, \widehat{g}^0_{n,\lambda} + w^\dagger)$. Then, the LRT statistic is $LRT_{n,\lambda} = \ell_{n,\lambda}(\widehat{f}^{H_0}_{n,\lambda}) - \ell_{n,\lambda}(\widehat{f}_{n,\lambda})$.

Given the inner product $\langle\cdot,\cdot\rangle$, we note that $\mathcal{H}_0$ is a closed subset in $\mathcal{H}$ and thus a Hilbert space. Hence, we will construct the projections of the two operators R_u and P_λ (associated with $\mathcal{H}$) onto the subspace $\mathcal{H}_0$, denoting them R^0_u and P^0_λ, respectively. Lemma 4.1 below provides a preliminary step for the construction. Its proof is similar to that of Proposition 2.1 and is thus omitted.

LEMMA 4.1. *For any $u = (x, z) \in \mathcal{U}$ and $q \in \mathbb{I}$, define*

$$H_{q,u} = (\Omega + \Sigma_\lambda)^{-1}(x - qA(z)) \text{ and } T_{q,u} = qK_z - A^T H_{q,u}.$$

Let $R_{q,u} \equiv (H_{q,u}, T_{q,u}) \in \mathcal{H}$. Then, for any $f \in \mathcal{H}$ and $u \in \mathcal{U}$, we have $\langle R_{q,u}, f\rangle = x^T\theta + qg(z)$.

Obviously, $R_{q,u}$ is a generalization of R_u defined in Proposition 2.1, i.e., $R_u = R_{1,u}$. Lemma 4.1 implies that the restricted parameter space $\mathcal{H}_0$ can be rewritten as

$$\mathcal{H}_0 = \{f = (\theta, g) \in \mathcal{H} | \langle R_{q_j,W_j}, f\rangle = 0,\ j = 1, \ldots, k\}, \tag{4.3}$$

where $W_j = (M_j, z_0)$. Define $H(Q, W) = (H_{q_1,W_1}, \ldots, H_{q_k,W_k})$, $T(Q, W) = (T_{q_1,W_1}(z_0), \ldots, T_{q_k,W_k}(z_0))$, and $M_K = MH(Q, W) + QT(Q, W)$. Construct the projections

$$R^0_u = R_u - \sum_{j=1}^{k} \rho_{u,j} R_{q_j,W_j} \text{ and } P^0_\lambda f = P_\lambda f - \sum_{j=1}^{k} \zeta_j(f) R_{q_j,W_j},$$

where $(\rho_{u,1} \ldots, \rho_{u,k})^T = M_K^{-1}(MH_u + QT_u(z_0))$ and $(\zeta_1(f), \ldots, \zeta_k(f))^T = M_K^{-1}(MH^*_g + QT^*_g(z_0))$. Recall that $R_u : u \mapsto (H_u, T_u)$ and $P_\lambda : (\theta, g) \mapsto (H^*_g, T^*_g)$ in Proposition 2.1. The invertibility of M_K is given in the proof of Proposition 4.2 below.

Proposition 4.2 below says that R^0_u and P^0_λ defined above are indeed what we need.

PROPOSITION 4.2. *Let $f = (\theta, g)$ and $\tilde{f} = (\tilde{\theta}, \tilde{g})$. For any $u = (x, z) \in \mathbb{I}^p \times \mathbb{I}$, $f, \tilde{f} \in \mathcal{H}_0$, we have $\langle R^0_u, f\rangle = x^T\theta + g(z)$ and $\langle P^0_\lambda f, \tilde{f}\rangle = \lambda J(g, \tilde{g})$.*

Based on Proposition 4.2, we can write down the Fréchet derivatives of $L_{n,\lambda}$ defined in (4.2) under $\mathcal{H}_0$ by modifying those of $\ell_{n,\lambda}$ as follows: replace θ, g, R_U, and P_λ by $\theta + \theta^\dagger$, $g + w^\dagger$, R^0_U,

and P_λ^0. For example,

$$\begin{aligned} DL_{n,\lambda}(f)\Delta f &= \frac{1}{n}\sum_{i=1}^{n}\dot{\ell}_a(Y_i; X_i^T\theta + g(Z_i) + X_i^T\theta^\dagger + w^\dagger)\langle R_{U_i}^0, \Delta f\rangle - \langle P_\lambda^0 f, \Delta f\rangle \\ &\equiv \langle S_n^0(f), \Delta f\rangle - \langle P_\lambda^0 f, \Delta f\rangle = \langle S_{n,\lambda}^0(f), \Delta f\rangle. \end{aligned}$$

Similarly, we have $S_{n,\lambda}^0(\widehat{f}_{n,\lambda}^0) = 0$. Also define $S^0(f) = E\{S_n^0(f)\}$ and $S_\lambda^0(f) = S^0(f) - P_\lambda^0(f)$. For the second derivative, we have $DS_{n,\lambda}^0(f)\Delta f_1\Delta f_2 = D^2L_{n,\lambda}(f)\Delta f_1\Delta f_2$ and $DS_\lambda^0(f)\Delta f_1\Delta f_2 = DS^0(f)\Delta f_1\Delta f_2 - \langle P_\lambda^0\Delta f_1, \Delta f_2\rangle$, where

$$DS^0(f)\Delta f_1\Delta f_2 = E\{\ddot{\ell}_a(Y; X^T\theta + g(Z) + X^T\theta^\dagger + w^\dagger)\langle R_U^0, \Delta f_1\rangle\langle R_U^0, \Delta f_2\rangle\}.$$

In Theorem 4.3 below, we present a new version of JBR that is restricted to the subspace $\mathcal{H}_0$. We need an additional assumption, A5, here. Let $f_0^0 \equiv (\theta_0 - \theta^\dagger, g_0 - w^\dagger)$, which belongs to $\mathcal{H}_0$ under H_0. Assumption A5 holds under mild conditions, i.e., (2.3), (2.4), and $\|\widehat{f}_{n,\lambda}^0 - f_0^0\|_{\mathcal{H}} = o_P(1)$.

ASSUMPTION A5. *Under H_0 specified in (4.1), $\|\widehat{f}_{n,\lambda}^0 - f_0^0\| = O_P((nh)^{-1/2} + h^m)$.*

THEOREM 4.3. *(Restricted Joint Bahadur Representation) Suppose that Assumptions A1, A2, A3, and A5 hold, and that $h = o(1)$ and $nh^2 \to \infty$ as $n \to \infty$. Under H_0 specified in (4.1), we have $\|\widehat{f}_{n,\lambda}^0 - f_0^0 - S_{n,\lambda}^0(f_0^0)\| = O_P(a_n \log n)$, where a_n is defined as in (2.17).*

Given the above preparatory results, we are ready to present general results for the null limit distribution of $-2n \cdot LRT_{n,\lambda}$ in Theorem 4.4. Define $r_n = (nh)^{-1/2} + h^m$, and let

$$\Phi_\lambda = \Lambda N^T M_K^{-1} N \Lambda^T,$$

where

$$\Lambda = \begin{pmatrix} (\Omega + \Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0, z_0)^{1/2} \end{pmatrix} \begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix}.$$

THEOREM 4.4. *(Joint Local Testing) Suppose that Assumptions A1 through A5 are satisfied. Also assume that the conditions in Lemma 3.2 hold, $h = o(1)$, $nh^2 \to \infty$, $r_n^2 h^{-1/2} = o(a_n)$, and $a_n = o(\min\{r_n, n^{-1}r_n^{-1}(\log n)^{-1}, n^{-1/2}h^{1/2}(\log n)^{-1}\})$, where a_n is defined as in (2.17). Furthermore, assume that, for any $z_0 \in [0,1]$, $\lim_{\lambda\to 0}\sqrt{n}(W_\lambda g_0)(z_0)/\sqrt{K(z_0,z_0)} = c_{z_0}$, $\lim_{\lambda\to 0}\Phi_\lambda = \Phi_0$, where Φ_0 is a fixed $(p+1)\times(p+1)$ positive semidefinite matrix, and*

$$\lim_{h\to 0} hV(K_{z_0}, K_{z_0}) \to \sigma_{z_0}^2 > 0, \ \lim_{h\to 0} h^{1/2}(W_\lambda A)(z_0) \to 0 \in \mathbb{R}^p, \tag{4.4}$$

$$\lim_{\lambda\to 0} E_Z\{B(Z)|K_{z_0}(Z)|^2\}/K(z_0,z_0) \equiv c_0 \in (0,1]. \tag{4.5}$$

Under H_0 specified in (4.1), we obtain: (i) $\|\widehat{f}_{n,\lambda} - \widehat{f}_{n,\lambda}^{H_0}\| = O_P(n^{-1/2})$; (ii) $-2n \cdot LRT_{n,\lambda} = n\|\widehat{f}_{n,\lambda} - \widehat{f}_{n,\lambda}^{H_0}\|^2 + o_P(1)$;

$$(iii) - 2n \cdot LRT_{n,\lambda} \xrightarrow{d} \upsilon^T \Phi_0 \upsilon, \tag{4.6}$$

where $\upsilon \sim N\left(\begin{pmatrix} 0 \\ c_{z_0} \end{pmatrix}, \begin{pmatrix} I_p & 0 \\ 0 & c_0 \end{pmatrix}\right)$.

The *parametric* convergence-rate result proved in (i) of Theorem 4.4 is reasonable since the null hypothesis imposes only a finite-dimensional constraint. By (2.9), it can be explicitly shown that

$$c_0 = \lim_{\lambda \to 0} \frac{Q_2(\lambda, z_0)}{Q_1(\lambda, z_0)}, \quad \text{where } Q_l(\lambda, z) \equiv \sum_{\nu \in \mathbb{N}} \frac{|h_\nu(z)|^2}{(1 + \lambda\gamma_\nu)^l} \text{ for } l = 1, 2. \tag{4.7}$$

It is well known that the reproducing kernel K is uniquely determined for any Hilbert space if it exists; see [13]. This implies that c_0 defined in (4.5) is also uniquely determined. Therefore, different choices of (h_ν, γ_ν) in (4.7) will give exactly the same value of c_0 although a particular choice may facilitate the computation of c_0. For example, in Case (I) of Example 5.1, we can explicitly calculate c_0 as 0.75 (0.83) when $m = 2$ (3) by choosing the trigonometric basis (5.2).

The null limit distribution derived in Theorem 4.4 cannot be directly used for inference because of the nontrivial estimation of c_{z_0}. Hence, in Corollary 4.5, we present a set of undersmoothing conditions under which the estimation bias of $\widehat{g}_{n,\lambda}$ can be removed, and thus $c_{z_0} = 0$.

COROLLARY 4.5. *Suppose that Assumptions A1 through A5 are satisfied and hypothesis H_0 holds. Let $m > 1 + \sqrt{3}/2 \approx 1.866$ and $G_1, \ldots, G_p$ satisfy (3.5) with $b > 1/(2m)$. 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 is implied by $g_0 \in H^{md}(\mathbb{I})$. Furthermore, if Φ_λ converges to some fixed $(p+1) \times (p+1)$ positive semidefinite matrix, i.e., Φ_0, and (4.4) and (4.5) are both satisfied for any $z_0 \in [0, 1]$, then (4.6) holds with $c_{z_0} = 0$ given that $h = h^* \asymp n^{-1/(2m+1)}$.*

Combining Theorem 4.4 with Corollary 4.5, we immediately obtain Corollary 4.6, which gives null limit distributions of the three commonly assumed joint hypotheses.

COROLLARY 4.6. *Suppose that the conditions in Corollary 4.5 hold. We have:*

(I) $H_0 : \theta = \theta_0$ and $g(z_0) = w_0$:

$$-2n \cdot LRT_{n,\lambda} \xrightarrow{d} \chi_p^2 + c_0 \chi_1^2,$$

where the two Chi-square distributions are independent. In this case, $N = I_{p+1}$, $\alpha = (\theta_0^T, w_0)^T$, and $\Phi_\lambda = \Phi_0 = I_{p+1}$.

(II) $H_0 : D\theta = \theta_0'$ *and* $g(z_0) = w_0$ *(*D *is an* $r \times p$ *matrix with* $0 < r \leq p$ *and rank*$(D) = r$*,* θ_0' *is an* r*-vector with* $0 < r < p$*):*

$$-2n \cdot LRT_{n,\lambda} \xrightarrow{d} \chi_r^2 + c_0\chi_1^2,$$

where the two Chi-square distributions are independent. In this case, $N = \begin{pmatrix} D & 0_r \\ 0_p^T & 1 \end{pmatrix}$, $\alpha = (\theta_0'^T, w_0)^T$, *and* $\Phi_0 = \begin{pmatrix} \mathcal{P}_r & 0_p \\ 0_p^T & 1 \end{pmatrix}$ *with the projection matrix (of rank* r*)* $\mathcal{P}_r = \Omega^{-1/2}D^T(D\Omega^{-1}D^T)^{-1}D\Omega^{-1/2}$.

(III) $H_0 : x_0^T\theta + g(z_0) = \alpha$ *(*α*,* x_0*, and* z_0 *are given):*

$$-2n \cdot LRT_{n,\lambda} \xrightarrow{d} c_0\chi_1^2.$$

In this case, $N = (x_0^T, 1)$ *and* $\Phi_0 = \begin{pmatrix} 0_{p\times p} & 0_p \\ 0_p^T & 1 \end{pmatrix}$.

The independence between the two Chi-square distributions in (I) and (II) follows from the asymptotic independence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$ proved in Theorem 3.3. In comparison with (I) and (II), we note that the null limit distribution in (III) is dominated by the effect from $g(z_0)$ because of its nonparametric nature, i.e., slower convergence rate.

As far as we are aware, Corollary 4.6 is the first semi-nonparametric version of the Wilks phenomenon. Note that the value of c_0 converges to one as $m \to \infty$. Therefore, this new type of Wilks phenomenon reverts to the classical version in the parametric setup as $m \to \infty$ by further consideration of the independence of the two Chi-squares. For example, the null limit distribution in (I) of Corollary 4.6 becomes χ_{p+1}^2 as $m \to \infty$.

5. Examples. In this section, we present three concrete examples together with simulations. In all the examples, the G_ks are sufficiently smooth for Theorem 3.3 and Corollary 4.6 to apply.

EXAMPLE 5.1. (*Partial Smoothing Spline*) Consider a partially linear regression model

$$Y = X^T\theta + g(Z) + \epsilon, \tag{5.1}$$

where $\epsilon \sim N(0, \sigma^2)$ with an unknown σ^2. Hence, $B(Z) = \sigma^{-2}$. For simplicity, Z is assumed to be uniformly distributed over $\mathbb{I}$. In this case, $V(g, \widetilde{g})$ becomes the usual L^2-norm. The function *ssr()* in the R package *assist* was used to select the smoothing parameter λ based on CV or GCV; see [15]. The unknown error variance can be consistently estimated by $\widehat{\sigma}^2 = n^{-1}\sum_i (Y_i - X_i^T\widehat{\theta}_{n,\lambda} - \widehat{g}_{n,\lambda}(Z_i))^2/(n - \text{trace}(A(\lambda)))$, where $A(\lambda)$ denotes the smoothing matrix (see [29]).

We next consider two separate cases: (I) $g \in H_0^m(\mathbb{I})$ and (II) $g \in H^m(\mathbb{I})$.

Case (I) $g \in H_0^m(\mathbb{I})$*:* We choose the following trigonometric eigensystem for $H_0^m(\mathbb{I})$:

$$h_\mu(z) = \begin{cases} \sigma, & \mu = 0, \\ \sqrt{2}\sigma\cos(2\pi kz), & \mu = 2k, k = 1, 2, \ldots, \\ \sqrt{2}\sigma\sin(2\pi kz), & \mu = 2k-1, k = 1, 2, \ldots, \end{cases} \tag{5.2}$$

with the corresponding eigenvalues $\gamma_0 = 0$ and $\gamma_{2k-1} = \gamma_{2k} = \sigma^2(2\pi k)^{2m}$ for $k \geq 1$.

It follows from (3.4) and (5.2) that the asymptotic variance of $\widehat{g}_{n,\lambda}(z_0)$ is expressed as

$$\sigma_{z_0}^2 = \lim_{h\to 0}\left\{\sigma^2 h\left(1 + 2\sum_{k=1}^{\infty}\frac{1}{(1+(2\pi h\sigma^{1/m}k)^{2m})^2}\right)\right\}.$$

Lemma 6.1 in [27] leads to, for $l = 1, 2$,

$$\sum_{k=1}^{\infty}\frac{1}{(1+(2\pi h\sigma^{1/m}k)^{2m})^l} \sim \frac{I_l}{2\pi h\sigma^{1/m}}, \tag{5.3}$$

where $I_l = \int_0^\infty (1+x^{2m})^{-l}dx$. Therefore, we have $\sigma_{z_0}^2 = (I_2\sigma^{2-1/m})/\pi$. According to Corollary 3.4, the 95% prediction interval for Y at a new observed covariate $u_0 = (x_0, z_0)$ is

$$\widehat{Y} \pm 1.96\sqrt{\widehat{\sigma}^{2-1/m}I_2/(\pi nh) + \widehat{\sigma}^2}, \tag{5.4}$$

where $\widehat{Y} = x_0^T\widehat{\theta}_{n,\lambda} + \widehat{g}_{n,\lambda}(z_0)$ is the predicted response. We next calculate c_0 based on (4.7). It follows from (5.2) and (5.3) that

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

for $l = 1, 2$. Hence, we obtain

$$c_0 = I_2/I_1. \tag{5.5}$$

Further calculations reveal that $c_0 = 0.75$ (0.83) when $m = 2$ (3).

In the simulations, we first verify the joint asymptotic phenomenon, i.e., (3.13), by investigating the (asymptotic) independence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z_0)$. Let $\theta_0 = (8, -8)^T$ and $g_0(z) = 0.6\beta_{30,17}(z) + 0.4\beta_{3,11}(z)$, where $\beta_{a,b}$ is the density function for $Beta(a, b)$. We estimated the nonparametric function g_0, which has many peaks and troughs, using periodic splines with $m = 2$; σ was set to one. To allow the linear and nonlinear covariates (X, Z) to be dependent, we generated them as follows: generate $U, V, Z \overset{i.i.d.}{\sim} Unif[0, 1]$, and set $X_1 = (U + 0.2Z)/1.2$,

$X_2 = (V + 0.2Z)/1.2$. This leads to $corr(X_1, Z) = corr(X_2, Z) \approx 0.20$, where $corr$ denotes the correlation coefficient. The dependence between $\widehat{\theta}_{n,\lambda}$ and $\widehat{g}_{n,\lambda}(z)$ was evaluated through the absolute values of the sample correlation coefficients (ACC) between $\widehat{\theta}_{n,\lambda} = (\widehat{\theta}_{n,\lambda,1}, \widehat{\theta}_{n,\lambda,2})^T$ and $\widehat{g}_{n,\lambda}(z)$ at ten evenly spaced z points in $[0, 1]$ based on 500 replicated data sets. The results are summarized in Figure 1 for sample sizes $n = 100, 300, 1000$. As n increases, it is easy to see that the ACC curves become uniformly closer to zero, which strongly indicates the desired asymptotic independence.

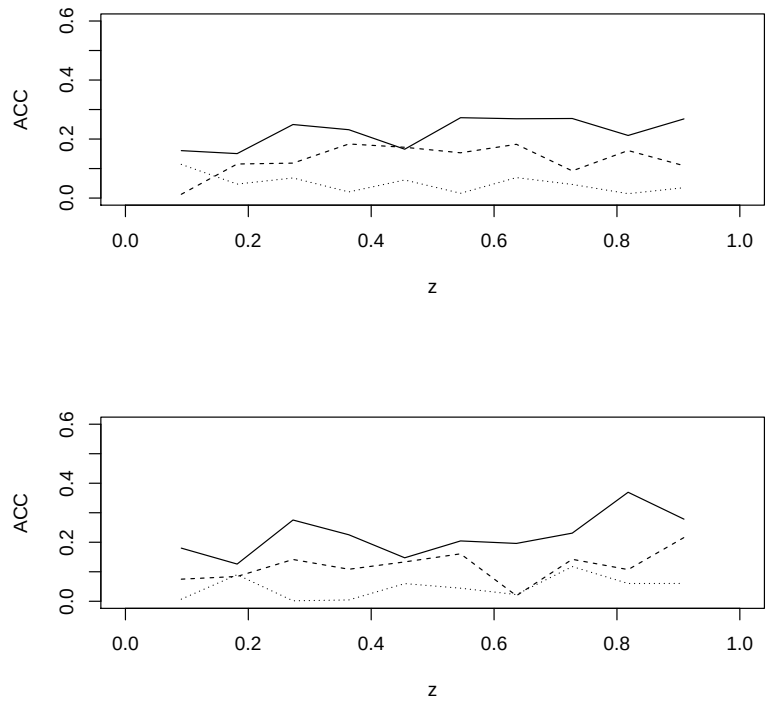

FIG 1. *Absolute values of correlation coefficients (ACC) between* $\widehat{\theta}_{n,\lambda,1}$ *and* $\widehat{g}_{n,\lambda}(z)$ *(the upper plot), and* $\widehat{\theta}_{n,\lambda,2}$ *and* $\widehat{g}_{n,\lambda}(z)$ *(the lower plot), at ten evenly spaced nonlinear covariates in Case (I) of Example 5.1. The three lines correspond to three sample sizes:* $n = 100$ *(solid),* $n = 300$ *(dashed),* $n = 1000$ *(dotted).*

To examine the performance of the 95% prediction intervals (5.4), we calculated the proportions of the prediction intervals covering the future response Y generated from model (5.1), i.e., the coverage proportion. The simulation setup is the same as before except that we assume a one-dimensional $\theta_0 = 4$ for simplicity. The new covariates are (x_0, z_0) with $x_0 = 1/4, 2/4, 3/4$ and z_0 being thirty evenly spaced points in $[0, 1]$. The coverage proportions were calculated based on 500 replications. We summarize our simulation results in Figure 2 for sample sizes $n = 100, 300, 1000$. As n grows, all the coverage proportions approach the nominal level, 95%. In addition, the predic-

tion interval lengths approach the theoretical value indicated in formula (5.4), i.e., $2 \times 1.96 = 3.92$.

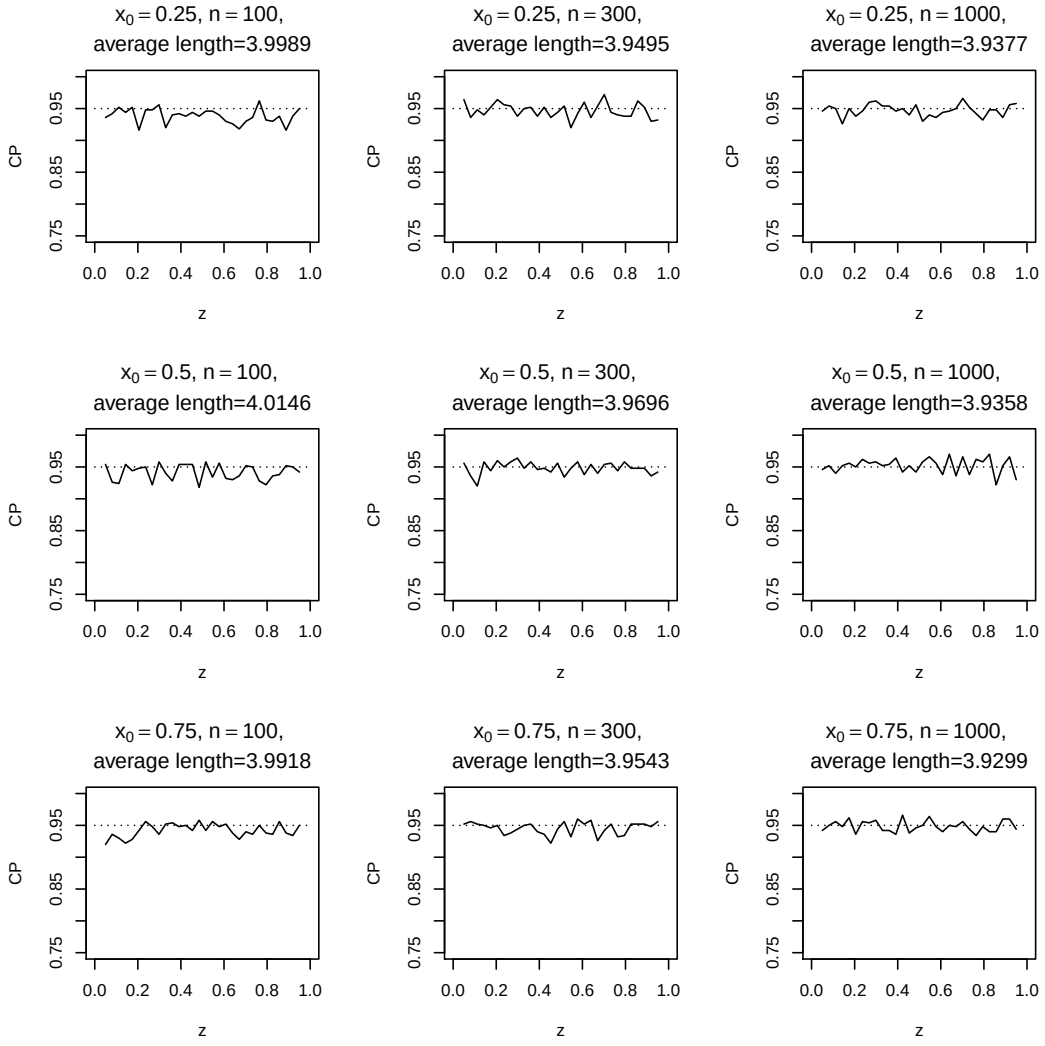

FIG 2. *Coverage proportion of 95% prediction intervals in Case (I) of Example 5.1.*

Finally, we tested $H_0 : x_0\theta + g(z_0) = 0$. The true parameters were chosen as $\theta_0 = -4$, $g_0(z) = \sin(\pi z)$, and $\sigma = 1$. The performance was demonstrated by calculating the powers for the nine combinations of $x_0 = 1/4, 2/4, 3/4$ and $z_0 = 1/4, 2/4, 3/4$ through 500 replicated data sets. In particular, H_0 is true when $x_0 = 1/4$ and $z_0 = 2/4$, and H_0 is false at the other values of (x_0, z_0). The results are summarized in Table 1 for sample sizes $n = 50, 100, 300, 500, 1000, 1500$. We observe that when $x_0 = 1/4$ and $z_0 = 2/4$, the power approaches the correct size 5%, while at the other values of (x_0, z_0), where H_0 does not hold, the power approaches one. This shows the validity of our local LRT test. The detailed computational algorithm for the constrained estimate under H_0 is given in the online supplementary document.

		$n=50$	$n=100$	$n=300$	$n=500$	$n=1000$	$n=1500$
$x_0=1/4$	$z_0=1/4$	43.00	56.60	77.60	90.40	97.80	98.60
	$z_0=2/4$	20.60	13.00	7.20	7.00	5.60	5.10
	$z_0=3/4$	42.00	50.00	77.60	89.60	97.80	99.20
$x_0=2/4$	$z_0=1/4$	98.60	99.80	100	100	100	100
	$z_0=2/4$	96.80	99.00	100	100	100	100
	$z_0=3/4$	98.80	99.80	100	100	100	100
$x_0=3/4$	$z_0=1/4$	99.80	100	100	100	100	100
	$z_0=2/4$	99.60	100	100	100	100	100
	$z_0=3/4$	99.60	100	100	100	100	100

TABLE 1

100 × Power of the local LRT test for nine combinations of x_0 and z_0 for Case (I) of Example 5.1.

Case (II) $g \in H^m(\mathbb{I})$*:* For this larger parameter space, we first construct an effective eigensystem that satisfies (2.10). Let $\tilde{h}_\nu$s and $\tilde{\gamma}_\nu$s be the normalized (with respect to the usual L_2-norm) eigenfunctions and eigenvalues of the boundary value problem $(-1)^m \tilde{h}_\nu^{(2m)} = \tilde{\gamma}_\nu \tilde{h}_\nu$, $\tilde{h}_\nu^{(j)}(0) = \tilde{h}_\nu^{(j)}(1) = 0$, $j = m, m+1, \ldots, 2m-1$. Then, we can construct $h_\nu = \sigma \tilde{h}_\nu$ and $\gamma_\nu = \sigma^2 \tilde{\gamma}_\nu$. Consequently,

$$Q_l(\lambda, z) = \sum_\nu \frac{|h_\nu(z)|^2}{(1+\lambda\gamma_\nu)^l} = \sigma^{2-1/m} h^{-1} \sum_\nu \frac{h\sigma^{1/m}|\tilde{h}_\nu(z)|^2}{(1+(h\sigma^{1/m})^{2m}\tilde{\gamma}_\nu)^l} \sim \sigma^{2-1/m} h^{-1} c_l(z), \tag{5.6}$$

where $c_l(z) = \lim_{h^\dagger \to 0} \sum_\nu \frac{h^\dagger |\tilde{h}_\nu(z)|^2}{(1+(h^\dagger)^{2m}\tilde{\gamma}_\nu)^l}$ and $h^\dagger = h\sigma^{1/m}$, for $l = 1, 2$. Hence, by (4.7), we have $c_0 = c_2(z_0)/c_1(z_0)$. In addition, by (3.4), we obtain the asymptotic variance of $\widehat{g}_{n,\lambda}(z_0)$ as $\sigma^{2-1/m} c_2(z_0)$, implying the following 95% prediction interval:

$$\widehat{Y} \pm 1.96 \sqrt{\widehat{\sigma}^{2-1/m} c_2(z_0)/(nh) + \widehat{\sigma}^2}.$$

The above discussion applies to general m. However, when $m = 2$, we can avoid estimating the $c_l(z_0)$s required in the inference by applying the equivalent kernel approach. Following the discussion in [27], we can actually obtain the same values of c_0 and $\sigma_{z_0}^2$ as in Case (I). The simulation setup is the same as before except that a different (nonperiodic) $g_0(z) = \sin(2.8\pi z)$ was used. Figure 3 displays the coverage proportion of the 95% prediction intervals for three sample sizes $n = 100, 300, 1000$. As n grows, all the coverage proportions approach the 95% nominal level, and the prediction interval lengths approach the theoretical value 3.92.

EXAMPLE 5.2. (*Semiparametric Gamma Model*) Consider a two-parameter exponential model

$$Y|X, Z \sim \text{Gamma}(\alpha, \exp(X^T\theta_0 + g_0(Z))),$$

where $\alpha > 0$ is known, $g_0 \in H_0^m(\mathbb{I})$, and $Z \sim Unif[0,1]$. It can be easily shown that $I(U) = \alpha$, and thus $B(Z) = \alpha$ in this model. Consequently, we can construct the basis functions h_ν as those defined in (5.2) with $\sigma = \alpha^{-1/2}$, and the eigenvalues as $\gamma_0 = 0$ and $\gamma_{2k-1} = \gamma_{2k} = \alpha^{-1}(2\pi k)^{2m}$ for $k \geq 1$. The remaining analysis is similar to Case (I) of Example 5.1, e.g., c_0 is given in (5.5).

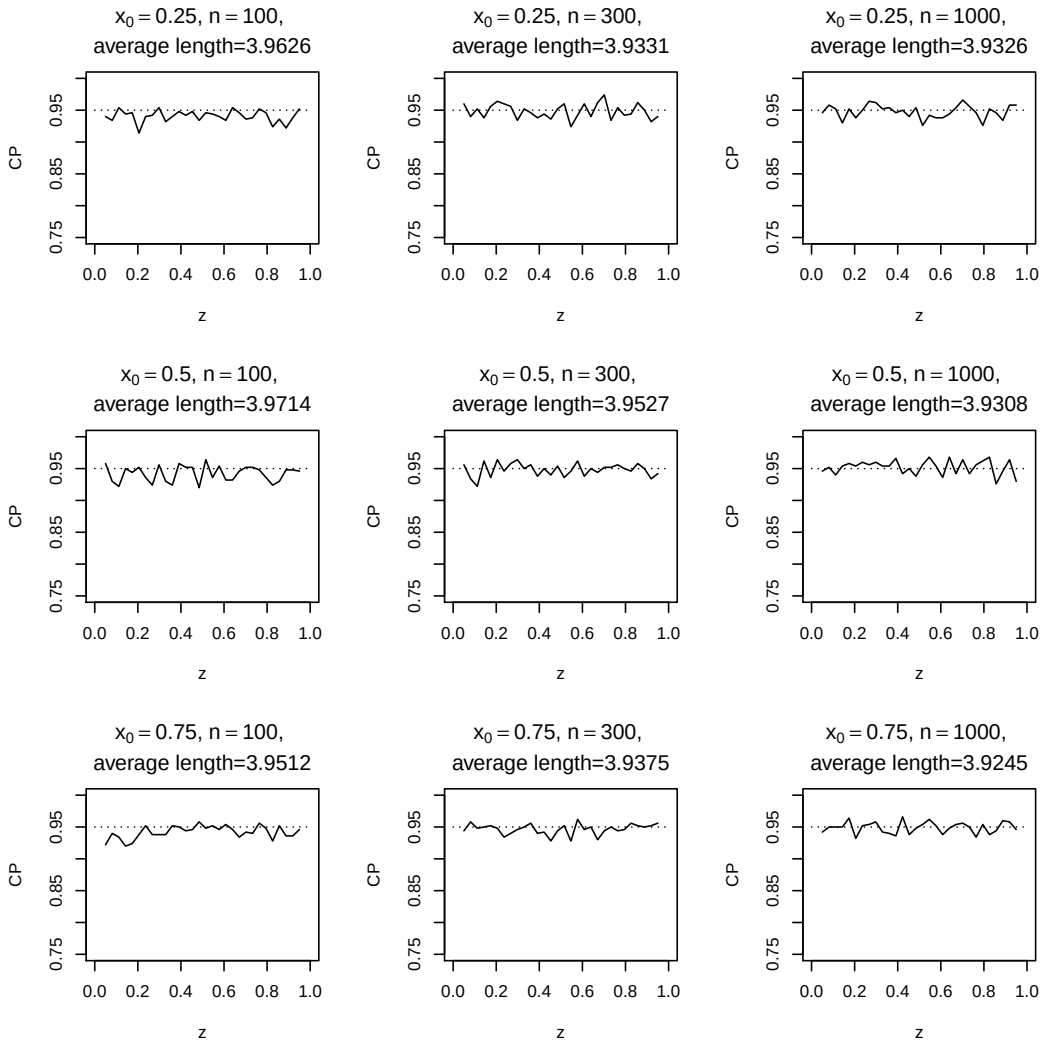

FIG 3. *Coverage proportion of 95% prediction intervals in Case (II) of Example 5.1.*

EXAMPLE 5.3. (*Semiparametric Logistic Regression*) For the binary response $Y \in \{0, 1\}$, we consider the following semiparametric logistic model

$$P(Y = 1|X = x, Z = z) = \frac{\exp(x^T\theta_0 + g_0(z))}{1 + \exp(x^T\theta_0 + g_0(z))}, \tag{5.7}$$

where $g_0 \in H^m(\mathbb{I})$. A straightforward calculation gives

$$I(U) = \frac{\exp(X^T\theta_0 + g_0(Z))}{(1 + \exp(X^T\theta_0 + g_0(Z)))^2} \text{ and } B(Z) = E\left\{ \frac{\exp(X^T\theta_0 + g_0(Z))}{(1 + \exp(X^T\theta_0 + g_0(Z)))^2} \middle| Z \right\}. \tag{5.8}$$

Similarly, we can solve (2.10) to construct the underlying eigensystem. However, in this model, we need to use consistent estimators of $B(\cdot)$ and $\pi(\cdot)$, e.g., $\widehat{B}(\cdot)$ is a plug-in estimator and $\widehat{\pi}(\cdot)$ is a kernel density estimator.

Given the length of this paper, we conducted simulations only for the CIs of the conditional mean defined in (5.7) at a number of (x_0, z_0) values, i.e., $x_0 = 1/4, 2/4, 3/4$ and thirty evenly spaced z_0 over $[0, 1]$. The true parameters are $\theta_0 = -0.5$ and $g_0(z) = 0.3(10^6)(1-z)^6 + (10^4)(1-z)^{10} - 2$. For simplicity, we generated $X, Z \overset{i.i.d.}{\sim} Unif[0, 1]$. Based on 500 replicated data sets, we constructed the 95% CIs and calculated their coverage proportions. The results are summarized in Figure 4 for various sample sizes $n = 400, 500, 700$. We observe that, as n increases, the coverage proportions approach the desired level, 95%, and the CI lengths shrink to zero.

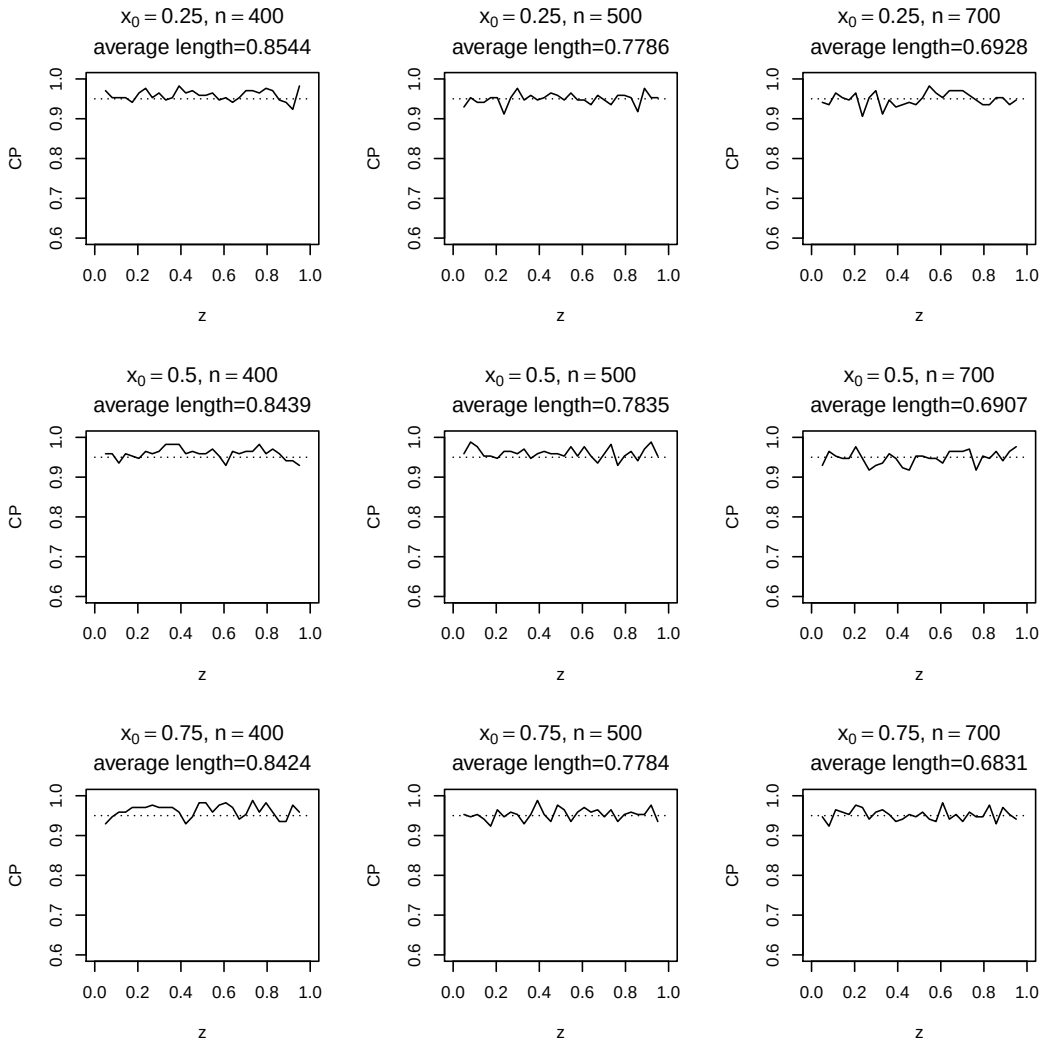

FIG 4. *Coverage proportion of 95% CIs for the conditional mean constructed at a variety of* (x, z) *values.*

REMARK 5.1. *We use this logistic regression model to illustrate the eigensystem difference between the semi-nonparametric context and the nonparametric context, which leads to different*

inference for the nonparametric components (except under some strong conditions, e.g., (5.9) below). This is slightly counterintuitive given that the parametric component can be estimated at a faster rate. As discussed above, the eigensystem for the semiparametric logistic model relies on $B(z)$ defined in (5.8). According to [27], the eigensystem for the nonparametric logistic model relies on $I'(z)$ defined as $\exp(g_0(z))/(1+\exp(g_0(z)))^2$. Therefore, the equivalence of the two eigensystems holds iff $B(z) = I'(z)$, i.e.,

$$E\left\{\frac{\exp(X^T\theta_0)}{(1+\exp(X^T\theta_0+g_0(z)))^2}\middle| Z=z\right\} = \frac{1}{(1+\exp(g_0(z)))^2}. \tag{5.9}$$

If $\theta_0 = 0$, it is clear that (5.9) is true. However, we argue that in general (5.9) may not hold. For instance, it does not hold when $g_0(z) = 0$ for some $z \in [0,1]$ because the above equation then simplifies to $E\left\{\frac{\exp(X^T\theta_0)}{(1+\exp(X^T\theta_0))^2}\right\} = 1$. This is not possible since $\frac{\exp(X^T\theta_0)}{(1+\exp(X^T\theta_0))^2} < 1$ almost surely. This concludes our argument. □

6. APPENDIX. For any $(\theta, g) \in \mathcal{H}$, define $f_{\theta,g} : (x,z) \mapsto x^T\theta + g(z)$, where $(x,z) \in \mathcal{S}$. Thus, (θ, g) can be viewed as a bivariate function defined on $\mathcal{U}$. Throughout the appendix, we will not distinguish (θ, g) and its associated function $f_{\theta,g}$. For instance, we use $(\theta, g) \in \mathcal{G}_0$ to mean $f_{\theta,g} \in \mathcal{G}_0$, some set of functions defined over $\mathcal{U}$.

A.1. *An Important Lemma.*

Lemma A.1.

$$\lim_{\lambda\to 0} E_Z\{B(Z)(G(Z)-A(Z))(G(Z)-A(Z))^T\} = 0. \tag{A.1}$$

$$\lim_{\lambda\to 0} E_Z\{B(Z)G(Z)(G(Z)-A(Z))^T\} = 0. \tag{A.2}$$

PROOF. The proofs of (A.1) and (A.2) are similar, so we only show (A.2) holds. Considering (2.11) and taking $g = h_\nu$, one has

$$V(G_k, h_\nu) = \langle A_k, h_\nu\rangle_1 = \langle \sum_\mu V(A_k, h_\mu)h_\mu, h_\nu\rangle_1 = (1+\lambda\gamma_\nu)V(A_k, h_\nu); \tag{A.3}$$

and, taking $g = K_z$, one has $V(G_k, K_z) = A_k(z)$. By (A.3), $A_k = \sum_\nu \frac{V(G_k,h_\nu)}{1+\lambda\gamma_\nu}h_\nu$ holds in $L_2(\mathbb{I})$. For any $k, j = 1, \ldots, p$, by a straightforward calculation, we have

$$E_Z\{B(Z)G_j(Z)(G_k(Z)-A_k(Z))\} = \sum_\nu V(G_j, h_\nu)V(G_k, h_\nu)\frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu}.$$

By square summability of $\{V(G_k, h_\nu)\}_{\nu\in\mathbb{N}}$ and dominated convergence theorem, the above sum converges to zero as $\lambda \to 0$. □

A.2. *Proof of Theorem 3.1.* Define

$$\widehat{f}^h_{n,\lambda} = (\widehat{\theta}_{n,\lambda}, h^{1/2}\widehat{g}_{n,\lambda}),\ f^{*h}_0 = (\theta^*_0, h^{1/2}g^*_0),\ R^h_u = (H_u, h^{1/2}T_u),$$

where recall $f^*_0 = (id - P_\lambda)f_0$, H_u, T_u were defined by (2.13) and P_λ is specified in Proposition 2.2. By Theorem 2.6,

$$Rem_n = \widehat{f}_{n,\lambda} - f^*_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i R_{U_i}$$

satisfies $\|Rem_n\| = O_P(a_n \log n)$, which will imply by Assumption A1 (b) that

$$\|\widehat{\theta}_{n,\lambda} - \theta^*_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i H_{U_i}\|_{l_2} = O_P(a_n \log n), \tag{A.4}$$

Define $Rem^h_n = \widehat{f}^h_{n,\lambda} - f^{*h}_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i R^h_{U_i}$, then it is easy to see that

$$Rem^h_n - h^{1/2}Rem_n = ((1-h^{1/2})(\widehat{\theta}_{n,\lambda} - \theta^*_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i H_{U_i}), 0).$$

Thus, by (A.4),

$$\|Rem^h_n - h^{1/2}Rem_n\| \le (1-h^{1/2}) \cdot O\left(\|\widehat{\theta}_{n,\lambda} - \theta^*_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i H_{U_i}\|_{l_2}\right) = O_P(a_n \log n).$$

Since by assumption $a_n \log n = o(n^{-1/2})$, $\|Rem^h_n\| = o_P(n^{-1/2})$. Next we will use Rem^h_n to obtain the target joint limiting distribution.

The idea is to employ Wald's device. For any $x \in \mathbb{I}^p$, we will obtain the limiting distribution of $n^{1/2}x^T(\widehat{\theta}_{n,\lambda} - \theta^*_0) + (nh)^{1/2}(\widehat{g}_{n,\lambda}(z_0) - g^*_0(z_0))$. Note that this is equal to $n^{1/2}\langle R_u, \widehat{f}^h_{n,\lambda} - f^{*h}_0\rangle$ with $u = (x, z_0)$. Using the fact that

$$\begin{aligned} & |n^{1/2}\langle R_u, \widehat{f}^h_{n,\lambda} - f^{*h}_0 - \frac{1}{n}\sum_{i=1}^n \epsilon_i R^h_{U_i}\rangle| \\ \le\ & n^{1/2}\|R_u\| \cdot \|Rem^h_n\| \\ =\ & O_P(n^{1/2}h^{-1/2}a_n \log n) = o_P(1), \end{aligned}$$

we just need to find the limiting distribution of $n^{1/2}\langle R_u, \frac{1}{n}\sum_{i=1}^n \epsilon_i R^h_{U_i}\rangle$, which is equal to

$$n^{1/2}\langle R_u, \frac{1}{n}\sum_{i=1}^n \epsilon_i R^h_{U_i}\rangle = n^{-1/2}\sum_{i=1}^n \epsilon_i (x^T H_{U_i} + h^{1/2}T_{U_i}(z_0)).$$

Next we will use CLT to find its limiting distribution. By Assumption A1 (c), i.e., $E\{\epsilon^2|U\} = I(U)$, we have that

$$\begin{aligned} & Var(\sum_{i=1}^n \epsilon_i (x^T H_{U_i} + h^{1/2}T_{U_i}(z_0))) \\ =\ & nE\{\epsilon^2 | x^T H_U + h^{1/2}T_U(z_0)|^2\} \\ =\ & nE\{E\{\epsilon^2|U\}|x^T H_U + h^{1/2}T_U(z_0)|^2\} \\ =\ & nE\{I(U)|x^T H_U + h^{1/2}T_U(z_0)|^2\}. \end{aligned}$$

A direct examination from (2.13) shows that

$$\begin{aligned} x^T H_U + h^{1/2} T_U(z) &= x^T(\Omega+\Sigma_\lambda)^{-1}(X-A(Z)) + h^{1/2}K_Z(z_0) - h^{1/2}A(z_0)^T(\Omega+\Sigma_\lambda)^{-1}(X-A(Z)) \\ &= h^{1/2}K_Z(z_0) + (x - h^{1/2}A(z_0))^T(\Omega+\Sigma_\lambda)^{-1}(X-A(Z)). \end{aligned}$$

Thus,

$$\begin{aligned} & E\{I(U)|x^T H_U + h^{1/2}T_U(z_0)|^2\} \\ = \; & hE\{I(U)|K_Z(z_0)|^2\} + 2h^{1/2}(x-h^{1/2}A(z_0))^T(\Omega+\Sigma_\lambda)^{-1}E\{I(U)K_Z(z_0)(X-A(Z))\} \\ & +(x-h^{1/2}A(z_0))^T E\{I(U)H_U H_U^T\}(x-h^{1/2}A(z_0)). \end{aligned} \tag{A.5}$$

Lemma A.1 tells us, as $\lambda \to 0$, $\Sigma_\lambda = E_Z\{B(Z)G(Z)(G(Z)-A(Z))^T\} \to 0$. It can be verified that

$$\begin{aligned} & E_U\{I(U)H_U H_U^T\} \\ = \; & (\Omega+\Sigma_\lambda)^{-1}E\{I(U)(X-A(Z))(X-A(Z))^T\}(\Omega+\Sigma_\lambda)^{-1} \\ = \; & (\Omega+\Sigma_\lambda)^{-1}E\{I(U)(X-G(Z)+G(Z)-A(Z))(X-G(Z)+G(Z)-A(Z))^T\}(\Omega+\Sigma_\lambda)^{-1} \\ = \; & (\Omega+\Sigma_\lambda)^{-1}\left(E\{I(U)(X-G(Z))(X-G(Z))^T\} + E\{I(U)(G(Z)-A(Z))(G(Z)-A(Z))^T\}\right) \\ & \cdot(\Omega+\Sigma_\lambda)^{-1} \to \Omega^{-1}, \end{aligned}$$

where the last limit follows by Lemma A.1. By assumption, as $\lambda \to 0$, $hE\{I(U)|K_Z(z_0)|^2\} = hV(K_{z_0},K_{z_0}) \to \sigma_{z_0}^2$, $h^{1/2}A(z_0) \to -\beta_{z_0}$, and

$$\begin{aligned} & h^{1/2}E\{I(U)K_Z(z_0)(X-A(Z))\} \\ = \; & h^{1/2}E\{B(Z)K_{z_0}(Z)(G(Z)-A(Z))\} \\ = \; & h^{1/2}(V(G,K_{z_0}) - V(A,K_{z_0})) \\ = \; & h^{1/2}(A(z_0) - V(A,K_{z_0})) \\ = \; & h^{1/2}(W_\lambda A)(z_0) \to \alpha_{z_0}. \end{aligned}$$

Thus, as λ approaches zero, the limit of (A.5) is

$$\sigma_{z_0}^2 + 2(x+\beta_{z_0})^T\Omega^{-1}\alpha_{z_0} + (x+\beta_{z_0})^T\Omega^{-1}(x+\beta_{z_0}) = (x^T,1)\Psi^*(x^T,1)^T,$$

where Ψ^* is defined in (3.3). This completes the proof.

A.3. *Proof of Lemma 3.2.* We have three steps to show (3.6).

(i). Show $\|V(G,W_\lambda g_0)\|_{l_2} = o(n^{-1/2})$. By (2.9), $V(G_k,W_\lambda g_0) = \sum_{\mu\in\mathbb{Z}} V(G_k,h_\mu)V(g_0,h_\mu)\frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu}$,

for any $k = 1, \ldots, p$. Then by Cauchy's inequality, we have

$$\begin{aligned} & |V(G_k, W_\lambda g_0)|^2 \\ \le & \sum_\mu |V(G_k, h_\mu)|^2 \frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu} \sum_\mu |V(g_0, h_\mu)|^2 \frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu} \\ \le & \text{const} \cdot \lambda \sum_\mu |V(G_k, h_\mu)|^2 \frac{\lambda\gamma_\mu}{1+\lambda\gamma_\mu} \\ = & \text{const} \cdot \lambda \sum_\mu |V(G_k, h_\mu)|^2 \gamma_\mu^b \left(\frac{\lambda\gamma_\mu^{1-b}}{1+\lambda\gamma_\mu} \right) \\ \le & \text{const} \cdot \lambda^{1+b}. \end{aligned}$$

Thus, when $n^{1/2}\lambda^{(1+b)/2} = n^{1/2}h^{m(1+b)} = o(1)$, $\|V(G, W_\lambda g_0)\|_{l_2} = o(n^{-1/2})$.

(ii). Show $\|A_k\|_{\sup} = O(1)$, for any $k = 1, \ldots, p$. Note for any $z \in \mathbb{I}$, by (2.11),

$$A_k(z) = \langle A_k, K_z \rangle_1 = V(G_k, K_z) = \sum_{\mu \in \mathbb{N}} \frac{V(G_k, h_\mu)}{1+\lambda\gamma_\mu} h_\mu(z).$$

By boundedness of h_νs (Assumption A3) and by Cauchy's inequality, uniformly for $z \in \mathbb{I}$,

$$\begin{aligned} |A_k(z)|^2 & \le \sum_\mu |V(G_k, h_\mu)|^2 (1+\gamma_\mu)^b |h_\mu(z)|^2 \cdot \sum_\mu \frac{1}{(1+\gamma_\mu)^b (1+\lambda\gamma_\mu)^2} \\ & = O\left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b} \right) = O(1), \end{aligned}$$

where the last equality follows by $\gamma_\mu \asymp \mu^{2m}$ and $2mb > 1$. This shows $\|A_k\|_{\sup} = O(1)$, implying $h^{1/2}A(z_0) = o(1)$. By (2.12), $(W_\lambda A)(z) = A(z) - \sum_\mu \frac{V(G,h_\mu)}{(1+\lambda\gamma_\mu)^2} h_\mu(z)$. Using the above derivations we can show that uniformly for $z \in \mathbb{I}$, $|\sum_\mu \frac{V(G,h_\mu)}{(1+\lambda\gamma_\mu)^2} h_\mu(z)|^2 = O\left(\sum_\mu \frac{1}{(1+\gamma_\mu)^b}\right) = O(1)$, implying $h^{1/2}(W_\lambda A)(z_0) = o(1)$.

(iii). By (i) and (ii), (3.6) follows by, as $n \to \infty$,

$$\begin{pmatrix} n^{1/2}(\theta_0^* - \theta_0) \\ (nh)^{1/2}(g_0^*(z) - g_0(z) + (W_\lambda g_0)(z)) \end{pmatrix} = \begin{pmatrix} n^{1/2}(\Omega + \Sigma_\lambda)^{-1} V(G, W_\lambda g_0) \\ -(nh)^{1/2} V(G^T, W_\lambda g_0)(\Omega + \Sigma_\lambda)^{-1} A(z) \end{pmatrix} \to 0.$$

A.4. *Proof of Theorem 4.4.* For notational convenience, denote $\widehat{f} = \widehat{f}_{n,\lambda}$, $\widehat{f}^0 = \widehat{f}_{n,\lambda}^{H_0}$, the constrained estimate of f under H_0, and $f = \widehat{f}^0 - \widehat{f} = (\theta, g)$. By Assumptions A4 and A5, with large probability, $\|f\| \le r_n$, where $r_n = M((nh)^{-1/2} + h^m)$ for some large M. By Assumption A1 (a), for some large constant $C > 0$, the event $B_n \equiv B_{n1} \cap B_{n2}$ has large probability, where $B_{n1} = \{\max_{1 \le i \le n} \sup_{a \in \mathcal{I}} |\ddot{\ell}_a(Y_i; a)| \le C \log n\}$ and $B_{n2} = \{\max_{1 \le i \le n} \sup_{a \in \mathcal{I}} |\ell_a'''(Y_i; a)| \le C \log n\}$. Let a_n be defined as in (2.17).

By Taylor's expansion,

$$\begin{aligned} LRT_{n,\lambda} &= \ell_{n,\lambda}(\widehat{f}^0) - \ell_{n,\lambda}(\widehat{f}) \\ &= S_{n,\lambda}(\widehat{f})f + \int_0^1\int_0^1 sDS_{n,\lambda}(\widehat{f}+ss'f)ffdsds' \\ &= \int_0^1\int_0^1 sDS_{n,\lambda}(\widehat{f}+ss'f)ffdads' \\ &= \int_0^1\int_0^1 s\{DS_{n,\lambda}(\widehat{f}+ss'f)ff - DS_{n,\lambda}(f_0)ff\}dsds' \\ &\quad +\frac{1}{2}(DS_{n,\lambda}(f_0)ff - E\{DS_{n,\lambda}(f_0)ff\}) \\ &\quad +\frac{1}{2}E\{DS_{n,\lambda}(f_0)ff\}, \end{aligned} \tag{A.6}$$

denote the above three sums by I_1, I_2 and I_3. Next we will study the asymptotic behavior of these sums. Denote $\widetilde{f} = \widehat{f} + ss'f - f_0 = (\widetilde{\theta}, \widetilde{g})$, for any $0 \le s, s' \le 1$. So $\|\widetilde{f}\| = O_P(r_n)$.

By calculations of the Frechét derivatives, we have

$$\begin{aligned} DS_{n,\lambda}(\widehat{f}+ss'f)ff &= DS_{n,\lambda}(\widetilde{f}+f_0)ff \\ &= \frac{1}{n}\sum_{i=1}^n \ddot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i) + X_i^T\widetilde{\theta} + \widetilde{g}(Z_i))(X_i^T\theta + g(Z_i))^2 - \langle P_\lambda f, f\rangle, \end{aligned}$$

and

$$DS_{n,\lambda}(f_0)ff = \frac{1}{n}\sum_{i=1}^n \ddot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i))(X_i^T\theta + g(Z_i))^2 - \langle P_\lambda f, f\rangle.$$

On B_n,

$$\begin{aligned} & |DS_{n,\lambda}(\widehat{f}+ss'f)ff - DS_{n,\lambda}(f_0)ff| \\ \le\ & \frac{1}{n}C(\log n)\|\widetilde{f}\|_{\sup}\sum_{i=1}^n (X_i^T\theta + g(Z_i))^2 \\ =\ & C(\log n)\|\widetilde{f}\|_{\sup}\langle \frac{1}{n}\sum_{i=1}^n (X_i^T\theta + g(Z_i))R_{U_i}, f\rangle \\ =\ & C(\log n)\|\widetilde{f}\|_{\sup}\langle \frac{1}{n}\sum_{i=1}^n (X_i^T\theta + g(Z_i))R_{U_i} - E_T\{(X^T\theta + g(Z))R_U\}, f\rangle \\ & +C(\log n)\|\widetilde{f}\|_{\sup}E_T\{(X^T\theta + g(Z))^2\}. \end{aligned} \tag{A.7}$$

Now we study $\frac{1}{n}\|\sum_{i=1}^n (X_i^T\theta + g(Z_i))R_{U_i} - E_T\{(X^T\theta + g(Z))R_U\}\|$. Let $d_n = c_m h^{-1/2} r_n$ and $\bar{f} = d_n^{-1}f/2 = (d_n^{-2}\theta/2, d_n^{-1}g/2) \equiv (\bar{\theta}, \bar{g})$. Consider $\psi(T; f) = X^T\theta + g(Z)$ and $\psi_n(T; \bar{f}) = (1/2)c_m^{-1}h^{1/2}d_n^{-1}\psi(T; 2d_n\bar{f})$. It is easy to see that $\psi_n(T; \bar{f})$, as a function of $\bar{f}$, satisfies the Lipschitz continuity condition (S.5).

Since $h = o(1)$ and $nh^2 \to \infty$, $d_n = o(1)$. Then by Lemma 2.4, on B_n, $\|\bar{f}\|_{\sup} \le 1/2$, which implies that for any $(x, z) \in \mathcal{U}$, $|x^T\bar{\theta} + \bar{g}(z)| \le 1/2$. Letting x approach zero, one gets that

$|\bar{g}(z)| \le 1/2$, and thus, $\|\bar{g}\|_{\sup} \le 1/2$, which further implies that $|x^T\bar{\theta}| \le \|\bar{g}\|_{\sup} + \|\bar{f}\|_{\sup} \le 1$ for any $x \in \mathbb{I}^p$. Also note that

$$\begin{aligned} J(\bar{g},\bar{g}) &= d_n^{-2}\lambda^{-1}(\lambda J(g,g))/4 \\ &\le d_n^{-2}\lambda^{-1}\|f\|^2/4 \\ &\le d_n^{-2}\lambda^{-1}r_n^2/4 \\ &< c_m^{-2}h\lambda^{-1}. \end{aligned}$$

Thus, when event B_n holds, $\bar{f} \equiv f_{\bar{\theta},\bar{g}}$ is an element in $\mathcal{G}$. Then by Lemma S.1 (in the supplementary material), with large probability

$$\begin{aligned} &\|\frac{1}{n}\sum_{i=1}^n[(X_i^T\theta + g(Z_i))R_{U_i} - E_T\{(X^T\theta + g(Z))R_U\}]\| \\ =\ & \frac{c_m h^{-1/2}d_n}{n}\|\sum_{i=1}^n[\psi_n(T_i;\bar{f})R_{U_i} - E_T\{\psi_n(T;\bar{f})R_U\}]\| \\ =\ & O_P(a_n'), \end{aligned} \tag{A.8}$$

where $a_n' = n^{-1/2}((nh)^{-1/2} + h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2}$. So by $a_n' = o(r_n)$,

$$\begin{aligned} |DS_{n,\lambda}(\widehat{f} + ss'f)ff - DS_{n,\lambda}(f_0)ff| &= \|\widetilde{f}\|_{\sup}(O_P(a_n'r_n\log n) + O_P(r_n^2\log n)) \\ &= h^{-1/2}r_nO_P(r_n^2\log n) \\ &= O_P(r_n^3h^{-1/2}\log n). \end{aligned} \tag{A.9}$$

Thus, $|I_1| = O_P(r_n^3h^{-1/2}\log n)$.

Next we approximate I_2. Define $\psi(T;f) = \ddot{\ell}_a(Y;X^T\theta_0 + g_0(Z))(X^T\theta + g(Z))$. Then by calculation of the Fréchet derivative (Section 2.2),

$$DS_{n,\lambda}(f_0)ff - E\{DS_{n,\lambda}(f_0)ff\} = \langle\frac{1}{n}\sum_{i=1}^n[\psi(T_i;f)R_{U_i} - E_T\{\psi(T;f)R_U\}], f\rangle.$$

Thus, $2|I_2| \le \frac{1}{n}\|\sum_{i=1}^n[\psi(T_i;f)R_{U_i} - E_T\{\psi(T;f)R_U\}]\| \cdot \|f\|$. So it is sufficient to approximate $\|\sum_{i=1}^n[\psi(T_i;f)R_{U_i} - E_T\{\psi(T;f)R_U\}]\|$. Let $\tilde{\psi}_n(T;\bar{f}) = (1/2)C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}\psi(T;2d_n\bar{f})$ and $\psi_n(T_i;\bar{f}) = \tilde{\psi}_n(T_i;\bar{f})I_{A_i}$, where $\bar{f} = d_n^{-1}f/2$ and $A_i = \{\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)| \le C\log n\}$ for $i = 1,\ldots,n$. By similar derivations as the ones below (A.7), it can be shown that on B_n, $\bar{f} \in \mathcal{G}$. Observe that B_n implies $\cap_i A_i$. A direct examination shows that ψ_n satisfies (S.5). By Lemma S.1, with large probability

$$\|\sum_{i=1}^n[\psi_n(T_i;\bar{f})R_{U_i} - E_T\{\psi_n(T;\bar{f})R_U\}]\| \le (n^{1/2}h^{-(2m-1)/(4m)} + 1)(5\log\log n)^{1/2}. \tag{A.10}$$

On the other hand, by Chebyshev's inequality

$$P(A_i^c) = \exp(-(C/C_0)\log n)E\{\exp(\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)|/C_0)\} \le C_1 n^{-C/C_0}.$$

Since $h = o(1)$ and $nh^2 \to \infty$, we may choose C to be large so that $(\log n)^{-1}n^{-C/(2C_0)} = o(a_n' h^{1/2} d_n^{-1})$, where $a_n' = n^{-1/2}((nh)^{-1/2}+h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2}$. By (2.3), which implies $E\{\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)|\big|U_i\} \le 2C_1C_0^2$, we have, on B_n, $E_T\{|\psi(T;2d_n\bar{f})|^2\} \le 2C_1C_0^2d_n^2$. So when n is large, on B_n, by Chebyshev's inequality

$$\begin{aligned}
& \|E_T\{\psi_n(T_i;\bar{f})R_{U_i}\} - E_T\{\tilde{\psi}_n(T_i;\bar{f})R_{U_i}\}\| \\
= \;& \|E_T\{\tilde{\psi}_n(T_i;\bar{f})R_{U_i}\cdot I_{A_i^c}\}\| \\
\le \;& (1/2)C^{-1}(\log n)^{-1}d_n^{-1}\left(E_T\{|\psi(T;2d_n\bar{f})|^2\}\right)^{1/2}P(A_i^c)^{1/2} \\
\le \;& (1/2)2^{1/2}C^{-1}C_0C_1(\log n)^{-1}n^{-C/(2C_0)} \\
= \;& o(a_n'h^{1/2}d_n^{-1}). \qquad \text{(A.11)}
\end{aligned}$$

Therefore, by (A.10) and (A.11), on B_n with large probability,

$$\begin{aligned}
& \frac{1}{n}\|\sum_{i=1}^n[\psi(T_i;f)R_{U_i} - E_T\{\psi(T;f)R_U\}]\| \\
= \;& \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}\|\sum_{i=1}^n[\tilde{\psi}_n(T_i;\bar{f})R_{U_i} - E_T\{\tilde{\psi}_n(T;\bar{f})R_U\}]\| \\
\le \;& \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}(\|\sum_{i=1}^n[\psi_n(T_i;\bar{f})R_{U_i} - E_T\{\psi_n(T;\bar{f})R_U\}]\| \\
& +n\|E_T\{\psi_n(T_i;\bar{f})R_{U_i}\} - E_T\{\tilde{\psi}_n(T_i;\bar{f})R_{U_i}\}\|) \\
\le \;& \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}\cdot[(n^{1/2}h^{-(2m-1)/(4m)}+1)(5\log\log n)^{1/2} + o(na_n'h^{1/2}d_n^{-1})] \\
\le \;& C'a_n'\log n.
\end{aligned} \qquad \text{(A.12)}$$

for some large constant $C' > 0$. Thus, $|I_2| = O_P(a_n' r_n \log n)$.

Note that $I_3 = -\|f\|^2/2$. Therefore, $-2n\cdot LRT_{n,\lambda} = n\|\widehat{f}^0-\widehat{f}\|^2 + O_P(nr_na_n'\log n + nr_n^3h^{-1/2}\log n) = n\|\widehat{f}^0-\widehat{f}\|^2 + O_P(nr_na_n\log n + nr_n^3h^{-1/2}\log n)$. By $r_n^2h^{-1/2} = o(a_n)$ and $nr_na_n = o((\log n)^{-1})$, it is easy to see that $O_P(nr_na_n\log n + nr_n^3h^{-1/2}\log n) = o_P(1)$. This shows $-2n\cdot LRT_{n,\lambda} = n\|\widehat{f}^0-\widehat{f}\|^2 + o_P(1)$. So we only focus on $n\|\widehat{f}^0-\widehat{f}\|^2$. By Theorems 2.6 and 4.3,

$$n^{1/2}\|\widehat{f}^0 - \widehat{f} - S_{n,\lambda}^0(f_0^0) + S_{n,\lambda}(f_0)\| = O_P(n^{1/2}a_n\log n) = o_P(1), \qquad \text{(A.13)}$$

so we just have to focus on $n^{1/2}\{S_{n,\lambda}^0(f_0^0) - S_{n,\lambda}(f_0)\}$. Recall that under H_0, $f_0^0 = (\theta_0^0, g_0^0) \in \mathcal{H}_0$,

so

$$\begin{aligned} S^0_{n,\lambda}(f^0_0) &= \frac{1}{n}\sum_{i=1}^n \dot{\ell}_a(Y_i; X_i^T\theta^0_0 + g^0_0(Z_i) + X_i^T\theta^\dagger + w^\dagger)R^0_{U_i} - P^0_\lambda f^0_0 \\ &= \frac{1}{n}\sum_{i=1}^n \dot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i))R^0_{U_i} - P^0_\lambda f^0_0 \\ &= \frac{1}{n}\sum_{i=1}^n \epsilon_i R^0_{U_i} - P^0_\lambda f^0_0, \end{aligned}$$

where $\epsilon_i = \dot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i))$, R^0_U and $P^0_\lambda f^0_0$ are defined in Section 4, and

$$S_{n,\lambda}(f_0) = \frac{1}{n}\sum_{i=1}^n \epsilon_i R_{U_i} - P_\lambda f_0.$$

Consequently,

$$\begin{aligned} & S^0_{n,\lambda}(f^0_0) - S_{n,\lambda}(f_0) \\ = \;& \frac{1}{n}\sum_{i=1}^n \epsilon_i(R^0_{U_i} - R_{U_i}) - (P^0_\lambda f^0_0 - P_\lambda f_0) \\ = \;& -\frac{1}{n}\sum_{i=1}^n \epsilon_i(\sum_{j=1}^k \rho_{U_i,j} R_{q_j,W_j}) + (\sum_{j=1}^k \zeta_j R_{q_j,W_j}) \\ = \;& -\frac{1}{n}\sum_{i=1}^n \epsilon_i(H(Q,W)\rho_{U_i}, (Q^TK_{z_0} - A^TH(Q,W))\rho_{U_i}) + (H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta) \\ = \;& (\xi,\beta) + (H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta), \end{aligned}$$

where $\beta = -\delta K_{z_0} - A^T\xi$, $\xi = -(1/n)\sum_{i=1}^n \epsilon_i H(Q,W)\rho_{U_i}$ and $\delta = (1/n)\sum_{i=1}^n \epsilon_i Q^T\rho_{U_i}$. Therefore,

$$\begin{aligned} \|S^0_{n,\lambda}(f^0_0) - S_{n,\lambda}(f_0)\|^2 &= \|(\xi,\beta)\|^2 + 2\langle(\xi,\beta), (H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta)\rangle \\ &\quad + \|(H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta)\|^2. \end{aligned}$$

We next evaluate the three items on the right side of the above equation. Denote $\Sigma_\lambda = E_U\{I(U)(G(Z) - A(Z))(G(Z) - A(Z))^T\}$. Note $E_Z\{B(Z)(G(Z) - A(Z))K_{z_0}(Z)\} = V(G,K_{z_0}) - V(A,K_{z_0}) =$

$\langle A, K_{z_0}\rangle_1 - V(A, K_{z_0}) = \langle W_\lambda A, K_{z_0}\rangle_1 = (W_\lambda A)(z_0)$. First,

$$\begin{aligned}
&\|(\xi,\beta)\|^2 \\
=\;& E_U\{I(U)(X^T\xi+\beta(Z))^2\} + \lambda J(\beta,\beta) \\
=\;& E_U\{I(U)[(X-A(Z))^T\xi - \delta K_{z_0}(Z)]^2\} + \lambda J(\beta,\beta) \\
=\;& \xi^T E_U\{I(U)(X-A(Z))(X-A(Z))^T\}\xi - 2\xi^T E_U\{I(U)(X-A(Z))K_{z_0}(Z)\}\delta \\
&+\delta^2 E_Z B(Z)|K_{z_0}(Z)|^2 + \langle W_\lambda(\delta K_{z_0} + A^T\xi), \delta K_{z_0} + A^T\xi\rangle_1 \\
=\;& \xi^T(\Omega+\Sigma_\lambda)\xi - 2\xi^T E_Z\{B(Z)(G(Z)-A(Z))K_{z_0}(Z)\}\delta + \delta^2 V(K_{z_0},K_{z_0}) \\
&+\delta^2\langle W_\lambda K_{z_0}, K_{z_0}\rangle_1 + 2\delta\xi^T\langle W_\lambda A, K_{z_0}\rangle_1 + \xi^T\langle W_\lambda A, A^T\rangle_1\xi \\
=\;& \xi^T\Gamma_\lambda\xi - 2\xi^T(W_\lambda A)(z_0)\delta + \delta^2 K(z_0,z_0) + 2\delta\xi^T(W_\lambda A)(z_0) \\
=\;& \xi^T\Gamma_\lambda\xi + \delta^2 K(z_0,z_0), \qquad \text{(A.14)}
\end{aligned}$$

where $\Gamma_\lambda = \Omega + \Sigma_\lambda + \langle W_\lambda A, A^T\rangle_1$ and $\Sigma_\lambda = E_Z\{B(Z)(G(Z)-A(Z))(G(Z)-A(Z))^T\}$. Second,

$$\begin{aligned}
&\langle(\xi,\beta), (H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta)\rangle \\
=\;& E_U\{I(U)[(X-A(Z))^T\xi - \delta K_{z_0}(Z)][(X-A(Z))^TH(Q,W)\zeta + Q^T\zeta K_{z_0}(Z)]\} \\
&+\langle W_\lambda\beta, Q^T\zeta K_{z_0} - A^TH(Q,W)\zeta\rangle_1 \\
=\;& \xi^T E_U\{I(U)(X-A(Z))(X-A(Z))^T\}H(Q,W)\zeta + \xi^T E_U\{I(U)(X-A(Z))K_{z_0}(Z)\}Q^T\zeta \\
&-\delta E_U\{I(U)K_{z_0}(Z)(X-A(Z))^T\}H(Q,W)\zeta - \delta Q^T\zeta V(K_{z_0},K_{z_0}) - \delta Q^T\zeta\langle W_\lambda K_{z_0}, K_{z_0}\rangle_1 \\
&+\delta(H(Q,W)\zeta)^T(W_\lambda A)(z_0) - Q^T\zeta\xi^T(W_\lambda A)(z_0) + \xi^T\langle W_\lambda A, A^T\rangle_1 H(Q,W)\zeta \\
=\;& \xi^T\Gamma_\lambda H(Q,W)\zeta - \delta Q^T\zeta K(z_0,z_0). \qquad \text{(A.15)}
\end{aligned}$$

Third, similar to the calculations in (A.14) and (A.15) we have

$$\begin{aligned}
&\langle(H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta), (H(Q,W)\zeta, (Q^TK_{z_0} - A^TH(Q,W))\zeta)\rangle \\
=\;& E_U\{I(U)[(X-A(Z))^TH(Q,W)\zeta + Q^T\zeta K_{z_0}(Z)]^2\} \\
&+\langle W_\lambda(Q^T\zeta K_{z_0} - A^TH(Q,W)\zeta), Q^T\zeta K_{z_0} - A^TH(Q,W)\zeta\rangle_1 \\
=\;& \zeta^TH(Q,W)^T\Gamma_\lambda H(Q,W)\zeta + (Q^T\zeta)^2K(z_0,z_0). \qquad \text{(A.16)}
\end{aligned}$$

It follows from (A.14) to (A.16) that

$$\begin{aligned}
\|S^0_{n,\lambda}(f^0_0) - S_{n,\lambda}(f_0)\|^2 &= (\xi + H(Q,W)\zeta)^T\Gamma_\lambda(\xi + H(Q,W)\zeta) + (\delta - Q^T\zeta)^2K(z_0,z_0) \\
&= \begin{pmatrix}\xi + H(Q,W)\zeta \\ \delta - Q^T\zeta\end{pmatrix}^T \begin{pmatrix}\Gamma_\lambda & 0 \\ 0 & K(z_0,z_0)\end{pmatrix}\begin{pmatrix}\xi + H(Q,W)\zeta \\ \delta - Q^T\zeta\end{pmatrix}. \qquad \text{(A.17)}
\end{aligned}$$

Next we find the limiting distribution of $n\|S^0_{n,\lambda}(f^0_0) - S_{n,\lambda}(f_0)\|^2$, which leads to the limiting distribution of $-2n \cdot LRT_{n,\lambda}$ in view of (A.13). By definition of ξ and the expressions of $H(Q,W)$, $T(Q,W)$, ρ_{U_i} and ζ in Section 4, we have

$$\begin{aligned}
\xi + H(Q,W)\zeta &= -\frac{1}{n}\sum_{i=1}^n \epsilon_i H(Q,W)M_K^{-1}(MH_{U_i} + QT_{U_i}(z_0)) + H(Q,W)M_K^{-1}(MH^*_{g_0} + QT^*_{g_0}(z_0)) \\
&= H(Q,W)M_K^{-1}N\left(-\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ T_{U_i}(z_0)\end{pmatrix} + \begin{pmatrix} H^*_{g_0} \\ T^*_{g_0}(z_0)\end{pmatrix}\right) \\
&= H(Q,W)M_K^{-1}N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1\end{pmatrix}\left(-\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ K_{z_0}(Z_i)\end{pmatrix} + \begin{pmatrix} H^*_{g_0} \\ (W_\lambda g_0)(z_0)\end{pmatrix}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\delta - Q^T\zeta &= Q^T M_K^{-1}N\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ T_{U_i}(z_0)\end{pmatrix} - \begin{pmatrix} H^*_{g_0} \\ T^*_{g_0}(z_0)\end{pmatrix}\right) \\
&= Q^T M_K^{-1}N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1\end{pmatrix}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ K_{z_0}(Z_i)\end{pmatrix} - \begin{pmatrix} H^*_{g_0} \\ (W_\lambda g_0)(z_0)\end{pmatrix}\right).
\end{aligned}$$

Therefore,

(A.18)
$$\begin{pmatrix} \xi + H(Q,W)\zeta \\ \delta - Q^T\zeta\end{pmatrix} = \begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix} M_K^{-1}N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1\end{pmatrix}\left(-\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ K_{z_0}(Z_i)\end{pmatrix} + \begin{pmatrix} H^*_{g_0} \\ (W_\lambda g_0)(z_0)\end{pmatrix}\right).$$

Define $\widetilde{M}_K = \begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix}^T \begin{pmatrix} \Gamma_\lambda & 0 \\ 0 & K(z_0,z_0)\end{pmatrix}\begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix}$, where recall that $\Gamma_\lambda = \Omega + \Sigma_\lambda + \langle W_\lambda A, A^T\rangle_1$. Since for any $1 \le j,k \le p$, $\langle W_\lambda A_k, A_j\rangle_1 = \lambda\sum_\nu V(A_j,h_\nu)V(A_k,h_n u)\gamma_\nu = O(\lambda) = o(1)$, we have as $\lambda \to 0$, $\langle W_\lambda A, A^T\rangle_1 \to 0$, a $p\times p$ zero matrix. Define λ_1 as the maximum eigenvalue of $\langle W_\lambda A, A^T\rangle_1$, and λ_2 as the minimum eigenvalue of $\Omega + \Sigma_\lambda$. Thus, $\lambda_1 = o(1)$. By equation (A.1) in Lemma A.1, λ_2 is asymptotically finitely upper bounded, and is lower bounded from zero. Note that

$$\begin{aligned}
\widetilde{M}_K - M_K &= \begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix}^T \begin{pmatrix} \langle W_\lambda A, A^T\rangle_1 & 0 \\ 0 & 0\end{pmatrix}\begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix} \\
&\le \frac{\lambda_1}{\lambda_2}\begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix}^T \begin{pmatrix} \Omega + \Sigma_\lambda & 0 \\ 0 & 0\end{pmatrix}\begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix} \\
&\le \frac{\lambda_1}{\lambda_2}\begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix}^T \begin{pmatrix} \Omega + \Sigma_\lambda & 0 \\ 0 & K(z_0,z_0)\end{pmatrix}\begin{pmatrix} H(Q,W) \\ -Q^T\end{pmatrix} = \frac{\lambda_1}{\lambda_2}M_K. \qquad \text{(A.19)}
\end{aligned}$$

Define

$$\Psi_\lambda = \begin{pmatrix} (\Omega+\Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix} \begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix} N^T M_K^{-1} \widetilde{M}_K M_K^{-1} N \begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1 \end{pmatrix} \begin{pmatrix} (\Omega+\Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix}.$$

Therefore, by (A.19),

$$0 \le \Psi_\lambda - \Phi_\lambda \le \frac{\lambda_1}{\lambda_2} \begin{pmatrix} (\Omega+\Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix} \begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix} N^T M_K^{-1} N \begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1 \end{pmatrix} \begin{pmatrix} (\Omega+\Sigma_\lambda)^{-1/2} & 0 \\ 0 & K(z_0,z_0)^{1/2} \end{pmatrix} = \frac{\lambda_1}{\lambda_2}\Phi_\lambda.$$

Since $\Phi_\lambda \le \mathrm{trace}(\Phi_\lambda) I_{p+1} = k I_{p+1}$, $\Psi_\lambda - \Phi_\lambda = o(1) I_{p+1}$. Thus, as $n \to \infty$, Ψ_λ approaches Φ_0.

Next we will conclude the proof by demonstrating the asymptotic distribution. It follows by Lemma 3.2 that $n^{1/2} H_{g_0}^* = o(1)$. Denote $N_U = (-H_U^T, K_Z(z_0)/K(z_0,z_0)^{1/2})^T$. By Assumption A1 (c),

$$E\{\epsilon^2 N_U N_U^T\} = E\left\{ I(U) \begin{pmatrix} H_U H_U^T & -H_U K_Z(z_0)/K(z_0,z_0)^{1/2} \\ -H_U^T K_Z(z_0)/K(z_0,z_0)^{1/2} & |K_Z(z_0)|^2/K(z_0,z_0) \end{pmatrix} \right\}.$$

To find the limit of this matrix, note that as $\lambda \to 0$, the following limits hold

- by Lemma A.1,

$$\begin{aligned} & E\{I(U) H_U H_U^T\} \\ = \; & (\Omega+\Sigma_\lambda)^{-1} E\{I(U)(X-A(Z))(X-A(Z))^T\}(\Omega+\Sigma_\lambda)^{-1} \\ = \; & (\Omega+\Sigma_\lambda)^{-1}(\Omega + E_Z\{B(Z)(G(Z)-A(Z))(G(Z)-A(Z))^T)(\Omega+\Sigma_\lambda)^{-1} \to \Omega^{-1}, \end{aligned}$$

- by $h^{1/2}(W_\lambda A)(z_0) \to 0$ and $hK(z_0,z_0) \to \sigma_{z_0}^2/c_0$ (by assumption (4.4)),

$$\begin{aligned} & E\{I(U) H_U K_Z(z_0)\}/K(z_0,z_0)^{1/2} \\ = \; & E\{I(U)(\Omega+\Sigma_\lambda)^{-1}(X-A(Z))K_Z(z_0)\}/K(z_0,z_0)^{1/2} \\ = \; & E\{B(Z)(G(Z)-A(Z))K_Z(z_0)\}/K(z_0,z_0)^{1/2} \\ = \; & (W_\lambda A)(z_0)/K(z_0,z_0)^{1/2} \to 0, \end{aligned}$$

- by assumption, $E\{B(Z)|K_Z(z_0)|^2\}/K(z_0,z_0) \to c_0$.

Thus, as $\lambda \to 0$, $E\{\epsilon^2 N_U N_U^T\} \to \begin{pmatrix} \Omega^{-1} & 0 \\ 0 & c_0 \end{pmatrix}$. So as $n \to \infty$,

$$n^{1/2} \begin{pmatrix} (\Omega+\Sigma_\lambda)^{1/2} & 0 \\ 0 & 1 \end{pmatrix} \left(-\frac{1}{n}\sum_{i=1}^n \epsilon_i \begin{pmatrix} H_{U_i} \\ \frac{K_{z_0}(Z_i)}{\sqrt{K(z_0,z_0)}} \end{pmatrix} + \begin{pmatrix} H_{g_0}^* \\ \frac{(W_\lambda g_0)(z_0)}{\sqrt{K(z_0,z_0)}} \end{pmatrix} \right) \xrightarrow{d} \upsilon, \tag{A.20}$$

where $\upsilon \sim N\left(\begin{pmatrix} 0 \\ c_{z_0} \end{pmatrix}, \begin{pmatrix} I_p & 0 \\ 0 & c_0 \end{pmatrix}\right)$. Therefore, it follows by (A.17), (A.18) and (A.20) that, as $n \to \infty$, $n\|S^0_{n,\lambda}(f^0_0) - S_{n,\lambda}(f_0)\|^2 \xrightarrow{d} \upsilon^T \Phi_0 \upsilon$. It immediately follows that $\|\widehat{f^0} - \widehat{f}\| = O_P(n^{-1/2})$. Besides, when $n \to \infty$, $-2n \cdot LRT_{n,\lambda} \xrightarrow{d} \upsilon^T \Phi_0 \upsilon$.

A.5. *Proof of Corollary 4.5.* By Fourier expansion of g_0 and $W_\lambda h_\nu = \frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu}$, we have $(W_\lambda g_0)(z_0) = \sum_\nu V(g_0, h_\nu)\frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu}h_\nu(z_0)$. By the assumption that $\sum_\nu |V(g_0, h_\nu)|^2\gamma_\nu^d < \infty$, one obtains the bound $|(W_\lambda g_0)(z_0)| = O((\lambda^d h^{-1})^{1/2}) = O(h^{md-1/2})$ by using Cauchy's inequality. Thus, by $h \asymp n^{-1/(2m+1)}$ and $d > 1 + 1/(2m)$, $(nh)^{1/2}(W_\lambda g_0)(z_0) = o(1)$. Direct calculations verify $h = o(1)$, $nh^2 \to \infty$, $a_n = o((nh)^{-1/2} + h^m)$, $a_n = o(n^{-1/2}h^{1/2}(\log n)^{-1})$, $a_n = o(n^{-1}((nh)^{-1/2} + h^m)^{-1}(\log n)^{-1})$, $a_n \gg ((nh)^{-1/2} + h^m)^2 h^{-1/2}$, and $n^{1/2}h^{m(1+b)} = o(1)$. Thus, the desired result follows from Theorem 4.3.

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. *Ann. Math. Statist.*, **37**, 577–581.

[3] Banerjee, M., Mukherjee, D. and Mishra, S. (2009). Semiparametric Binary Regression Models under Shape Constraints with an Application to Indian Schooling Data. *Journal of Econometrics*, **149**, 101–117.

[4] Boente, G., He, X. and Zhou, J. (2006). Robust estimates in generalized partially linear models. *Annals of Statistics*, **34**, 2856–2878.

[5] Bickel, P., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York.

[6] Cheng, G. (2009) Semiparametric Additive Isotonic Regression. *Journal of Statistical Planning and Inference*, **139**, 1980-1991

[7] Cheng, G. and Huang, J.Z. (2010) Bootstrap Consistency for General Semiparametric M-estimation. *Annals of Statistics*, **38**, 2884-2915

[8] Cheng, G. and Kosorok, M. R. (2008). General Frequentist Properties of the Posterior Profile Distribution. *Annals of Statistics*. **36**, 1819–1853.

[9] Cheng, G. and Kosorok, M. R. (2009). The penalized profile sampler. *Journal of Multivariate Analysis*. **100**, 345–362.

[10] Cheng, G. and Wang, X. (2011). Semiparametric additive transformation model under current status data. *Electronic Journal of Statistics*, **5**, 1735–1764.

[11] Cheng, G (2013). How many iterations are sufficient for efficient semiparametric estimation? *Scandinavian Journal of Statistics*, To Appear.

[12] Cox, D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.*, **18**, 1676–1695.

[13] Davis, P. J. (1963). Interpolation and Approximation, Blaisdell, New York.

[14] Gu, C. (2002). Smoothing Spline ANOVA Models. Springer-Verlag.

[15] Ke, C. and Wang, Y. (2002) ASSIST: A Suite of S-plus functions Implementing Spline smoothing Techniques. Preprint.

[16] Kim, J. and Pollard, D. (1990). Cube Root Asymptotics. *Annals of Statistics*, **18**, 191–219.

[17] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer: New York.

[18] Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics*, **32**, 1448–1491.

[19] Li, Y., Prentice, R. L. and Lin, X. (2008). Semiparametric maximum likelihood estimation in normal transformation models for bivariate survival data. *Biometrika*, **95**, 947–960.

[20] Mammen, E. and van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, **25**, 1014–1035.

[21] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd Edition. London: Chapman and Hall.

[22] Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood (with discussion). *J. Amer. Statist. Assoc.*, **95**, 449 – 485

[23] Nychka, D. (1995). Splines as local smoothers. *Ann. Statist.*, **23**, 1175–1197.

[24] Radchenko, P. (2008). Mixed-rates Asymptotics. *Ann. Statist.*, **36**, 287–309.

[25] Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Prob.*, **22**, 1679–1706.

[26] Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association*, **89**, 501–511

[27] Shang, Z. and Cheng, G. (2013). Local and Global Asymptotic Inference in Smoothing Spline Models. *Annals of Statistics*, To Appear.

[28] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer.

[29] Wahba, G. (1990). Spline Models for Observational Data. SIAM, Philidelphia.

[30] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Submitted to the Annals of Statistics

Supplementary materials to

JOINT ASYMPTOTICS FOR SEMI-NONPARAMETRIC MODELS UNDER PENALIZATION

By Guang Cheng [‡] and Zuofeng Shang[§]

Department of Statistics, Purdue University

November 11, 2013

In this document, we give the proofs of several results that were not given in the main text of this paper. The reference labels of the equations, Theorems, Propositions and Lemmas in this document are consistent with those in the main text of the paper.

This supplement document is organized as follows. Sections S.1 and S.2 contain the proofs of Propositions 2.1 and 2.2. Section S.4 contains the statement and proof of a useful concentration inequality, i.e., Lemma S.1. In Section S.5, we prove Proposition 2.5, i.e., the rates of convergence of the estimate $\widehat{f}_{n,\lambda}$. In Section S.6, we prove the *joint Bahadur representation*, i.e., Theorem 2.6. Sections S.7 and S.8 contain the proofs of Corollary 3.4 and Proposition 4.2. In Section S.9, we prove the limiting distribution of the local LRT test, i.e., Theorem 4.3. Section S.10 presents an algorithm in computing the constrained estimate under a local hypothesis.

S.1. *Proof of Proposition 2.1.* The idea is to use property (2.6) in the main paper to solve H_u and T_u. We use $E\{\}$ to denote the expectation with respect to the random pair $U = (X, Z)$. For

[‡]Corresponding Author. Associate Professor, Research Sponsored by NSF, DMS-0906497, CAREER Award DMS-1151692

[§]Visiting Assistant Professor

any $f = (\theta, g) \in \mathcal{H}$, by (2.5) in the main paper

$$\begin{aligned}
& \langle (H_u, T_u), f \rangle \\
= \; & E\{I(U)(X^T\theta + g(Z))(X^T H_u + T_u(Z))\} + \lambda J(T_u, g) \\
= \; & E\{I(U)(\theta^T X X^T H_u + \theta^T X T_u(Z) + g(Z) X^T H_u + g(Z) T_u(Z))\} + \lambda J(T_u, g) \\
= \; & \theta^T E\{I(U) X X^T\} H_u + \theta^T E\{I(U) X T_u(Z)\} + E\{I(U) g(Z) X^T\} H_u + E\{I(U) g(Z) T_u(Z)\} \\
& + \lambda J(T_u, g) \\
= \; & \theta^T (E\{I(U) X X^T\} H_u + E_Z\{B(Z) G(Z) T_u(Z)\}) + E\{B(Z) G(Z)^T g(Z)\} H_u + \langle g, T_u \rangle_1 \\
= \; & \theta^T (E\{I(U) X X^T\} H_u + E_Z\{B(Z) G(Z) T_u(Z)\}\}) + \langle A^T H_u, g \rangle_1 + \langle g, T_u \rangle_1 \\
= \; & \theta^T (E\{I(U) X X^T\} H_u + E_Z\{B(Z) G(Z) T_u(Z)\}) + \langle A^T H_u + T_u, g \rangle_1 \\
= \; & \theta^T x + \langle K_z, g \rangle_1.
\end{aligned}$$

Therefore, by arbitrariness of θ and g, we have the equations

$$\text{(S.1)} \qquad \begin{cases} E\{I(U) X X^T\} H_u + E_Z\{B(Z) G(Z) T_u(Z)\} = x \\ A^T H_u + T_u = K_z. \end{cases}$$

Note that from the second equation and the linearity property of W, we have $T_u = K_z - A^T H_u$. Then substitute it to the first equation, we have

$$\begin{aligned}
& x \\
= \; & E\{I(U) X X^T\} H_u + E\{B(Z) G(Z)(K_z(Z) - A(Z)^T H_u)\} \\
= \; & E\{I(U) X X^T\} H_u + E\{B(Z) G(Z) K_z(Z)\} \\
& - E\{B(Z) G(Z) G(Z)^T\} H_u + E\{B(Z) G(Z)(G(Z) - A(Z))^T\} H_u \\
= \; & (\Omega + \Sigma_\lambda) H_u + E\{B(Z) G(Z) K_z(Z)\} \\
= \; & (\Omega + \Sigma_\lambda) H_u + A(z),
\end{aligned}$$

where recall that $\Sigma_\lambda = E\{B(Z)G(Z)(G(Z) - A(Z))^T\}$ is a $p \times p$ matrix. In view of Lemma A.1 in the main paper, Σ_λ is asymptotically negligible, thus, $\Omega + \Sigma_\lambda$ is asymptotically positive definite. This gives us

$$\text{(S.2)} \qquad H_u = (\Omega + \Sigma_\lambda)^{-1}(x - A(z)),$$

Then from the second equation we have

$$T_u = K_z - A^T H_u = K_z - A^T (\Omega + \Sigma_\lambda)^{-1}(x - A(z)).$$

By this way we have constructed explicitly the $R_u = (H_u, T_u)$ such that for any $u = (x, z) \in \mathcal{S}$ and any $f = (\theta, g) \in \mathcal{H}$,

$$\langle R_u, f \rangle = x^T \theta + g(z).$$

S.2. *Proof of Proposition 2.2.* The idea is similar to the proof of (2.6) in the main paper. Denote $P_\lambda f = (H_g^*, T_g^*)$. Then we will use property (2.7) in the main paper to solve H_g^* and T_g^*. From (2.7) in the main paper,

$$\begin{aligned}
& \langle P_\lambda f, \widetilde{f}\rangle \\
= \;& E\{I(U)(X^T H_g^* + T_g^*(Z))(X^T\widetilde{\theta} + \widetilde{g}(Z))\} + \lambda J(T_g^*, \widetilde{g}) \\
= \;& (H_g^*)^T E\{I(U)XX^T\}\widetilde{\theta} + (H_g^*)^T E\{I(U)X\widetilde{g}(Z)\} + E\{I(U)T_g^*(Z)X^T\}\widetilde{\theta} \\
& + E\{I(U)T_g^*(Z)\widetilde{g}(Z)\} + \lambda J(T_g^*, \widetilde{g}) \\
= \;& ((H_g^*)^T E\{I(U)XX^T\} + E\{B(Z)G(Z)^T T_g^*(Z)\})\widetilde{\theta} \\
& + (H_g^*)^T E\{B(Z)G(Z)\widetilde{g}(Z)\} + \langle T_g^*, \widetilde{g}\rangle_1 \\
= \;& ((H_g^*)^T E\{I(U)XX^T\} + E\{B(Z)G(Z)^T T_g^*(Z)\})\widetilde{\theta} + \langle (H_g^*)^T A, \widetilde{g}\rangle_1 + \langle T_g^*, \widetilde{g}\rangle_1 \\
= \;& \langle W_\lambda g, \widetilde{g}\rangle_1.
\end{aligned}$$

So we get the equations

$$\begin{cases} (H_g^*)^T E\{I(U)XX^T\} + E\{B(Z)G(Z)^T T_g^*(Z)\} = 0 \\ (H_g^*)^T A + T_g^* = W_\lambda g. \end{cases}$$

The solution of the above equations is

$$\begin{cases} H_g^* = -(\Omega + \Sigma_\lambda)^{-1} E\{B(Z)G(Z)(W_\lambda g)(Z)\} \\ T_g^* = E\{B(Z)G(Z)^T(W_\lambda g)(Z)\}(\Omega + \Sigma_\lambda)^{-1} A + W_\lambda g. \end{cases}$$

So $P_\lambda f = (H_g^*, T_g^*)$ satisfies property (2.7) in the main paper. Obviously P_λ is linear and self-adjoint, i.e., $\langle P_\lambda f, \widetilde{f}\rangle = \langle f, P_\lambda \widetilde{f}\rangle$. The boundedness of P_λ follows from the inequality (2.14) in the main paper.

S.3. *Proof of Lemma 2.4.* For any $u = (x, z) \in \mathcal{S}$, $|\langle R_u, f\rangle| \le \|R_u\| \cdot \|f\|$, so we only need to find the bound for $\|R_u\|$. By definition of R_u,

$$\begin{aligned}
\langle R_u, R_u\rangle &= x^T x_u + T_u(z) \\
&= K(z, z) + (x - E\{B(Z)G(Z)K_z(Z)\})^T(\Omega + \Sigma_\lambda)^{-1}(x - E\{B(Z)G(Z)K_z(Z)\}).
\end{aligned}$$

By (2.9) and the boundedness of h_μs,

$$K(z, z) = \sum_{\mu\in\mathbb{Z}} \frac{|h_\mu(z)|^2}{1 + \lambda\gamma_\mu} \le C \sum_{\mu\in\mathbb{Z}} \frac{1}{1 + \lambda\gamma_\mu} \le c\lambda^{-1/(2m)} = ch^{-1}, \tag{S.3}$$

where c is a constant that does not rely on z. On the other hand,

$$\begin{aligned} & |E\{B(Z)G_k(Z)K_z(Z)\}|^2 \\ = \; & |V(G_k, K_z)| \\ = \; & |\sum_\mu V(G_k, h_\mu)\frac{h_\mu(z)}{1+\lambda\gamma_\mu}|^2 \\ \le \; & \sum_\mu |V(G_k, h_\mu)|^2 \sum_\mu \frac{|h_\mu(z)|^2}{(1+\lambda\gamma_\mu)^2} \\ \le \; & c' h^{-1}, \end{aligned}$$

where c' is a constant that does not rely on z. So there exists a constant $c_m > 0$ which does not rely on u such that $\|R_u\| \le c_m h^{-1/2}$. Then the conclusion of Lemma 2.4 holds.

S.4. *A Concentration Inequality.* The *concentration inequality* Lemma S.1 stated below will be used as a preliminary step in obtaining JBR.

Define $\mathcal{G}_1 = \{g_1(x) = x^T\theta : x \in \mathbb{I}^p, \|g_1\|_{\sup} \le 1, \theta \in \mathbb{R}^p\}$ and $\mathcal{G}_2 = \{g_2(z) \in H^m(\mathbb{I}) : \|g_2\|_{\sup} \le 1, J(g_2, g_2) \le c_m^{-2} h\lambda^{-1}\}$, where the constant c_m is specified in Lemma 2.4. Let $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 \equiv \{g_1(x) + g_2(z) : g_1 \in \mathcal{G}_1 \text{ and } g_2 \in \mathcal{G}_2\}$. For any $f \in \mathcal{G}$, define the empirical processes $Z_n(f)$ as

$$Z_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [\psi_n(T_i; f)R_{U_i} - E_T(\psi_n(T; f)R_U)], \tag{S.4}$$

where $\psi_n(T; f)$ is a real-valued function (possibly depending on n) defined on $\mathcal{T} \times \mathcal{G}$. Recall that $U_i = (X_i, Z_i)$'s are covariates and $T_i = (Y_i, X_i, Z_i)$s denote the full data variables with domain $\mathcal{T}$.

Lemma S.1. *Suppose that ψ_n satisfies the following Lipschitz continuity:*

$$|\psi_n(T; f) - \psi_n(T; \tilde{f})| \le c_m^{-1} h^{1/2} \|f - \tilde{f}\|_{\sup} \quad \textit{for any } f, \tilde{f} \in \mathcal{G}, \tag{S.5}$$

where c_m is specified in Lemma 2.4. Then we have

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

We first establish and prove a preliminary lemma for proving Lemma S.1. Denote $N(\delta, \mathcal{G}, \|\cdot\|_{\sup})$ as the δ-covering number of the function class $\mathcal{G}$ in terms of the uniform norm.

Lemma S.2. *Suppose that $c_m^{-2} h\lambda^{-1} > 1$. Then for any $\delta > 0$,*

$$\log N(\delta, \mathcal{G}, \|\cdot\|_{\sup}) \le C(h\lambda^{-1})^{1/(2m)} \delta^{-1/m},$$

where $C > 0$ is an universal constant.

PROOF OF LEMMA S.2. By [17, Lemma 9.25],

$$N(2\delta, \mathcal{G}, \|\cdot\|_{\sup}) \le N(\delta, \mathcal{G}_1, \|\cdot\|_{\sup}) N(\delta, \mathcal{G}_2, \|\cdot\|_{\sup}).$$

Since $N(\delta, \mathcal{G}_1, \|\cdot\|_{\sup})$ is dominated by $N(\delta, \mathcal{G}_2, \|\cdot\|_{\sup})$, it is sufficient to bound $N(\delta, \mathcal{G}_2, \|\cdot\|_{\sup})$. Note that by $c_m^{-2} h\lambda^{-1} > 1$,

$$\mathcal{G}_2 = (c_m^{-2}h\lambda^{-1})^{1/2} \cdot \{g_2 \in H^m(\mathbb{I}) | \|g_2\|_{\sup} \le (c_m^{-2}h\lambda^{-1})^{-1/2}, J(g_2, g_2) \le 1\} \subset (c_m^{-2}h\lambda^{-1})^{1/2}\mathcal{T},$$

where $\mathcal{T} = \{g \in H^m(\mathbb{I}) | \|g\|_{\sup} \le 1, J(g,g) \le 1\}$. So by [28],

$$\begin{aligned}
\log N(\delta, \mathcal{G}_2, \|\cdot\|_{\sup}) &\le \log N(\delta, (c_m^{-2}h\lambda^{-1})^{1/2}\mathcal{T}, \|\cdot\|_{\sup}) \\
&= \log N((c_m^{-2}h\lambda^{-1})^{-1/2}\delta, \mathcal{T}, \|\cdot\|_{\sup}) \\
&\le c((c_m^{-2}h\lambda^{-1})^{-1/2}\delta)^{-1/m} \\
&= c c_m^{-1/m} (h\lambda^{-1})^{1/(2m)} \delta^{-1/m}.
\end{aligned}$$

□

Now we prove Lemma S.1. For any $g, f \in \mathcal{G}$, by Lemma 2.4,

$$\begin{aligned}
\|(\psi_n(T; f) - \psi_n(T; g))R_U\| &\le c_m^{-1} h^{1/2} \|f - g\|_{\sup} \cdot \|R_U\| \\
&\le c_m^{-1} h^{1/2} \|f - g\|_{\sup} \cdot c_m h^{-1/2} = \|f - g\|_{\sup}.
\end{aligned}$$

By Theorem 3.5 of [25], for any $t > 0$

$$P\left(\|Z_n(f) - Z_n(g)\| \ge t\right) \le 2\exp\left(-\frac{t^2}{8\|f - g\|_{\sup}^2}\right).$$

Then by Lemma 8.1 in [17], we have

$$\left\| \|Z_n(g) - Z_n(f)\| \right\|_{\psi_2} \le 8\|g - f\|_{\sup},$$

where $\|\cdot\|_{\psi_2}$ denotes the Orlicz norm associated with $\psi_2(s) \equiv \exp(s^2) - 1$. Then it follows by Lemma S.2 and Theorem 8.4 of [17] that for arbitrary $\delta > 0$,

$$\begin{aligned}
&\left\| \sup_{\substack{g,f \in \mathcal{G} \\ \|g-f\|_{\sup} \le \delta}} \|Z_n(g) - Z_n(f)\| \right\|_{\psi_2} \\
&\le C' \left(\int_0^\delta \sqrt{\log(1 + N(\delta, \mathcal{G}, \|\cdot\|_{\sup}))} + \delta \sqrt{\log(1 + N(\delta, \mathcal{G}, \|\cdot\|_{\sup})^2)} \right) \\
&\asymp h^{-(2m-1)/(4m)} \delta^{1-1/(2m)}.
\end{aligned}$$

So, again, by Lemma 8.1 in [17],

$$P\left(\sup_{\substack{g\in\mathcal{G}\\ \|g\|_{\sup}\le\delta}}\|Z_n(g)\|\ge t\right)\le 2\exp(-h^{(2m-1)/(2m)}\delta^{-2+1/m}t^2). \tag{S.6}$$

Let $b_n = n^{1/2}h^{-(2m-1)/(4m)}$, $\varepsilon = b_n^{-1}$, $\gamma = 1-1/(2m)$, $T_n = (5\log\log n)^{1/2}$, and $Q_\varepsilon = [\gamma - \log_2\varepsilon - 1]$, where $[a]$ denotes the integer part of a. Then by (S.6),

$$\begin{aligned}
& P\left(\sup_{\substack{g\in\mathcal{G}\\ \|g\|_{\sup}\le 2}}\frac{\sqrt{n}\|Z_n(g)\|}{a_n\|g\|_{\sup}^{\gamma}+1}\ge T_n\right)\\
\le\ & P\left(\sup_{\substack{g\in\mathcal{G}\\ \|g\|_{\sup}\le\varepsilon^{1/\gamma}}}\frac{\sqrt{n}\|Z_n(g)\|}{a_n\|g\|_{\sup}^{\gamma}+1}\ge T_n\right)\\
& +\sum_{l=0}^{Q_\varepsilon}P\left(\sup_{\substack{g\in\mathcal{G}\\ (2^l\varepsilon)^{1/\gamma}\le\|g\|_{\sup}\le(2^{l+1}\varepsilon)^{1/\gamma}}}\frac{\sqrt{n}\|Z_n(g)\|}{a_n\|g\|_{\sup}^{\gamma}+1}\ge T_n\right)\\
\le\ & P\left(\sup_{\substack{g\in\mathcal{G}\\ \|g\|_{\sup}\le\varepsilon^{1/\gamma}}}\sqrt{n}\|Z_n(g)\|\ge T_n\right)\\
& +\sum_{l=0}^{Q_\varepsilon}P\left(\sup_{\substack{g\in\mathcal{G}\\ \|g\|_{\sup}\le(2^{l+1}\varepsilon)^{1/\gamma}}}\sqrt{n}\|Z_n(g)\|\ge (1+2^l)T_n\right)\\
\le\ & 2\exp\left(-h^{(2m-1)/(2m)}(\varepsilon^{1/\gamma})^{-2+1/m}T_n^2/n\right)\\
& +\sum_{l=0}^{Q_\varepsilon}2\exp\left(-h^{(2m-1)/(2m)}[(2^{l+1}\varepsilon)^{1/\gamma}]^{-2+1/m}T_n^2(2^l+1)^2/n\right)\\
=\ & 2\exp\left(-T_n^2\right)+\sum_{l=0}^{Q_\varepsilon}2\exp\left(-2^{-2(l+1)}T_n^2(2^l+1)^2\right)\\
\le\ & 2(Q_\varepsilon+2)\exp\left(-T^2/4\right)\le \text{const}\cdot\log n(\log n)^{-5/4}\to 0,
\end{aligned}$$

as $n\to\infty$. This proves the result.

S.5. *Proof of Proposition 2.5. Recall that we have used (θ, g) without distinguishing to denote the bivariate function $f_{\theta,g}:(x,z)\mapsto x^T\theta+g(z)$*

Consider the function classes

$$\mathcal{F}_1=\{f_1(x)=x^T\theta, x\in\mathbb{I}^p|\|f_1\|_{\sup}\le 1,\theta\in\mathbb{R}^p\},\quad\text{and}$$

$$\mathcal{F}_2=\{f_2(z)\in H^m(\mathbb{I})|\|f_2\|_{\sup}\le 1, J(g_2,g_2)\le 1\}.$$

Denote $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$. By a modification of Lemma S.2, it can be shown that for any $\delta > 0$,

$$\log N(\delta, \mathcal{F}, \|\cdot\|_{\sup}) \le c\delta^{-1/m},$$

where c is some universal constant. Then a modification of Lemma S.1 leads to

Lemma S.3. *Suppose that ψ_n satisfies Lipschitz continuity, namely,*

$$|\psi_n(T; f) - \psi_n(T; g)| \le c_m^{-1} h^{1/2} \|f - g\|_{\sup}, \text{ for all } f, g \in \mathcal{F}, \tag{S.7}$$

where c_m is specified in Lemma 2.4. Then we have

$$\lim_{n\to\infty} P\left(\sup_{\substack{g\in\mathcal{F} \\ \|g\|_{\sup}\le 1}} \frac{\|Z_n(g)\|}{\|g\|_{\sup}^{1-1/(2m)} + n^{-1/2}} \le (5\log\log n)^{1/2} \right) = 1,$$

where the empirical process $Z_n(f)$ is defined in (S.4).

Denote $f = \widehat{f}_{n,\lambda} - f_0 = (\theta, g)$. By the consistency of $\widehat{f}_{n,\lambda}$ in $\|\cdot\|_{\mathcal{H}}$-norm and Sobolev embedding Theorem (see [1]), we know that $x^T\widehat{\theta} + \widehat{g}(z)$ falls in $\mathcal{I}$ for any $(x, z) \in \mathcal{U}$ and large enough n. By Taylor's expansion,

$$\ell_{n,\lambda}(f_0 + f) - \ell_{n,\lambda}(f_0) = S_{n,\lambda}(f_0)f + \frac{1}{2}DS_{n,\lambda}(f_0)ff + \frac{1}{6}D^2S_{n,\lambda}(f^\star)fff \ge 0,$$

where $f^\star = f_0 + t^\star f$ for some $t^\star \in [0, 1]$. Denote the three sums on the right side of the above equation by I_1, I_2, I_3. Next we will study the rates for these terms. Denote $A_i = \{\sup_{a\in\mathcal{I}} |\ell_a'''(Y_i; a)| \le C\log n\}$. By (2.4) in the main paper, we may choose C to be large so that $\cap_i A_i$ has large probability, and $P(A_i^c) = O(n^{-1})$. Then on $\cap_i A_i$,

$$\begin{aligned}
|6I_3| &\le \frac{1}{n}\sum_{i=1}^n \sup_{a\in\mathcal{I}} |\ell_a'''(Y_i; a)| \cdot |X_i^T\theta + g(Z_i)|^3 \\
&\le \frac{1}{n}\|f\|_{\sup} \sum_{i=1}^n \sup_{a\in\mathcal{I}} |\ell_a'''(Y_i; a)| \cdot (X_i^T\theta + g(Z_i))^2 \\
&= \frac{1}{n}\|f\|_{\sup} \langle \sum_{i=1}^n \psi(T_i; f) R_{U_i}, f\rangle \\
&= \frac{1}{n}\|f\|_{\sup} \langle \sum_{i=1}^n [\psi(T_i; f)R_{U_i} - E_T\{\psi(T; f)R_U\}], f\rangle + \|f\|_{\sup} E_T\{\psi(T; f)(X^T\theta + g(Z))\},
\end{aligned}$$

where $\psi(T_i; f) = \sup_{a\in\mathcal{I}} |\ell_a'''(Y_i; a)|(X_i^T\theta + g(Z_i))I_{A_i}$. Let $\psi_n(T_i; f) = (C\log n)^{-1}c_m^{-1}h^{1/2}\psi(T_i; f)$, which satisfies (S.7). Thus, by Lemma S.3, for large n and with large probability,

$$\|\sum_{i=1}^n [\psi_n(T_i; f)R_{U_i} - E_T\{\psi_n(T; f)R_U\}]\| \le (n^{1/2}\|f\|_{\sup}^{1-1/(2m)} + 1)(5\log\log n)^{1/2}.$$

So

$$\begin{aligned} & \langle \sum_{i=1}^{n} [\psi(T_i; f) R_{U_i} - E_T\{\psi(T; f) R_U\}], f \rangle \\ \leq \; & \|f\| \cdot (n^{1/2} \|f\|_{\sup}^{1-1/(2m)} + 1)(5 \log\log n)^{1/2}. \end{aligned}$$

On the other hand, by Assumption A1 (a)

$$E_T\{\psi(T; f)(X^T\theta + g(Z))\} \leq E_T\{\sup_{a \in \mathcal{I}} |\ell'''_a(Y; a)|(X^T\theta + g(Z))^2\} \leq 2C_0^2 C_1 \|f\|^2.$$

By $(n^{1/2}h)^{-1}(\log\log n)^{m/(2m-1)}(\log n)^{2m/(2m-1)} = o(1)$, which implies $(n^{1/2}h)^{-1}(\log\log n)^{1/2}\log n = o(1)$, we have

$$\begin{aligned} & |6I_3| \\ \leq \; & \frac{1}{n}\|f\|_{\sup} \cdot \|f\|(C \log n) c_m h^{-1/2}(n^{1/2}\|f\|_{\sup}^{1-1/(2m)} + 1)(5\log\log n)^{1/2} + 2C_0^2 C_1 \|f\|_{\sup} \cdot \|f\|^2 \\ = \; & c_m^2 C (n^{1/2}h)^{-1}(\log\log n)^{1/2}(\log n)\|f\|_{\sup}^{1-1/(2m)} + c_m^2 C (nh)^{-1}(\log\log n)^{1/2}(\log n)\|f\|^2 \\ \text{(S.8)} \; & o_P(1) \cdot \|f\|^2. \end{aligned}$$

To approximate I_2, by Cauchy's inequality we have

$$\begin{aligned} & \left| E_T\{\ddot{\ell}_a(Y; X^T\theta_0 + g_0(Z)) I_{A^c}(X^T\theta + g(Z))^2\} \right| \\ \leq \; & E_T\{|\ddot{\ell}_a(Y; X^T\theta_0 + g_0(Z))|^2 I_{A^c}(X^T\theta + g(Z))^4\}^{1/2} \cdot P(A^c)^{1/2} \\ \leq \; & O(1) \cdot \log n \|f\|_{\sup}\|f\| n^{-1} = \|f\|^2 O((nh)^{-1/2}) = o(1)\|f\|^2. \end{aligned}$$

By changing ψ and ψ_n in the proof of (S.8) to $\psi(T_i; f) = \ddot{\ell}_a(Y_i; X_i^T\theta_0 + g_0(Z_i))(X_i^T\theta + g(Z_i))I_{A_i}$ and $\psi_n(T_i; f) = (C\log n)^{-1} c_m^{-1} h^{1/2}\psi(T_i; f)$, and using an argument similar to the proof of (S.8), it can be shown that

$$\begin{aligned} & |[DS_{n,\lambda}(f_0) - E_T\{DS_{n,\lambda}(f_0)\}]ff| \\ \leq \; & C c_m h^{-1+1/(4m)} n^{-1/2}(\log\log n)^{1/2}(\log n)\|f\|^{2-1/(2m)} \\ & + C c_m (nh^{1/2})^{-1}(\log\log n)^{1/2}(\log n)\|f\| + o_P(1)\|f\|^2. \end{aligned}$$

Thus,

$$\begin{aligned} 2I_2 \; = \; & -\|f\|^2 + C c_m h^{-1+1/(4m)} n^{-1/2}(\log\log n)^{1/2}(\log n)\|f\|^{2-1/(2m)} \\ & + C c_m (nh^{1/2})^{-1}(\log\log n)^{1/2}(\log n)\|f\| + o_P(1)\|f\|^2. \end{aligned} \tag{S.9}$$

Note that $E\{\|\sum_{i=1}^n \epsilon_i R_{U_i}\|^2\} = O(nh^{-1})$, and by (2.14), $\|P_\lambda f_0\| \leq \sqrt{\lambda J(g_0, g_0)} = O(\lambda^{1/2})$, we have $\|S_{n,\lambda}(f_0)\| = O_P((nh)^{-1/2} + \lambda^{1/2})$. Combining (S.8) and (S.9), and the fact that

$$(nh^{1/2})^{-1}(\log\log n)^{1/2}(\log n) = o((nh)^{-1/2}),$$

we have for some large C'

$$(1+o_P(1))\|f\|^2 \le C'((nh)^{-1/2}+\lambda^{1/2})\|f\| + Cc_m h^{-1+1/(4m)} n^{-1/2}(\log\log n)^{1/2}(\log n)\|f\|^{2-1/(2m)}.$$

Solving this inequality, and using $(n^{1/2}h)^{-1}(\log\log n)^{m/(2m-1)}(\log n)^{2m/(2m-1)} = o(1)$, we get $\|f\| = O_P((nh)^{-1/2}+\lambda^{1/2})$.

S.6. *Proof of Theorem 2.6.* By Assumption A1 (a), it is not difficult to check the following

$$\max_{1\le i\le n}\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)| = O_P(\log n). \tag{S.10}$$

So we can let $C > C_0$ be sufficiently large so that the event $B_{n1} = \{\max_{1\le i\le n}\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)| \le C\log n\}$ has large probability.

Denote $f = \widehat{f}_{n,\lambda} - f_0 \equiv (\theta, g)$. By Assumption A4, the event $B_{n2} = \{\|f\| \le r_n \equiv M((nh)^{-1/2} + h^m)\}$ has large probability with some preselected large M. So $B_n = B_{n1}\cap B_{n2}$ has large probability. Define $\bar{f} \equiv (\bar{\theta}, \bar{g}) = d_n^{-1}f/2$ (so $\bar{\theta} = d_n^{-1}\theta/2$ and $\bar{g} = d_n^{-1}g/2$), where $d_n = c_m r_n h^{-1/2}$. Since $h = o(1)$ and $nh^2 \to \infty$, $d_n = o(1)$. Then by Lemma 2.4, on B_n, $\|\bar{f}\|_{\sup} \le 1/2$, which implies that for any $(x,z)\in\mathcal{S}$, $|x^T\bar{\theta} + \bar{g}(z)| \le 1/2$. Letting x approach zero, one gets that $|\bar{g}(z)| \le 1/2$, and thus, $\|\bar{g}\|_{\sup} \le 1/2$, which further implies that $|x^T\bar{\theta}| \le \|\bar{g}\|_{\sup} + \|\bar{f}\|_{\sup} \le 1$ for any $x\in\mathbb{I}^p$. Also note that

$$\begin{aligned} J(\bar{g},\bar{g}) &= d_n^{-2}\lambda^{-1}(\lambda J(g,g))/4 \\ &\le d_n^{-2}\lambda^{-1}\|f\|^2/4 \\ &\le d_n^{-2}\lambda^{-1}r_n^2/4 \\ &< c_m^{-2}h\lambda^{-1}. \end{aligned}$$

Thus, when event B_n holds, $\bar{f} \equiv f_{\bar{\theta},\bar{g}}$ is an element in $\mathcal{G}$.

Define $\psi(T;f) = \dot{\ell}_a(Y; X^T\theta + g(Z) + X^T\theta_0 + g_0(Z)) - \dot{\ell}_a(Y; X^T\theta_0 + g_0(Z))$. By the definition of $S_{n,\lambda}$ and S_n, and a direct calculation, one can verify that

$$S_n(f+f_0) - S(f+f_0) - (S_n(f_0) - S(f_0)) = \frac{1}{n}\sum_{i=1}^n [\psi(T_i;f)R_{U_i} - E_T(\psi(T;f)R_U)].$$

Let $\tilde{\psi}_n(T;\bar{f}) = (1/2)C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}\psi(T;2d_n\bar{f})$ and $\psi_n(T_i;\bar{f}) = \tilde{\psi}_n(T_i;\bar{f})I_{A_i}$, where $A_i = \{\sup_{a\in\mathcal{I}}|\ddot{\ell}_a(Y_i;a)| \le C\log n\}$ for $i = 1,\ldots,n$. Observe that B_n implies $\cap_i A_i$.

Next we show that ψ_n satisfies (S.5) in the main paper. For any $\bar{f}_1 = (\theta_1, g_1)$, $\bar{f}_2 = (\theta_2, g_2) \in \mathcal{G}$, and $(x,z)\in\mathcal{S}$, since $x^T\theta_0 + g_0(z) \in \mathcal{I}_0$ and $d_n = o(1)$, both $x^T\theta_0 + g_0(z) + 2d_n(x^T\theta_1 + g_1(z))$ and $x^T\theta_0 + g_0(z) + 2d_n(x^T\theta_2 + g_2(z))$ fall in $\mathcal{I}$ when n is sufficiently large (recall that $\mathcal{I}_0$ and $\mathcal{I}$ are

specified in Assumption A1). Therefore,

$$\begin{aligned}
& |\psi_n(T_i;\bar f_1)-\psi_n(T_i;\bar f_2)| \\
= \;& (1/2)C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}|\psi(T_i;2d_n\bar f_1)-\psi(T_i;2d_n\bar f_2)|\cdot I_{A_i} \\
= \;& (1/2)C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}\Big|\int_{X_i^T\theta_0+g_0(Z_i)}^{X_i^T\theta_0+g_0(Z_i)+2d_n(X^T\theta_1+g_1(Z_i))}\ddot\ell_a(Y_i;a)\cdot I_{A_i}da \\
& -\int_{X_i^T\theta_0+g_0(Z_i)}^{X_i^T\theta_0+g_0(Z_i)+2d_n(X_i^T\theta_1+g_1(Z_i))}\ddot\ell_a(Y_i;a)\cdot I_{A_i}da\Big| \\
\le \;& C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}\cdot d_n\|\bar f_1-\bar f_2\|_{\sup}\cdot\sup_{a\in\mathcal{I}}|\ddot\ell_a(Y_i;a)|\cdot I_{A_i} \\
\le \;& C^{-1}c_m^{-1}(\log n)^{-1}h^{1/2}d_n^{-1}\cdot d_n\cdot C\log n\cdot\|\bar f_1-\bar f_2\|_{\sup} \\
= \;& c_m^{-1}h^{1/2}\|\bar f_1-\bar f_2\|_{\sup}.
\end{aligned}$$

Thus, ψ_n satisfies (S.5). By Lemma S.1, on B_n with large probability

$$\|\sum_{i=1}^n[\psi_n(T_i;\bar f)R_{U_i}-E_T\{\psi_n(T;\bar f)R_U\}]\|\le(n^{1/2}h^{-(2m-1)/(4m)}+1)(5\log\log n)^{1/2}. \tag{S.11}$$

On the other hand, by Chebyshev's inequality

$$P(A_i^c)=\exp(-(C/C_0)\log n)E\{\exp(\sup_{a\in\mathcal{I}}|\ddot\ell_a(Y_i;a)|/C_0)\}\le C_1n^{-C/C_0}.$$

Since $h=o(1)$ and $nh^2\to\infty$, we may choose C to be large so that $(\log n)^{-1}n^{-C/(2C_0)}=o(a_n'h^{1/2}d_n^{-1})$, where $a_n'=n^{-1/2}((nh)^{-1/2}+h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2}$. By (2.3) in the main paper, which implies $E\{\sup_{a\in\mathcal{I}}|\ddot\ell_a(Y_i;a)|\big|U_i\}\le 2C_1C_0^2$, we have, on B_n, $E_T\{|\psi(T;d_ng)|^2\}\le 2C_1C_0^2d_n^2$. So when n is large, on B_n, by Chebyshev's inequality

$$\begin{aligned}
& \|E_T\{\psi_n(T_i;\bar f)R_{U_i}\}-E_T\{\tilde\psi_n(T_i;\bar f)R_{U_i}\}\| \\
= \;& \|E_T\{\tilde\psi_n(T_i;\bar f)R_{U_i}\cdot I_{A_i^c}\}\| \\
\le \;& C^{-1}(\log n)^{-1}d_n^{-1}\left(E_T\{|\psi(T;2d_n\bar f)|^2\}\right)^{1/2}P(A_i^c)^{1/2} \\
\le \;& 2^{1/2}C^{-1}C_0C_1(\log n)^{-1}n^{-C/(2C_0)} \\
\le \;& Ma_n'h^{1/2}d_n^{-1}.
\end{aligned}$$

Therefore, by (S.11) and on B_n,

$$\begin{aligned}
& \|S_n(f+f_0)-S(f+f_0)-(S_n(f_0)-S(f_0))\| \\
= \;& \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}\|\sum_{i=1}^n[\tilde{\psi}_n(T_i;\bar{f})R_{U_i}-E_T\{\tilde{\psi}_n(T;\bar{f})R_U\}]\| \\
\le \;& \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}(\|\sum_{i=1}^n[\psi_n(T_i;\bar{f})R_{U_i}-E_T\{\psi_n(T;\bar{f})R_U\}]\| \\
& +n\|E_T\{\psi_n(T_i;\bar{f})R_{U_i}\}-E_T\{\tilde{\psi}_n(T_i;\bar{f})R_{U_i}\}\|) \\
\le \;& \frac{2Cc_m(\log n)h^{-1/2}d_n}{n}\cdot[(n^{1/2}h^{-(2m-1)/(4m)}+1)(5\log\log n)^{1/2}+nMa_n'h^{1/2}d_n^{-1}] \\
\le \;& C'c_m a_n'\log n, \qquad \text{(S.12)}
\end{aligned}$$

for some large constant $C'>0$.

By Taylor's expansion, by the fact that $S_{n,\lambda}(f+f_0)=0$, and by Proposition 2.3,

$$\begin{aligned}
& \|S_n(f+f_0)-S(f+f_0)-(S_n(f_0)-S(f_0))\| \\
= \;& \|S_{n,\lambda}(f+f_0)-S_\lambda(f+f_0)-S_{n,\lambda}(f_0)+S_\lambda(f_0)\| \\
= \;& \|S_\lambda(f+f_0)+S_{n,\lambda}(f_0)-S_\lambda(f_0)\| \\
= \;& \|DS_\lambda(f_0)f+\int_0^1\int_0^1 sD^2S_\lambda(f_0+ss'f)ffdsds'+S_{n,\lambda}(f_0)\| \\
= \;& \|-f+\int_0^1\int_0^1 sD^2S_\lambda(f_0+ss'f)ffdsds'+S_{n,\lambda}(f_0)\| \\
\ge \;& \|-f+S_{n,\lambda}(f_0)\|-\|\int_0^1\int_0^1 sD^2S_\lambda(f_0+ss'f)ffdsds'\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \|f-S_{n,\lambda}(f_0)\| \\
\le \;& \|S_n(f+f_0)-S(f+f_0)-(S_n(f_0)-S(f_0))\| \\
& +\|\int_0^1\int_0^1 sD^2S_\lambda(f_0+ss'f)ffdsds'\| \\
\le \;& \|S_n(f+f_0)-S(f+f_0)-(S_n(f_0)-S(f_0))\| \\
& +\int_0^1\int_0^1 s\|D^2S_\lambda(f_0+ss'f)ff\|dsds'. \qquad \text{(S.13)}
\end{aligned}$$

Next we find an upper bound for $\|D^2S_\lambda(f_0+ss'f)ff\|$. The Frechét derivative of DS_λ is found to be $D^2S_\lambda=D^2S$, therefore,

$$\begin{aligned}
D^2S_\lambda(f_0+ss'f)ff &= D^2S(f_0+ss'f)ff \\
&= E(\ell_a'''(Y;X^T(\theta_0+ss'\theta)+(g_0+ss'g)(Z))(X^T\theta+g(Z))^2R_U).
\end{aligned}$$

Hence, by (2.4) in the main paper, on B_n,

$$\begin{aligned} & \|D^2 S_\lambda(f_0+ss'f)ff\| \\ = \; & \|E\{\ell'''_a(Y;X^T(\theta_0+ss'\theta)+(g_0+ss'g)(Z))(X^T\theta+g(Z))^2 R_U\}\| \\ \le \; & E\{E\{\sup_{a\in\mathcal{I}}|\ell'''_a(Y;a)|\big|U\}(X^T\theta+g(Z))^2\|R_U\|\} \\ \le \; & C_\ell c_m h^{-1/2}\|f\|^2, \end{aligned} \tag{S.14}$$

where $C_\ell = \sup_{u\in\mathcal{U}} E\{\sup_{a\in\mathcal{I}}|\ell'''_a(Y;a)|\big|U=u\}$. Thus, from (S.12), (S.13) and (S.14), with large probability

$$\|f - S_{n,\lambda}(f_0)\| \le C' c_m a'_n \log n + C_\ell c_m h^{-1/2}((nh)^{-1/2}+h^m)^2.$$

This completes the proof of Theorem 2.6.

S.7. *Proof of Corollary 3.4.* We first show part (i). By $\ell'''_a(y;a)=0$ for any y and a, that is, in (2.17) of the main paper $C_\ell = 0$, we obtain $a_n = n^{-1/2}((nh)^{-1/2}+h^m)h^{-(6m-1)/(4m)}(\log\log n)^{1/2}$. Since $h \asymp n^{-1/(4m+1)}$ and $b > 1+1/(2m)$, we have $h=o(1)$, $nh^2\to\infty$, and $n^{1/2}h^{m(1+b)}=o(1)$. By $m > 1+\sqrt{3}/2$, it can be verified that $a_n \log n = o(n^{-1/2}h^{1/2})$.

On the other hand, by expression of K in terms of h_νs (see (2.9)), as $h\to 0$,

$$\begin{aligned} & \int_0^1 g_0^{(2m)}(z)K(z_0,z)dz - g_0^{(2m)}(z_0)/\pi(z_0) \\ = \; & \sum_\nu \frac{1}{1+\lambda\gamma_\nu} V(g_0^{(2m)}/\pi, h_\nu)h_\nu(z_0) - \sum_\nu V(g_0^{(2m)}/\pi,h_\nu)h_\nu(z_0) \\ = \; & -\sum_\nu \frac{\lambda\gamma_\nu}{1+\lambda\gamma_\nu} V(g_0^{(2m)}/\pi,h_\nu)h_\nu(z_0) \to 0, \end{aligned} \tag{S.15}$$

where the limit in (S.15) follows from $\sum_\nu |V(g_0^{(2m)},h_\nu)h_\nu(z_0)|<\infty$ and dominated convergence theorem. Then, by (3.11) in the main paper and by integration by parts, it can be shown that

$$\begin{aligned} & (W_\lambda g_0)(z_0) = \langle W_\lambda g_0, K_{z_0}\rangle_1 = \lambda J(g_0,K_{z_0}) \\ = \; & (-1)^m h^{2m}\int_0^1 g_0^{(2m)}(z)K(z_0,z)dz = (-1)^m h^{2m}(g_0^{(2m)}(z_0)/\pi(z_0)+o(1)). \end{aligned} \tag{S.16}$$

So, as $n\to\infty$, $(nh)^{1/2}(W_\lambda g_0)(z_0) \to (-1)^m g_0^{(2m)}(z_0)/\pi(z_0)$. Therefore all the assumptions in Theorem 3.3 hold. Then (3.12) directly follows from (3.10).

The proof of (3.13) is similar to that of (3.12). One only notes, by (S.16) and $h\asymp n^{-d}$ for $\frac{1}{4m+1} < d \le \frac{2m}{10m-1}$, $(nh)^{1/2}(W_\lambda g_0)(z_0) = O((nh)^{1/2}h^{2m}) = o(1)$. Then (3.13) follows from (3.10).

The proof of part (ii) is similar to that of part (i). The only difference is that since g_0 does not satisfy the boundary conditions, and by integration by parts, (S.16) should be replaced by the

following

(S.17)
$$(W_\lambda g_0)(z_0) = h^{2m}\sum_{j=1}^{m}(-1)^{j-1}\left[\left(\frac{\partial^{m-j}}{\partial z^{m-j}}K_{z_0}^{m-j}(z)\right)\cdot g_0^{(m+j-1)}(z)\Big|_0^1\right]+(-1)^m h^{2m}\int_0^1 g_0^{(2m)}(z)K(z_0,z)dz.$$

The first sum, by (3.14), is $o(h^{2m})$. The second sum, by (S.15), is $(-1)^m h^{2m}(g_0^{(2m)}(z_0)/\pi(z_0)+o(1))$. Thus, $(W_\lambda g_0)(z_0) = (-1)^m h^{2m} g_0^{(2m)}(z_0)/\pi(z_0)+o(h^{2m})$. Note this is not true for $z_0 = 0$ or 1. Then the proof can be finished by similar arguments in the proof of part (i). Note that the derivation of Ψ can be referred to Example 5.1.

S.8. *Proof of Proposition 4.2.* Denote the orthogonal complement of $\mathcal{H}_0$ and the subspace spanned by $\{R_{q_j,W_j}\}_{j=1}^k$ as $\mathcal{H}_0^\perp$ and $\text{span}\{R_{q_j,W_j}\}_{j=1}^k$, respectively. We first show that

$$\text{span}\{R_{q_j,W_j}\}_{j=1}^k = \mathcal{H}_0^\perp. \tag{S.18}$$

It is easy to see by (4.3) that $\text{span}\{R_{q_j,W_j}\}_{j=1}^k \subset \mathcal{H}_0^\perp$. On the other hand, any element belonging to the orthogonal complement of $\text{span}\{R_{q_j,W_j}\}_{j=1}^k$, which ought to be orthogonal to each R_{q_j,W_j}, is an element in $\mathcal{H}_0$. Therefore, $\mathcal{H}_0^\perp \subset \text{span}\{R_{q_j,W_j}\}_{j=1}^k$. The above analysis implies (S.18).

Since $R_u - R_u^0 \in \mathcal{H}_0^\perp$, we can write $R_u - R_u^0 = \sum_{j=1}^k \rho_{u,j}R_{q_j,W_j}$. For simplicity, let $\rho_u = (\rho_{u,1},\ldots,\rho_{u,k})^T \in \mathbb{R}^k$. Hence, it amounts to finding the form of ρ_u. Since $R_u - \sum_{j=1}^k \rho_{u,j}R_{q_j,W_j} = (H_u - \sum_{j=1}^k \rho_{u,j}H_{q_j,W_j}, T_u - \sum_{j=1}^k \rho_{u,j}T_{q_j,W_j})$ belongs to $\mathcal{H}_0$, we have $M(H_u - \sum_{j=1}^k \rho_{u,j}H_{q_j,W_j}) + Q(T_u(z_0) - \sum_{j=1}^k \rho_{u,j}T_{q_j,W_j}(z_0)) = 0$. Then, we have $M_K\rho_u = MH_u + QT_u(z_0)$, which implies $\rho_u = M_K^{-1}(MH_u + QT_u(z_0))$. The invertibility of M_K follows from the direct calculations below, based on (4.3),

$$\begin{aligned} M_K &= K(z_0,z_0)QQ^T + (M - QA(z_0)^T)(\Omega+\Sigma_\lambda)^{-1}(M - QA(z_0)^T)^T \\ &= N\begin{pmatrix} I_p & 0 \\ -A(z_0)^T & 1 \end{pmatrix}\begin{pmatrix} (\Omega+\Sigma_\lambda)^{-1} & 0 \\ 0 & K(z_0,z_0) \end{pmatrix}\begin{pmatrix} I_p & -A(z_0) \\ 0 & 1 \end{pmatrix}N^T, \end{aligned}$$

where N is assumed to be full rank. With the above explicit ρ_u, we have obtained that $R_u^0 = R_u - \sum_{j=1}^k \rho_{u,j}R_{q_j,W_j}$. As for P_λ^0, we can write $P_\lambda^0 f = P_\lambda f - \sum_{j=1}^k \zeta_j(f)R_{q_j,W_j}$ based on (S.18), for some $\zeta(f) \equiv (\zeta_1(f),\ldots,\zeta_k(f))^T \in \mathbb{R}^k$. Similar arguments as above give that $\zeta(f) = M_K^{-1}(MH_g^* + QT_g^*(z_0))$.

S.9. *Proof of Theorem 4.3.* The proof is similar to those in Theorem 2.6, so we only sketch the idea. Let $f = \widehat{f}_{n,\lambda}^0 - f_0^0$. Assumption A5 guarantees that with large probability, $\|f\| \le r_n \equiv M((nh)^{-1/2} + h^m)$ for a proper large M. By a modification of the proof of Lemma S.1, we have the following lemma.

Lemma S.4. *Suppose that ψ_n satisfies Lipschitz continuity, namely, there exists a constant $C_\psi > 0$ such that*

$$|\psi_n(T; f_1) - \psi_n(T; f_2)| \le c_m^{-1} h^{1/2} \|f_1 - f_2\|_{\sup}, \text{ for all } f_1, f_2 \in \mathcal{H}_0, \tag{S.19}$$

where recall that $T = (Y, X, Z)$ denotes the full data variable. Then we have

$$\lim_{n\to\infty} P\left(\sup_{\substack{f\in\mathcal{G}_0 \\ \|f\|_{\sup}\le 1}} \frac{\|Z_n^0(f)\|}{n^{1/2}h^{-(2m-1)/(4m)}\|f\|_{\sup}^{1-1/(2m)} + 1} \le (5\log\log n)^{1/2} \right) = 1,$$

where $\mathcal{G}_0 = \mathcal{H}_0 \cap \mathcal{G}$, and $Z_n^0(f) = \sum_{i=1}^n [\psi_n(T_i; f) R_{U_i}^0 - E_T\{\psi_n(T; f) R_U^0\}]$.

By a reexamination of the proof of Theorem 2.6, we have, with large probability, $f = (\theta, g) \in \mathcal{G}_0$ and ψ_n satisfies Lipschitz continuity (S.19), where

$$\psi_n(T; f) = C^{-1} c_m^{-1} (\log n)^{-1} h^{1/2} d_n^{-1} \{\dot\ell_a(Y; X^T(\theta_0^0 + d_n\theta) + g_0^0(Z) + d_n g(Z)) - \dot\ell_a(Y; X^T\theta_0^0 + g_0^0(Z))\},$$

and $d_n = c_m r_n h^{-1/2}$. This leads to, with large probability,

$$\|\sum_{i=1}^n [\psi_n(T_i; g) R_{U_i}^0 - E_T\{\psi_n(T; f) R_U^0\}]\|_1 \le (n^{1/2} h^{-(2m-1)/(4m)} + 1)(5\log\log n)^{1/2}.$$

The remainder of the proof follows by (S.10), and by an argument similar to (S.12) through (S.14).

S.10. *A Computational Algorithm for the Constrained Estimate Under A Local Hypothesis.* Consider testing $H_0 : x_0\theta + g(z_0) = 0$ where $x_0 \in (0, 1)$ and θ_0 are both one-dimensional. We demonstrate how to find constrained estimate of $\widehat f_{n,\lambda}^0 = (\widehat\theta_{n,\lambda}^0, \widehat g_{n,\lambda}^0)$ under H_0, i.e., to find the solution to

$$\arg \min_{(\theta,g): x_0\theta + g(z_0) = 0} n^{-1} \sum_{i=1}^n (Y_i - X_i\theta - g(Z_i))^2/(2\widehat\sigma^2) + \lambda' J(g, g). \tag{S.20}$$

Let $R(s, t)$ be the reproducing kernel function associated with the roughness penalty J. Under H_0, the quadratic function in (S.20) is rewritten as

$$n^{-1} \sum_{i=1}^n (Y_i + X_i g(z_0)/x_0 - g(Z_i))^2/(2\widehat\sigma^2) + \lambda' J(g, g). \tag{S.21}$$

Let $\tilde K$ be the reproducing kernel for $H_0^m(\mathbb{I})$ associated with the usual inner product $\langle\cdot,\cdot\rangle_0$ defined in [29] . Define $\eta_i = \tilde K_{Z_i} - X_i \tilde K_{z_0}/x_0$, for $i = 1, \ldots, n$. Then $\langle\eta_i, g\rangle_0 = g(Z_i) - X_i g(z_0)/x_0$. It follows from Theorem 1.3.1 of [29] that the minimizer of (S.21) has the form $\widehat g_{n,\lambda}^* = \alpha + \sum_{i=1}^n \beta_i \mathcal{P}\eta_i$, where α, β_i are constants and $\mathcal{P}$ is the projection from $H_0^m(\mathbb{I})$ to the complement of the null space of J. It

can be seen from Chapter 2 of [29] that $\mathcal{P}\tilde{K}_z = R_z$, for any $z \in \mathbb{I}$. Then $P\eta_i = R_{Z_i} - X_i R_{z_0}/x_0$. Let Q be a symmetric $n \times n$ matrix with entries $Q_{ij} = R(Z_i, Z_j) - X_j R(Z_i, z_0)/x_0 - X_i R(Z_j, z_0)/x_0 + X_i X_j R(z_0, z_0)/x_0^2$, S be an n-vector with elements $S_i = 1 - X_i/x_0$. Let $Y = (Y_1, \ldots, Y_n)^T$ and $\beta = (\beta_1, \ldots, \beta_n)^T$. Minimizing (S.21) can thus be achieved by optimizing the following function of α, β:

$$\|Y - S\alpha - Q\beta\|^2 + \widehat{\sigma}^2 n\lambda\beta^T Q\beta. \tag{S.22}$$

The optimization of (S.22) can be achieved by similar arguments in [14]. Specifically, the solution has the form

$$\beta = F_2(F_2^T Q F_2 + n\lambda\widehat{\sigma}^2 I)^{-1} F_2^T Y, \quad \alpha = (S^T S)^{-1} S^T (Y - (Q + n\lambda\widehat{\sigma}^2 I)\beta).$$

Department of Statistics
Purdue University
250 N. University Street
West Lafayette, IN 47906
Email: chengg@purdue.edu; shang9@purdue.edu