The Annals of Statistics
1993, Vol. 21, No. 1, 217–234

SMOOTHING SPLINE DENSITY ESTIMATION: THEORY

BY CHONG GU[1] AND CHUNFU QIU[2]

Purdue University

In this article, a class of penalized likelihood probability density estimators is proposed and studied. The true log density is assumed to be a member of a reproducing kernel Hilbert space on a finite domain, not necessarily univariate, and the estimator is defined as the unique unconstrained minimizer of a penalized log likelihood functional in such a space. Under mild conditions, the existence of the estimator and the rate of convergence of the estimator in terms of the symmetrized Kullback–Leibler distance are established. To make the procedure applicable, a semiparametric approximation of the estimator is presented, which sits in an adaptive finite dimensional function space and hence can be computed in principle. The theory is developed in a generic setup and the proofs are largely elementary. Algorithms are yet to follow.

1. Introduction. Let X_i, $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) samples from an unknown probability density f on a domain $\mathscr{X}$. The estimation of f from the samples is of lasting interest to statisticians. When the density f is known to belong to a finite dimensional parametric family, say $P_\theta = \{f(\theta)\colon \theta \in \Theta\}$, where the form of f is known up to a finite dimensional parameter θ, density estimation reduces to parameter estimation, and the maximum likelihood (ML) method is the standard technique which possesses many favorable properties. Note that a parametric approach puts rigid constraints on the estimator. When a parametric form is not available, however, a naive ML density estimator without any nonintrinsic constraint (see below for the intrinsic constraints) is a sum of delta function spikes at the sample points, which apparently is not an appealing estimator when the domain $\mathscr{X}$ is continuous. The middle ground between these two extremes is where the nonparametric/semiparametric methods come in to play. Of course all estimators have to be bound by the intrinsic positivity constraint that $f \geq 0$ and the unity constraint that $\int_{\mathscr{X}} f = 1$.

In their pioneering article, Good and Gaskins (1971) introduced the idea of penalized likelihood density estimation. The idea is to minimize a penalized minus log likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) + (\lambda/2)J(f),$$

Received April 1991; revised May 1992.
[1]Research supported by NSF Grant DMS-91-01730.
[2]Research supported by NSF Grant DMS-87-17799.
AMS 1991 *subject classifications*. Primary 62G07; secondary 65D07, 65D10, 41A25, 41A65.
Key words and phrases. Density estimation, penalized likelihood, rate of convergence, reproducing kernel Hilbert space, semiparametric approximation, smoothing splines.

where the $J(f)$ is a roughness penalty and the λ is called the smoothing parameter. The log likelihood dictates the estimate to adapt to the data, the roughness penalty counteracts by demanding less variation (hence less adaptiveness) in f, and the smoothing parameter controls the tradeoff between the two conflicting goals. Presumably the $J(f)$ evaluates infinity at the delta sum, which is then effectively ruled out by the procedure. The null space of J, say $J_\perp$, is usually of finite dimension, so as $\lambda \to \infty$ the method reduces to the standard parametric ML estimation with $P_\theta = J_\perp$.

In implementing the Good–Gaskins method, a main concern is the incorporation of the positivity and unity constraints. Leonard (1978) introduced the logistic density transform $f = e^g / \int e^g$ and proposed to estimate g via minimizing

$$-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \log\int e^g + (\lambda/2)J(g), \tag{1.1}$$

which is constraint free. However, note that f determines g only up to a constant but the constant function is usually in $J_\perp$, so the operating criterion (1.1) may not have a unique solution if care is not taken. Silverman (1982) proposed to estimate the log density $g = \log f$ which is free of the positivity constraint, and to augment (1.1) by a functional $\int e^g$ to effectively enforce the unity constraint, ending up solving a constraint-free problem with a unique solution. Note that when g is a log density the second term in (1.1) disappears, so in appearance Silverman's method replaces $\log \int e^g$ in (1.1) by $\int e^g$. Silverman (1982) then developed a theory for his estimator, including the existence, the asymptotic convergence rates under various function norms, and the asymptotic Gaussian process approximation. Cox and O'Sullivan (1990) developed a general asymptotic theory for penalized likelihood estimators, which applies also to Silverman's method. Leonard's (1978) and Silverman's (1982) treatments are mainly in a univariate context. To the authors' knowledge, the multivariate counterpart is largely unexplored, although the Cox–O'Sullivan theory may well apply if one were available.

In this article, we propose to enforce a one-to-one logistic density transform by imposing a one dimensional side condition concerning the constant on the space of the log likelihood g. The minimizer $\hat{g}$ of (1.1) is then unique in the restricted function space when it exists, and the minimization problem is free of constraint. We shall formulate the problem in a general reproducing kernel Hilbert space, and develop a general theory in parallel to that of Silverman (1982).

The rest of the article is organized as follows. In Section 2, we formulate the problem, discuss the basic ingredients of the method and give examples. Section 3 illustrates the estimation of inhomogeneous Poisson intensity via density estimation and notes that the estimator studied in this article agrees with Silverman's (1982) estimator. Section 4 establishes the existence of the estimator under the condition that the ML solution exists in $J_\perp$. Section 5 establishes the asymptotic convergence rate of the "ideal" estimator $\hat{g}$ in

terms of the symmetrized Kullback–Leibler distance between the truth and the estimator using quadratic approximations adapted from Silverman (1982) and Cox and O'Sullivan (1990). Section 6 describes a semiparametric approximation $\hat{g}_n$ which is computable and shows that this estimator is asymptotically as good as $\hat{g}$. Computational methods for the semiparametric approximation is further developed in Gu (1993) and portable code available.

2. The definition of the estimator and examples. A reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions in which the evaluation functional is continuous. We will confine the log likelihood g to a RKHS $\mathscr{H}$ of functions on $\mathscr{X}$, and define the estimator as a minimizer of (1.1) in $\mathscr{H}$, where $J(g)$ is a square seminorm in $\mathscr{H}$ with a finite dimensional null space $J_\perp$. A Hilbert space carries a metric and a geometry which helps us in conducting theoretical and numerical calculations. Continuous evaluation ensures the continuity of the log likelihood part of (1.1) under mild conditions. A finite dimensional $J_\perp$, presumably of dimension less than n, prevents "interpolation," a conceptual equivalent of the delta function sum. The choice of J as a quadratic form is for practicality and by convention. Following Wahba's (1990) general definition, the estimator, as a minimizer of a functional involving a quadratic penalty in a Hilbert space, is a smoothing spline. This justifies the title of the article. Wahba (1990) has a thorough treatment of smoothing splines in the regression setup.

To ensure a one-to-one logistic density transform $f = e^g/\int e^g$, one needs to eliminate from $\mathscr{H}$ all but one log likelihood which differ only by a constant from each other. This can be done by enforcing a side condition $\int_B g\,d\nu = 0$ for members of $\mathscr{H}$, where ν is a measure on $\mathscr{X}$ with $\nu(B) > 0$. For example, such a side condition could be $g(x_0) = 0$ for a certain $x_0 \in \mathscr{X}$. In most applications of the smoothing spline technique, $\mathscr{H}$ is taken as $\{g\colon J(g) < \infty\}$ and $J_\perp \supseteq \{1\}$, and the norm in $\mathscr{H}$ is defined via augmenting $J(g)$ by a norm in $J_\perp$, say $\|g\|_\perp$, which is usually a sum of norms in one-dimensional spaces $\operatorname{span}\{\phi_\nu\}$ where ϕ_ν span $J_\perp$. When 1 is one of the ϕ_ν, which is usually the case, a RKHS $\mathscr{H} = \{g\colon J(g) < \infty\} \ominus \{1\}$, which is to be our requirement, can be easily obtained by dropping $\{1\}$ from the conventional construction. Examples follow after a bit more general treatment.

Throughout the remaining of the article, it is assumed that $1 \notin \mathscr{H}$ and $J_\perp = \{g\colon g \in \mathscr{H},\ J(g) = 0\}$. Let $L(g) = -(1/n)\sum_{i=1}^n g(X_i) + \log \int e^g$.

LEMMA 2.1. *$L(g)$ is strictly convex in $\mathscr{H}$. Consequently, the minimizer of (1.1) in $\mathscr{H}$, if it exists, is unique.*

PROOF. By Hölder's inequality, for $\alpha, \beta > 0$, $\alpha + \beta = 1$ and $g, h \in \mathscr{H}$,

$$\log \int e^{\alpha g + \beta h} \le \alpha \log \int e^g + \beta \log \int e^h,$$

where the equality holds only when $e^g \propto e^h$, which amounts to $g = h$ in $\mathscr{H}$. □

LEMMA 2.2. *If $e^{|g|}$ are Riemann integrable on $\mathscr{X}$ for all $g \in \mathscr{H}$, then $L(g)$ is continuous in $\mathscr{H}$. Furthermore, $L(g + \alpha h)$, $\forall\ g, h \in \mathscr{H}$ and $\alpha \in R$, is infinitely differentiable as a function of α.*

PROOF. The claims follow from the Riemann sum approximations of related integrals and the continuity of evaluation. □

Riemann integrability can be assured by a finite dimensional bounded domain $\mathscr{X}$ and the continuity of $g \in \mathscr{H}$ on $\mathscr{X}$. Later developments only require $L(g + \alpha h)$ to be twice differentiable.

Before proceeding with examples, we review a few basic properties of RKHS. Details are to be found in Aronszajn (1950). An equivalent defining property of a RKHS $\mathscr{H}$ is that $\mathscr{H}$ possesses a reproducing kernel (RK) $R(\cdot, \cdot)$, a positive definite bivariate function on $\mathscr{X}$, such that $R(x, \cdot) = R(\cdot, x) \in \mathscr{H}$, $\forall\ x \in \mathscr{X}$, and $\langle R(x, \cdot), f(\cdot) \rangle = f(x)$ (the reproducing property), $\forall\ f \in \mathscr{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathscr{H}$. As a matter of fact, starting from any positive definite function $R(\cdot, \cdot)$ on the domain $\mathscr{X}$, one can construct a RKHS $\mathscr{H} = \text{span}\{R(x, \cdot), \forall\ x \in \mathscr{X}\}$ with an inner product satisfying $\langle R(x, \cdot), R(y, \cdot) \rangle = R(x, y)$, which has $R(\cdot, \cdot)$ as its RK. Another useful result is that when $\mathscr{H} = \mathscr{H}_0 \oplus \mathscr{H}_1$ is a RKHS with the RK R, then $\mathscr{H}_i$ are RKSH's with RK's R_i, $i = 0, 1$, where $R_0 + R_1 = R$. Note that the RK and the norm in $\mathscr{H}$ define each other uniquely, though explicit formulas are not always available simultaneously for both. An explicit norm, or J for the purpose of this article, provides the most direct intuition about the notion of smoothness in the estimation procedure. An explicit RK, on the other hand, is the only thing needed to perform numerical calculations. We mention in passing that the quadratic penalty J is equivalent to a mean zero partially improper Gaussian process prior for g on $\mathscr{X}$, where the Gaussian process has two independent components, one diffuses in $J_\perp$ and the other has covariance function R_J, the RK associated with the norm J in $\mathscr{H} \ominus J_\perp$; see, for example, Wahba (1978) and Leonard (1978).

We present a few examples in the remaining of the section.

EXAMPLE 2.1 (Cubic spline on $[0, 1]$). Let $\mathscr{X} = [0, 1]$ and $J(g) = \int \ddot{g}^2$. The null space of J without side condition is $\{1, x\}$. There are at least two different formulations which lead to the same estimated density $e^g / \int e^g$, but the two estimated log likelihoods belong to two different RKHS's.

The first formulation specifies $\int g = 0$. The accompanying norm in $J_\perp = \{x - 0.5\}$ is $\|g\|_\perp = (\int \dot{g})^2$ and the associated RK in $\mathscr{H} \ominus J_\perp$ is $R_J(x, y) = k_2(x)k_2(y) - k_4(|x - y|)$ where $k_\nu = B_\nu / \nu!$ and B_ν is the νth Bernoulli polynomial; see Craven and Wahba (1979). It can be verified that $\int_0^1 R_J(x, y)\, dy = 0$ and $\int_0^1 (\partial^2 R_J(x, y)/\partial y^2) \ddot{g}(y)\, dy = g(x)$, $\forall\ g \in \mathscr{H} \ominus J_\perp$.

The second formulation specifies $g(0) = 0$. The accompanying norm in $J_\perp = \{x\}$ is $\|g\|_\perp = (\dot{g}(0))^2$ and the associated RK in $\mathscr{H} \ominus J_\perp$ is $R_J(x, y) = \int_0^1 (x - u)_+ (y - u)_+\, du$, where $(\cdot)_+$ is the positive part of $(\cdot)$. It is easy to

check that

$$R_J(x,0) = 0$$

and

$$\int_0^1 \left(\partial^2 R_J(x,y)/\partial y^2\right)\ddot{g}(y)\,dy = g(x),$$

$$\forall\, g \in \mathscr{H} \ominus J_\perp = \{g\colon g(0) = \dot{g}(0) = 0,\ J(g) < \infty\}.$$

EXAMPLE 2.2 (Thin plate splines). Thin plate splines are defined on $\mathscr{X} = R^d$ with

$$J(g) = J_m^d(g)$$

$$= \sum_{\alpha_1 + \cdots + \alpha_d = m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^m g}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}\right)^2 \times dx_1 \cdots dx_d. \tag{2.1}$$

Technically one needs $2m > d$ to make the space $\tilde{\mathscr{H}} = \{f\colon J(f) < \infty\}$ a RKHS. Note that $J_m^d(f)$ are rotation invariant, so thin plate splines are rotation invariant. The null space of J without side condition is $J_\perp^* = \{g\colon J(g) = 0\} = \{x_1^{\alpha_1} \cdots x_d^{\alpha_d}\}_{\alpha_1, \ldots, \alpha_d = 0}^{\alpha_1 + \cdots + \alpha_d < m}$ of dimension $M = \binom{d+m-1}{d}$.

A convenient side condition is $Ag = \sum_{s \in \mathscr{S}} g(s)/\#(\mathscr{S}) = 0$, where $\mathscr{S} \subset R^d$ is a collection of finite number of points called a normalizing mesh. Accordingly, a basis $\{\phi_\nu\}_{\nu=1}^M$ for $J_\perp^*$ can be constructed such that $\phi_1 = 1$ and $A(\phi_\nu \phi_\mu) = \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta. Let $\|g\|_\perp^* = \sum_{\nu=1}^M (A(g\phi_\nu))^2$ be the norm in $J_\perp^*$ and let $Pg = \sum_{\nu=1}^M (A(g\phi_\nu))\phi_\nu$. It can be shown that $J(f)$ is a norm in $\{g\colon \|g\|_\perp^* = 0,\ J(f) < \infty\}$ with the associated RK $R_J(\mathbf{x}, \mathbf{y}) = (I-P)_{\mathbf{x}}(I-P)_{\mathbf{y}} E_m^d(|\mathbf{x} - \mathbf{y}|)$, where I is the identity operator and $(I-P)_{\mathbf{x}}$ means $(I-P)$ applying on what follows as a function of $\mathbf{x}$, $|\cdot|$ is the Euclidean norm in R^d, and $E_m^d(\cdot)$ are known functions whose expressions can be found in, for example, Wahba and Wendelberger (1980). This construction works in setups more general than thin plate splines; see, for example, Wahba (1978) and Gu and Wahba (1993) for more details. Dropping $\phi_1 = 1$ from $J_\perp^*$ one gets $J_\perp$ complying with the side condition, and $\mathscr{H} = J_\perp \oplus \{g\colon \|g\|_\perp^* = 0,\ J(f) < \infty\}$ restricted to a bounded domain $\mathscr{X} \supset \mathscr{S}$ is an appropriate RKHS for the purpose of this article.

A few remarks follow. Note again that although the estimated log likelihood g varies with the normalizing mesh $\mathscr{S}$, the estimated density $e^g/\int e^g$ remains the same. The equivalence of J_m^d and R_J only holds on the domain $\mathscr{X} = R^d$. On a bounded domain $\mathscr{X} \subset R^d$, R_J as constructed above still defines a RKHS, but the corresponding J no longer has a nice explicit expression like (2.1). The essential element here is a positive definite function $R(\mathbf{x}, \mathbf{y})$ complying with a side condition $A_{\mathbf{y}} R(\mathbf{x}, \mathbf{y}) = 0$, where A is an averaging operator. Such an R

defines a norm J in span$\{R(\mathbf{x}, \cdot), \mathbf{x} \in \mathscr{X}\}$. This space can be further augmented by a finite dimensional Hilbert space $J_\perp$, $J_\perp \cap \text{span}\{R(\mathbf{x}, \cdot), \mathbf{x} \in \mathscr{X}\} = \{0\}$, whose members also comply with the side condition $Ag = 0$. Thin plate splines belong to a class that $R(\mathbf{x}, \mathbf{y}) = R(|\mathbf{x} - \mathbf{y}|)$ is radial and $J_\perp$ is invariant with respect to rotation and shift of $\mathscr{X}$ in R^d.

EXAMPLE 2.3 (Tensor product splines). The structure of tensor product splines depends on product RKHS's. The constructions of $\mathscr{H}$ and J for tensor product splines in the regression setup are well understood; details can be found in, for example, Gu and Wahba (1991a, b, 1993). With the previous two examples in mind where the drop-constant procedure is illustrated, the construction of $\mathscr{H}$ for log likelihood estimation should be straightforward following the lines in the literature. Here we only observe a few simple implications concerning their use in the current setup.

Consider a domain $\mathscr{X} = [0, 1]^3$. A tensor product spline on $\mathscr{X}$ can be decomposed as

$$g(x_1, x_2, x_3) = C + g_1(x_1) + g_2(x_2) + g_3(x_3) + g_{1,2}(x_1, x_2) + g_{1,3}(x_1, x_3) + g_{2,3}(x_2, x_3) + g_{1,2,3}(x_1, x_2, x_3),$$

where the C is a constant, the g_i are the main effects, the $g_{i,j}$ are the bivariate interactions, and the $g_{1,2,3}$ is the trivariate interaction. Accordingly, $J(g) = \sum_{I \subseteq \{1,2,3\}} J_I(g_I)$. Main effects and interactions satisfy side conditions $A_i g_i = A_i g_{i,j} = A_i g_{1,2,3} = 0$, where A_i are averaging operators on the axes. For log likelihood estimation one sets $C = 0$. The remaining seven components can all be included or excluded separately, resulting in 2^7 possible models of different complexities. The main-effect-only model, also called an additive model, implies the independence of the three coordinates. It is reassuring to see that the main-effect-only model fitted via minimizing (1.1) is equivalent to solutions of three separate problems with $J = J_i$ on each axis. Some less trivial probability structures may also be built in via selective inclusion of the terms in a tensor product spline. For example, the conditional independence of x_1 and x_2 given x_3 may be incorporated by excluding $g_{1,2}$ and $g_{1,2,3}$ from the model; a bit more discussion can be found in Gu [(1993), Section 7]. See, for example, Whittaker (1990) for a general discussion.

3. Poisson intensity and Silverman's estimator. We shall briefly note the estimation of Poisson intensity via density estimation in this section, and as a byproduct, show that our estimator as defined in Section 2 agrees with Silverman's when $J(g)$ annihilates constant.

Observing N occurrences X_i, $i = 1, \ldots, N$, from a Poisson counting process on $\mathscr{X}$ with an intensity function $\lambda(x)$, where $\lambda(x)$ is not to be confused with the smoothing parameter λ, the joint likelihood of N and X_i can be shown to be

$$\left(\prod_{i=1}^{N} \lambda(X_i)\right) \exp\left(-\int_{\mathscr{X}} \lambda(x)\right) = \left(\prod_{i=1}^{N} \lambda_0(X_i)\right) (\Lambda^N e^{-\Lambda}),$$

where $\Lambda = \int_{\mathscr{X}} \lambda(x)$ is the overall intensity of the process on $\mathscr{X}$ and $\lambda_0(x) = \lambda(x)/\Lambda$ is the occurrence density; see, for example, Snyder [(1975), Section 2.3]. N is statistically sufficient for Λ and has a Poisson distribution with mean Λ, and $X_i|N$ are conditionally i.i.d. with a probability density $\lambda_0(x)$. A penalized likelihood estimator of Poisson intensity may be defined as the minimizer of

$$(3.1) \qquad -\sum_{i=1}^{N} \log \lambda_0(X_i) - N \log \Lambda + \Lambda + J(\log \lambda_0(x) + \log \Lambda),$$

for $\log \lambda(x) \in \tilde{\mathscr{H}} \supset \{1\}$, where $\tilde{\mathscr{H}}$ is a general reproducing kernel Hilbert space and the smoothing parameter is absorbed into J to avoid confusion with the intensity $\lambda(x)$. Decompose $\tilde{\mathscr{H}} = \{1\} \oplus \mathscr{H}$, where $\mathscr{H}$ carries a side condition, and write $\log \lambda(x) = C + g$, where C is a constant and $g \in \mathscr{H}$. Noting that $\log \lambda_0 = g - \log \int_{\mathscr{X}} e^g$ and $\log \Lambda = C + \log \int_{\mathscr{X}} e^g$, (3.1) can be written as

$$(3.2) \qquad \begin{aligned} &\left[-\sum_{i=1}^{N} g(X_i) + N \log \int_{\mathscr{X}} e^g + J(C + g)\right] \\ &\quad + \left[-N\left(C + \log \int_{\mathscr{X}} e^g\right) + \exp\left(C + \log \int_{\mathscr{X}} e^g\right)\right]. \end{aligned}$$

Naturally J should annihilate constant since smoothing should only apply to the occurrence density, so $J(C + g) = J(g)$. The minimization of (3.2) can then be achieved in two steps, first to minimize the sum in the first pair of square brackets in (3.2) with respect to $g \in \mathscr{H}$ to estimate the occurrence density $\lambda_0(x)$, then to minimize the sum in the second pair of square brackets with respect to C to estimate the overall intensity Λ. The former is a smoothing spline density estimation based on data X_i, $i = 1, \ldots, N$, and the latter is a Poisson density estimation based on a single sample N.

When J annihilates constant, the two-step estimation in (3.2) may be manipulated to enforce arbitrary positive value on Λ by modifying the second part accordingly. Specifically, replacing $-N \log \Lambda + \Lambda$ in (3.2) by $-N \log \Lambda + N\Lambda$ and dividing the whole thing by N result in Silverman's estimator, which is our estimator multiplied by $\Lambda = 1$. Were a probability density defined to integrate to two, Silverman (1982) might have used $\int_{\mathscr{X}} e^g/2$ instead of $\int_{\mathscr{X}} e^g$ to augment (1.1) to enforce the "unity" constraint.

4. The existence of the estimator. In this section, we shall prove the following theorem, which guarantees the existence of the estimator. Without loss of generality, $\lambda = 2$ is assumed in this section.

THEOREM 4.1. *Suppose $L(g)$ is a continuous and strictly convex functional in a Hilbert space $\mathscr{H} = J_\perp \oplus \mathscr{H}_J$, where $\mathscr{H}_J$ has a square norm $J(g)$ and $J_\perp$ is the null space of $J(g)$ of finite dimension. If $L(g)$ has a minimizer in $J_\perp$, then $L(g) + J(g)$ has a unique minimizer in $\mathscr{H}$.*

A referee suggests that the theorem is likely to be known in the optimization literature, but we could not find it in the references we had our hands on so a proof is provided here. The proof builds on the following two lemmas. $L(g)$ and $J(g)$ below are the same as in Theorem 4.1.

LEMMA 4.1. *If a continuous strictly convex functional $A(g)$ has a minimizer in $\mathcal{J}_\perp$, then it has a minimizer in the cylinder area $C_\rho = \{g: g \in \mathcal{H}, J(g) \le \rho\}$, $\forall\, \rho \in (0, \infty)$.*

LEMMA 4.2. *If $L(g) + J(g)$ has a minimizer in $C_\rho = \{g: g \in \mathcal{H}, J(g) \le \rho\}$, $\forall\, \rho \in (0, \infty)$, then it has a minimizer in $\mathcal{H}$.*

The rest of the section are the proofs.

PROOF OF LEMMA 4.1. Let $\|\cdot\|_\perp$ be the norm in $\mathcal{J}_\perp$, and g_0 be the minimizer of $A(g)$ in $\mathcal{J}_\perp$. By Theorem 4 of Tapia and Thompson [(1978), page 162], $A(g)$ has a minimizer in a "rectangle" $R_{\rho,\gamma} = \{g: g \in \mathcal{H}, J(g) \le \rho, \|g - g_0\|_\perp \le \gamma\}$. Now if the lemma is not true, that is, that $A(g)$ has no minimizer in C_ρ for some ρ, then the minimizer g_γ of $A(g)$ in $R_{\rho,\gamma}$ should satisfy $\|g_\gamma - g_0\|_\perp = \gamma$. By the convexity of $A(g)$ and that $A(g_\gamma) < A(g_0)$,

$$A\big(\alpha g_\gamma + (1-\alpha)g_0\big) < \alpha A(g_\gamma) + (1-\alpha)A(g_0) < A(g_0) \tag{4.1}$$

for $\alpha \in (0,1)$. Now take a sequence $\gamma_i \to \infty$ and take $\alpha_i = \gamma_i^{-1}$, and write $\alpha_i g_{\gamma_i} + (1-\alpha_i)g_0 = g_i^0 + g_i^*$ where $g_i^0 \in \mathcal{J}_\perp$ and $g_i^* \in \mathcal{H} \ominus \mathcal{J}_\perp$. It is easy to check that $\|g_i^0 - g_0\|_\perp = 1$ and $J(g_i^*) \le \alpha_i^2 \rho$. Since $\mathcal{J}_\perp$ is finite dimensional, $\{g_i^0\}$ has a convergent subsequence converging to, say, $g_1 \in \mathcal{J}_\perp$, and $\|g_1 - g_0\|_\perp = 1$. It is apparent that $g_i^* \to 0$. By the continuity of $A(g)$ and (4.1), $A(g_1) \le A(g_0)$, which contradicts the fact that g_0 is the unique minimizer of $A(g)$ in $\mathcal{J}_\perp$. Hence, $\|g_\gamma - g_0\|_\perp = \gamma$ cannot be true for all $\gamma \in (0, \infty)$. This completes the proof. □

PROOF OF LEMMA 4.2. Without loss of generality we assume $L(0) = 0$. If the lemma is not true, then the minimizer g_ρ of $L(g) + J(g)$ in C_ρ must fall on the boundary of C_ρ for every ρ, that is, $J(g_\rho) = \rho$, $\forall\, \rho \in (0, \infty)$. By the convexity of $L(g)$,

$$L(\alpha g_\rho) < \alpha L(g_\rho), \tag{4.2}$$

for $\alpha \in (0,1)$. By the definition of g_ρ,

$$L(g_\rho) + J(g_\rho) < L(\alpha g_\rho) + J(\alpha g_\rho). \tag{4.3}$$

By combining (4.2) and (4.3) and substituting $J(g_\rho) = \rho$, one obtains

$$L(\alpha g_\rho)/\alpha + \rho < L(\alpha g_\rho) + \alpha^2 \rho,$$

which after some algebra leads to

$$L(\alpha g_\rho) < -\alpha(1+\alpha)\rho. \tag{4.4}$$

Now choose $\alpha = \rho^{-1/2}$. Since $J(\alpha g_\rho) = 1$, (4.4) leads to $L(g_1) < -(\rho^{1/2} + 1)$, which is impossible for large enough ρ. This proves the lemma. □

PROOF OF THEOREM 4.1. Applying Lemma 4.1 on $A(g) = L(g) + J(g)$ leads to the condition of Lemma 4.2, and Lemma 4.2 in turn yields the theorem. □

5. The convergence of the estimator. We shall establish the asymptotic convergence rates of the estimator in this section. Let g_0 be the true log likelihood and let $\hat{g}$ be the minimizer of (1.1) in $\mathscr{H}$. Throughout this section, it is assumed that $J(g_0) < \infty$ and that $L(g)$ has a minimizer in $J_\perp$. Denote by $J(g, h)$ the (semi) inner product associated with the square (semi) norm $J(g)$, write $\mu_g(h)$ as the mean of $h(X)$ where X has a log likelihood g. We intend to show that $\mu_{g_0}(g_0 - \hat{g}) + \mu_{\hat{g}}(\hat{g} - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, where the left-hand side is the symmetrized Kullback–Leibler distance between the truth and the estimator, and the r in the right-hand side codes the smoothness implied by $J(g)$ (cf. Assumption A.2). First we show that

$$\mu_{g_0}(g_0 - \hat{g}) + \mu_{\hat{g}}(\hat{g} - g_0) = \lambda J(\hat{g}, g_0 - \hat{g}) + \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{g} - g_0)(X_i) - \mu_{g_0}(\hat{g} - g_0)\right] \tag{5.1}$$

and then use a quadratic approximation device of Silverman (1982) and Cox and O'Sullivan (1990) to bound the right-hand side of (5.1).

Define $A_{g,h}(\alpha) = L(g + \alpha h) + (\lambda/2)J(g + \alpha h)$.

LEMMA 5.1. $\dot{A}_{g,h}(0) = -(1/n)\sum_{i=1}^{n} h(X_i) + \mu_g(h) + \lambda J(g, h)$.

The proof of the lemma is straightforward. Set $g = \hat{g}$ and $h = \hat{g} - g_0$ in Lemma 5.1. Note that $\dot{A}_{\hat{g},h}(0) = 0$. Equation (5.1) follows.

Following Silverman (1982) and Cox and O'Sullivan (1990), we first introduce an approximation of $\hat{g}$, say g_1, which is the minimizer of

$$L_1(g) + \left(\frac{\lambda}{2}\right)J(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \mu_{g_0}(g) + \left(\frac{1}{2}\right)V(g - g_0) + \left(\frac{\lambda}{2}\right)J(g), \tag{5.2}$$

where $V(h) = V_{g_0}(h)$ and $V_g(h)$ is the variance of $h(X)$ where X has a log likelihood g. $L_1(g)$ is basically the quadratic approximation of $L(g)$ at g_0. It can be seen that $V(h)$ is a norm in $\mathscr{H}$. We also write $V(g, h)$ as the inner product associated with the norm $V(g)$, the covariance of $g(X)$ and $h(X)$.

A bilinear form B is said to be completely continuous with respect to another bilinear form A if for any $\varepsilon > 0$, there exist finite number of linear functionals $l_1, \ldots, l_k$ such that $l_j(g) = 0$, $\forall\ j = 1, \ldots, k$, implies $B(g) \le \varepsilon A(g)$; see Weinberger [(1974), Section 3.3]. Note that to avoid interpolation,

the smoothness penalty J must restrict the solution of $L(g) + \lambda J(g)$ to an effectively finite dimensional space, and as $n \to \infty$, such restriction may be gradually relaxed by letting $\lambda \to 0$.

ASSUMPTION A.1. V is completely continuous with respect to $V + J$.

Under A.1, by Theorem 3.1 of Weinberger [(1974), page 52], there exists a sequence of eigenvalues λ_ν and the associated eigenfunctions ψ_ν of V with respect to $V + J$ such that

$$V(\psi_\nu, \psi_\mu) = \lambda_\nu \delta_{\nu,\mu} \quad \text{and} \quad (V + J)(\psi_\nu, \psi_\mu) = \delta_{\nu,\mu},$$

where $\delta_{\nu,\mu}$ is the Kronecker delta and $1 \geq \lambda_\nu \to 0$. Defining $\phi_\nu = \lambda_\nu^{-1/2}\psi_\nu$, it follows that

$$V(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu} \quad \text{and} \quad J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu},$$

where $0 \leq \rho_\nu = \lambda_\nu^{-1} - 1 \to \infty$. The smoothness of the functions in space $\mathscr{H}$ can be characterized by the rate of decay of λ_ν, and the convergence rates of the estimator depend on this rate of decay.

ASSUMPTION A.2. $\rho_\nu = c_\nu \nu^r$, where $r > 1$ and $c_\nu \in (\beta_1, \beta_2) \subset (0, \infty)$.

In Example 2.1, if g_0 is bounded from above and below, then Assumption A.2 is satisfied with $r = 4$; see Silverman [(1982), page 802].

Denote $g = \Sigma_\nu g_\nu \phi_\nu$ and $g_0 = \Sigma_\nu g_{\nu,0}\phi_\nu$, where $g_\nu = V(g, \phi_\nu)$ are the Fourier coefficients of g with basis ϕ_ν. Equation (5.2) becomes

$$\begin{aligned} &- \sum_\nu g_\nu \left[\frac{1}{n} \sum_{i=1}^n \phi_\nu(X_i) - \mu_{g_0}(\phi_\nu) \right] \\ &\quad + \left(\frac{1}{2}\right) \sum_\nu (g_\nu - g_{\nu,0})^2 + \left(\frac{\lambda}{2}\right) \sum_\nu \rho_\nu g_\nu^2. \end{aligned} \tag{5.3}$$

Writing $\beta_\nu = (1/n)\Sigma_{i=1}^n \phi_\nu(X_i) - \mu_{g_0}(\phi_\nu)$ and minimizing (5.3) with respect to g_ν, it follows that

$$g_{\nu,1} = (\beta_\nu + g_{\nu,0})/(1 + \lambda\rho_\nu).$$

Now

$$V(g_1 - g_0) = \sum_\nu (g_{\nu,1} - g_{\nu,0})^2 = \sum_\nu \frac{\beta_\nu^2 - 2\beta_\nu \lambda \rho_\nu g_{\nu,0} + \lambda^2 \rho_\nu^2 g_{\nu,0}^2}{(1 + \lambda\rho_\nu)^2},$$

$$\lambda J(g_1 - g_0) = \lambda \sum_\nu \rho_\nu (g_{\nu,1} - g_{\nu,0})^2 = \sum_\nu \lambda\rho_\nu \frac{\beta_\nu^2 - 2\beta_\nu \lambda \rho_\nu g_{\nu,0} + \lambda^2 \rho_\nu^2 g_{\nu,0}^2}{(1 + \lambda\rho_\nu)^2}.$$

Note that $E(\beta_\nu) = 0$ and $E(\beta_\nu^2) = n^{-1}$, one gets

$$E(V(g_1 - g_0)) = n^{-1}\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2} + \lambda\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}\rho_\nu g_{\nu,0}^2,$$

(5.4)

$$E(\lambda J(g_1 - g_0)) = n^{-1}\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} + \lambda\sum_\nu \frac{(\lambda\rho_\nu)^2}{(1+\lambda\rho_\nu)^2}\rho_\nu g_{\nu,0}^2.$$

LEMMA 5.2. *Under Assumption A.2, as* $\lambda \to 0$,

$$\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} = O(\lambda^{-1/r}),$$

$$\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2} = O(\lambda^{-1/r}),$$

$$\sum_\nu \frac{1}{1+\lambda\rho_\nu} = O(\lambda^{-1/r}).$$

PROOF. We only prove the first equation. The other two are similar:

$$\begin{aligned}\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} &= \left(\sum_{\nu<\lambda^{-1/r}} + \sum_{\nu\ge\lambda^{-1/r}}\right)\frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}\\ &= O(\lambda^{-1/r}) + O\left(\int_{\lambda^{-1/r}}^\infty \frac{\lambda x^r}{(1+\lambda x^r)^2}\,dx\right)\\ &= O(\lambda^{-1/r}) + \lambda^{-1/r}O\left(\int_1^\infty \frac{x^r}{(1+x^r)^2}dx\right)\\ &= O(\lambda^{-1/r}).\end{aligned}$$

□

THEOREM 5.1. *Under Assumptions A.1 and A.2, as* $n \to \infty$ *and* $\lambda \to 0$,

$$E(V(g_1 - g_0)) = O(n^{-1}\lambda^{-/r} + \lambda),$$

$$E(\lambda J(g_1 - g_0)) = O(n^{-1}\lambda^{-1/r} + \lambda).$$

PROOF. Note that $\Sigma_\nu \rho_\nu g_{\nu,0}^2 = J(g_0) < \infty$. The theorem follows from (5.4) and Lemma 5.2. □

We now turn to the approximation error $\hat{g} - g_1$. Set $g = \hat{g}$ and $h = \hat{g} - g_1$ in Lemma 5.1, it follows that

$$-\frac{1}{n}\sum_{i=1}^n (\hat{g} - g_1)(X_i) + \mu_{\hat{g}}(\hat{g} - g_1) + \lambda J(\hat{g}, \hat{g} - g_1) = 0. \tag{5.5}$$

Defining $B_{g,h}(\alpha) = L_1(g + \alpha h) + (\lambda/2)J(g + \alpha h)$, similarly one gets

$$-\frac{1}{n}\sum_{i=1}^{n}(\hat{g} - g_1)(X_i) + \mu_{g_0}(\hat{g} - g_1) + V(g_1 - g_0, \hat{g} - g_1) + \lambda J(g_1, \hat{g} - g_1) = \dot{B}_{g_1, \hat{g}-g_1}(0) = 0. \tag{5.6}$$

Combining (5.5) and (5.6) yields

$$\begin{aligned}\mu_{\hat{g}}(\hat{g} - g_1) - \mu_{g_1}(\hat{g} - g_1) + \lambda J(\hat{g} - g_1) &= V(g_1 - g_0, \hat{g} - g_1) \\ &\quad + \mu_{g_0}(\hat{g} - g_1) - \mu_{g_1}(\hat{g} - g_1).\end{aligned} \tag{5.7}$$

Now define $C(\alpha) = \mu_{g_0+\alpha(g_1-g_0)/\sigma}(\hat{g} - g_1) - \mu_{g_0}(\hat{g} - g_1)$, where $\sigma = V^{1/2}(g_1 - g_0) = o_p(1)$. A Taylor expansion gives $C(\alpha) = \alpha(1 + o(1))V(g_1 - g_0, \hat{g} - g_1)/\sigma$, where $o(1)$ is with respect to $\alpha \to 0$. This results in

$$\mu_{g_1}(\hat{g} - g_1) - \mu_{g_0}(\hat{g} - g_1) = C(\sigma) = V(g_1 - g_0, \hat{g} - g_1)\big(1 + o_p(1)\big), \tag{5.8}$$

as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$. Now define $D(\alpha) = \mu_{g_1+\alpha(\hat{g}-g_1)}(\hat{g} - g_1)$. It can be shown that $\dot{D}(\alpha) = V_{g_1+\alpha(\hat{g}-g_1)}(\hat{g} - g_1)$. By the mean value theorem,

$$\begin{aligned}\mu_{\hat{g}}(\hat{g} - g_1) - \mu_{g_1}(\hat{g} - g_1) &= D(1) - D(0) \\ &= \dot{D}(\alpha) \\ &= V_{g_1+\alpha(\hat{g}-g_1)}(\hat{g} - g_1),\end{aligned}$$

where $0 \le \alpha \le 1$.

ASSUMPTION A.3. For g in a convex set B_0 around g_0 containing $\hat{g}$ and g_1,

$$\exists\, c_1 \in (0, \infty) \text{ such that } c_1 V \le V_g \text{ uniformly.}$$

Assumption A.3 is satisfied when the members of B_0 have uniform upper and lower bounds on $\mathscr{X}$.

THEOREM 5.2. *Under Assumptions* A.1–A.3, *as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$,

$$\begin{aligned}V(\hat{g} - g_1) &= o_p(n^{-1}\lambda^{-1/r} + \lambda), \\ \lambda J(\hat{g} - g_1) &= o_p(n^{-1}\lambda^{-1/r} + \lambda).\end{aligned}$$

Consequently,

$$\begin{aligned}V(\hat{g} - g_0) &= O_p(n^{-1}\lambda^{-1/r} + \lambda), \\ \lambda J(\hat{g} - g_0) &= O_p(n^{-1}\lambda^{-1/r} + \lambda).\end{aligned}$$

PROOF. From (5.7), (5.8) and Assumption A.3,

$$c_1 V(\hat{g} - g_1) + \lambda J(\hat{g} - g_1) \le o_p(V(g_1 - g_0, \hat{g} - g_1)) = o_p(V^{1/2}(\hat{g} - g_1)V^{1/2}(g_1 - g_0)).$$

The theorem follows from Theorem 5.1 after trivial manipulation. □

Now we are ready to state and prove the main convergence result.

THEOREM 5.3. *Under Assumptions A.1–A.3, as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$,

$$\mu_{g_0}(g_0 - \hat{g}) + \mu_{\hat{g}}(\hat{g} - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda).$$

PROOF. Since $\lambda J(\hat{g}, g_0 - \hat{g}) \le (\lambda J(\hat{g}))^{1/2}(\lambda J(\hat{g} - g_0))^{1/2}$ and $\lambda J(\hat{g}) \le \lambda J(g_0) + \lambda J(\hat{g} - g_0)$, by Theorem 5.2, the first term of the right-hand side of (5.1) is of order $O_p(n^{-1}\lambda^{-1/r} + \lambda)$. For the second term, write

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g} - g_0)(X_i) - \mu_{g_0}(\hat{g} - g_0) = \sum_{\nu}(\hat{g}_\nu - g_{\nu,0})\beta_\nu,$$

where $\hat{g}_\nu$ are the Fourier coefficients of $\hat{g}$. By Cauchy–Schwarz,

$$\sum_{\nu}|(\hat{g}_\nu - g_{\nu,0})\beta_\nu| \le \left(\sum_{\nu}\alpha_\nu^2(\hat{g}_\nu - g_{\nu,0})^2\right)^{1/2}\left(\sum_{\nu}\alpha_\nu^{-2}\beta_\nu^2\right)^{1/2}, \tag{5.9}$$

for some sequence α_ν. Taking $\alpha_\nu^2 = 1 + \lambda\rho_\nu$, $\sum_\nu(1 + \lambda\rho_\nu)(\hat{g}_\nu - g_{\nu,0})^2 = (V + \lambda J)(\hat{g} - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ by Theorem 5.2, and $E(\sum_\nu(1 + \lambda\rho_\nu)^{-1}\beta_\nu^2) = O(n^{-1}\lambda^{-1/r})$ by Lemma 5.2 and the fact that $E(\beta_\nu^2) = n^{-1}$. Substituting these in (5.9) leads to

$$\sum_{\nu}|(\hat{g}_\nu - g_{\nu,0})\beta_\nu| = O_p(n^{-1}\lambda^{-1/r} + n^{-1/2}\lambda^{-1/2r+1/2}). \tag{5.10}$$

Combining this with the bound for the first term yields the theorem. □

6. The semiparametric approximation of the estimator. The minimizer $\hat{g}$ of (1.1) in $\mathscr{H}$ is not computable. For practical applications, one must find an appropriate finite dimensional approximating space and calculate an estimate in the approximating space. Let $\mathscr{H}_n = J_\perp \oplus \text{span}\{R_J(X_i, \cdot),\ i = 1, \ldots, n\}$, where R_J is the RK in $\mathscr{H} \ominus J_\perp$ associated with J. We shall prove in this section that $\mathscr{H}_n$ is an appropriate approximating space under the following extra condition.

ASSUMPTION A.4. $\exists\ c_2 \in (0, \infty)$ such that $V(\phi_\nu\phi_\mu) \le c_2$ uniformly for all ν and μ.

That is, we show that if $\hat{g}_n \in \mathscr{H}_n$ minimizes (1.1) then it is as good an estimator as $\hat{g}$ in the sense that $\mu_{g_0}(g_0 - \hat{g}_n) + \mu_{\hat{g}_n}(\hat{g}_n - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

LEMMA 6.1. *Under Assumptions A.1, A.2 and A.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(h) = O_p(\lambda J(h))$, $\forall\ h \in \mathscr{H} \ominus \mathscr{H}_n$.

Note that with an optimal smoothing parameter $\lambda = O(n^{-r/(1+r)})$, $n\lambda^{2/r} = O(n^{1-2/(1+r)}) \to \infty$ as $n \to \infty$.

PROOF. Note that $h \perp J_\perp$ and $h(X_i) = J(R_J(X_i, \cdot\ ), h) = 0$, so $\sum_{i=1}^n h^2(X_i) = 0$. Write $h = \sum_\nu h_\nu \phi_\nu$. It follows that

$$\begin{aligned} V(h) \le \mu_{g_0}(h^2) &= \sum_\nu \sum_\mu h_\nu h_\mu \mu_{g_0}(\phi_\nu \phi_\mu) \\ &= \sum_\nu \sum_\mu h_\nu h_\mu \left(\mu_{g_0}(\phi_\nu \phi_\mu) - \frac{1}{n} \sum_{i=1}^n \phi_\nu(X_i)\phi_\mu(X_i) \right) \\ &\le \left(\sum_\nu \sum_\mu (1+\lambda\rho_\nu)^{-1}(1+\lambda\rho_\mu)^{-1} \right. \\ &\qquad \left. \times \left(\frac{1}{n} \sum_{i=1}^n \phi_\nu(X_i)\phi_\mu(X_i) - \mu_{g_0}(\phi_\nu\phi_\mu) \right)^2 \right)^{1/2} \\ &\qquad \times \left(\sum_\nu \sum_\mu (1+\lambda\rho_\nu)(1+\lambda\rho_\mu) h_\nu^2 h_\mu^2 \right)^{1/2} \\ &= O_p(n^{-1/2}\lambda^{-1/r})(V+\lambda J)(h), \end{aligned}$$

where Lemma 5.2 and

$$E\left(\frac{1}{n} \sum_{i=1}^n \phi_\nu(X_i)\phi_\mu(X_i) - \mu_{g_0}(\phi_\nu\phi_\mu) \right)^2 \le c_2 n^{-1}$$

are used. The lemma follows. □

Let g_n be the projection of $\hat{g}$ in $\mathscr{H}_n$. $\hat{g} - g_n \in \mathscr{H} \ominus \mathscr{H}_n$.

THEOREM 6.1. *Under Assumptions A.1–A.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$V(\hat{g} - g_n) = O_p(n^{-1}\lambda^{-1/r}),$$

$$\lambda J(\hat{g} - g_n) = O_p(n^{-1}\lambda^{-1/r}).$$

PROOF. Set $g = \hat{g}$ and $h = \hat{g} - g_n$ in Lemma 5.1. Note that $(\hat{g} - g_n)(X_i) = 0$ and $J(g_n, \hat{g} - g_n) = 0$. It follows that

$$\mu_{\hat{g}}(\hat{g} - g_n) + \lambda J(\hat{g} - g_n) = 0.$$

By the continuity of μ and the fact that $\hat{g} \to g_0$, $\mu_{\hat{g}}(\hat{g} - g_n) = \mu_{g_0}(\hat{g} - g_n)(1 + o_p(1))$. Similar to (5.10), it can be shown that

$$\mu_{g_0}(\hat{g} - g_n) = O_p(n^{-1/2}\lambda^{-1/2r})\big(V^{1/2} + (\lambda J)^{1/2}\big)(\hat{g} - g_n). \tag{6.1}$$

So

$$\lambda J(\hat{g} - g_n) = O_p(n^{-1/2}\lambda^{-1/2r})\big(V^{1/2} + (\lambda J)^{1/2}\big)(\hat{g} - g_n).$$

This and Lemma 6.1 yield the theorem. □

Let $\hat{g}_n$ be the minimizer of (1.1) in $\mathscr{H}_n$, the semiparametric approximation we are aiming at. Note that our existence result also applies to $\mathscr{H}_n$, so $\hat{g}_n$ exists.

THEOREM 6.2. *Assume $\hat{g}_n$ and g_n also belong to the convex set B_0 in Assumption A.3. Under Assumptions A.1–A.4, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,*

$$V(\hat{g}_n - g_n) = O_p(n^{-1}\lambda^{-1/r}),$$
$$\lambda J(\hat{g}_n - g_n) = O_p(n^{-1}\lambda^{-1/r}).$$

PROOF. Set $g = \hat{g}_n$ and $h = \hat{g}_n - g_n \in \mathscr{H}_n$ in Lemma 5.1. It follows that

$$-\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n - g_n)(X_i) + \mu_{\hat{g}_n}(\hat{g}_n - g_n) + \lambda J(\hat{g}_n, \hat{g}_n - g_n) = 0. \tag{6.2}$$

Setting $g = \hat{g}$ and $h = \hat{g} - \hat{g}_n$ in Lemma 5.1, one gets

$$-\frac{1}{n}\sum_{i=1}^{n}(\hat{g} - \hat{g}_n)(X_i) + \mu_{\hat{g}}(\hat{g} - \hat{g}_n) + \lambda J\big(\hat{g}, \hat{g} - \hat{g}_n\big) = 0. \tag{6.3}$$

Note that $(\hat{g} - g_n)(X_i) = 0$ and $J(\hat{g} - g_n, g_n) = J(\hat{g} - g_n, \hat{g}_n) = 0$. Equation (6.3) leads to

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n}(g_n - \hat{g}_n)(X_i) + \mu_{\hat{g}}(\hat{g} - \hat{g}_n) \\ + \lambda J(\hat{g} - g_n) + \lambda J(g_n, g_n - \hat{g}_n) = 0. \end{aligned} \tag{6.4}$$

Adding (6.2) and (6.4), some algebra leads to

$$\begin{aligned} &\mu_{\hat{g}_n}(\hat{g}_n - g_n) - \mu_{g_n}(\hat{g}_n - g_n) + \lambda J(\hat{g}_n - g_n) + \lambda J(\hat{g} - g_n) \\ &\quad = \mu_{\hat{g}}(g_n - \hat{g}) + \mu_{\hat{g}}(\hat{g}_n - g_n) - \mu_{g_n}(\hat{g}_n - g_n). \end{aligned} \tag{6.5}$$

Now by Assumption A.3,

$$\mu_{\hat{g}_n}(\hat{g}_n - g_n) - \mu_{g_n}(\hat{g}_n - g_n) \geq c_1 V(\hat{g}_n - g_n).$$

By (6.1) and Theorem 6.1,

$$\mu_{\hat{g}}(g_n - \hat{g}) = \mu_{g_0}(g_n - \hat{g})\big(1 + o_p(1)\big) = O_p(n^{-1}\lambda^{-1/r})$$

and

$$\begin{aligned}\mu_{\hat{g}}(\hat{g}_n - g_n) - \mu_{g_n}(\hat{g}_n - g_n) &= V(\hat{g} - g_n, \hat{g}_n - g_n)\big(1 + o_p(1)\big)\\ &= O_p\big(V^{1/2}(\hat{g} - g_n)V^{1/2}(\hat{g}_n - g_n)\big)\\ &= O_p(n^{-1/2}\lambda^{-1/2r})V^{1/2}(\hat{g}_n - g_n).\end{aligned} \tag{6.6}$$

Hence,

$$\begin{aligned}&c_1 V(\hat{g}_n - g_n) + \lambda J(\hat{g}_n - g_n) + \lambda J(\hat{g} - g_n)\\ &\quad = O_p(n^{-1}\lambda^{-1/r}) + O_p(n^{-1/2}\lambda^{-1/2r})V^{1/2}(\hat{g}_n - g_n).\end{aligned}$$

The theorem follows immediately. □

We are now ready for the main result of this section.

THEOREM 6.3. *Under Assumptions A.1–A.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$\begin{aligned}V(\hat{g}_n - g_0) &= O_p(n^{-1}\lambda^{-1/r} + \lambda),\\ \lambda J(\hat{g}_n - g_0) &= O_p(n^{-1}\lambda^{-1/r} + \lambda).\end{aligned}$$

Consequently,

$$\mu_{g_0}(g_0 - \hat{g}_n) + \mu_{\hat{g}_n}(\hat{g}_n - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda).$$

PROOF. The first part of the theorem follows from Theorems 5.2, 6.1 and 6.2. To prove the second part, set $g = \hat{g}$ and $h = \hat{g}_n - g_0$ in Lemma 5.1. This yields

$$-\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n - g_0)(X_i) + \mu_{\hat{g}}(\hat{g}_n - g_0) + \lambda J(\hat{g}, \hat{g}_n - g_0) = 0.$$

Hence,

$$\begin{aligned}&\mu_{g_0}(g_0 - \hat{g}_n) + \mu_{\hat{g}_n}(\hat{g}_n - g_0)\\ &\quad = \mu_{g_0}(g_0 - \hat{g}_n) + \mu_{\hat{g}_n}(\hat{g}_n - g_0)\\ &\qquad + \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n - g_0)(X_i)\\ &\qquad - \mu_{\hat{g}}(\hat{g}_n - g_0) + \lambda J(\hat{g}, g_0 - \hat{g}_n)\\ &\quad = \lambda J(\hat{g}, g_0 - \hat{g}_n)\\ &\qquad + \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_n - g_0)(X_i) - \mu_{g_0}(\hat{g}_n - g_0)\right]\\ &\qquad + \big[\mu_{\hat{g}_n}(\hat{g}_n - g_0) - \mu_{\hat{g}}(\hat{g}_n - g_0)\big].\end{aligned}$$

The first two terms in the preceding expression can be shown to be of order $O_p(n^{-1}\lambda^{-1/r} + \lambda)$ by the same techniques used in proving Theorem 5.3.

Similar to (6.6), it is straightforward to show that the third term is of the same order. This completes the proof. □

7. Discussion. We offer a few remarks before concluding this article. It is apparent that the existence theorem of Section 4 is generally applicable to smoothing spline estimators in many contexts, and possibly also to other regularization problems. For the convergence results, we deviated from the conventional mean square error and instead concentrated on the symmetrized Kullback–Leibler, which might be a more natural criterion for the density estimation problem. We omitted the Gaussian process approximation discussed in Silverman [(1982), Section 9]; although the structure of the estimator is almost identical to that of Silverman's, we were not able to validate a version of Silverman's Lemma 5.5 because of the generic setup, note that derivatives are not even defined for $g(x) \in \mathscr{H}$ on $\mathscr{X}$. From a practical point of view, Theorem 6.3 is the most important result of this article, for it allows practical application of the proposed method, especially in multivariate setups where a general purpose basis does not exist. The semiparametric approximation is motivated by the semiparametric expression of the exact solution in the regression setup, which played a central role in the computation of multivariate smoothing spline regression models. A dimensionless algorithm for calculating $\hat{g}_n$ with a data-driven automatic λ has been developed in a further work by Gu (1993) with portable code available from chong@stat.purdue.edu. Yet further algorithmic developments are needed to speed up the algorithm in one dimension and to incorporate automatic multiple smoothing parameters in multidimension for the calculation of the tensor product spline estimators of Example 2.3.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

GU, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88**.

GU, C. and WAHBA, G. (1991a). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.* **12** 383–398.

GU, C. and WAHBA, G. (1991b). Discussion of "Multivariate adaptive regression splines" by J. Friedman. *Ann. Statist.* **19** 115–123.

GU, C. and WAHBA, G. (1993). Semiparametric ANOVA with tensor product thin plate splines. *J. Roy. Statist. Soc. Ser. B* **55**.

LEONARD, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.

SNYDER, D. L. (1975). *Random Point Processes*. Wiley, New York.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation*. Johns Hopkins Univ. Press.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

WAHBA, G. and WENDELBERGER, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review* **108** 1122–1145.

WEINBERGER, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. SIAM, Philadelphia.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907