The Annals of Statistics
1990, Vol. 18, No. 2, 717–741

LARGE-SAMPLE INFERENCE FOR LOG-SPLINE MODELS[1]

BY CHARLES J. STONE

University of California, Berkeley

Let f be a continuous and positive unknown density on a known compact interval $\mathscr{Y}$. Let F denote the distribution function of f and let $Q = F^{-1}$ denote its quantile function. A finite-parameter exponential family model based on B-splines is constructed. Maximum-likelihood estimation of the parameters of the model based on a random sample of size n from f yields estimates $\hat{f}$, $\hat{F}$ and $\hat{Q}$ of f, F and Q, respectively. Under mild conditions, if the number of parameters tends to infinity in a suitable manner as $n \to \infty$, these estimates achieve the optimal rate of convergence. The asymptotic behavior of the corresponding confidence bounds is also investigated. In particular, it is shown that the standard errors of $\hat{F}$ and $\hat{Q}$ are asymptotically equal to those of the usual empirical distribution function and empirical quantile function.

1. Introduction. Let Y be a random variable that ranges over a subinterval $\mathscr{Y}$ of $\mathbb{R}$ and has unknown density f on $\mathscr{Y}$. Let $Y_1, \ldots, Y_n$ be a random sample of size n from the distribution of Y. This random sample can be used for inference about the density of Y and other aspects of the distribution of Y.

The classical parametric approach is to start by "assuming" a fixed parametric model for f involving a J-dimensional vector $\boldsymbol{\theta} \in \Theta$ of unknown parameters. We write the resulting model as $f(\cdot\,; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. The maximum-likelihood estimate $\bar{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by maximizing the log-likelihood function

$$l(\boldsymbol{\theta}) = \sum_i \log(f(Y_i, \boldsymbol{\theta})), \qquad \boldsymbol{\theta} \in \Theta.$$

This estimate is especially attractive when the parametric model for f has the form of an exponential family in which $\boldsymbol{\theta}$ is a natural parameter.

To obtain such a form, choose a J-dimensional vector space $\mathscr{S}$ of functions on $\mathscr{Y}$ such that (for identifiability) the zero function on $\mathscr{Y}$ is the only function in $\mathscr{S}$ that equals a constant almost everywhere on $\mathscr{Y}$. Let $B_1, \ldots, B_J$ be a basis of $\mathscr{S}$. Given the column vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^t \in \mathbb{R}^J$, set

$$s(\cdot\,; \boldsymbol{\theta}) = \sum_1^J \theta_j B_j \quad \text{and} \quad c(\boldsymbol{\theta}) = \log\left[\int_{\mathscr{Y}} \exp(s(y; \boldsymbol{\theta}))\, dy\right].$$

Then $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^J\colon c(\boldsymbol{\theta}) < \infty\}$ is a convex subset of $\mathbb{R}^J$, which is assumed to be nonempty and open. The corresponding J-parameter exponential family is

Received September 1986; revised June 1989.
[1]Research supported in part by NSF grants MCS83-01257 and DMS-8600409.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62F12.
Key words and phrases. Functional inference, exponential families, B-splines, maximum likelihood, rates of convergence.

given by

$$f(\cdot;\boldsymbol{\theta}) = \exp(s(\cdot;\boldsymbol{\theta}) - c(\boldsymbol{\theta})), \qquad \boldsymbol{\theta} \in \Theta.$$

For $\boldsymbol{\theta} \in \Theta$, let $\mathbf{H}(\boldsymbol{\theta})$ denote the Hessian matrix of $c(\cdot)$ at $\boldsymbol{\theta}$, which is the symmetric $J \times J$ matrix having entry $\partial^2 c(\boldsymbol{\theta})/\partial\theta_j\,\partial\theta_k$ in row j and column k for $1 \le j, k \le J$. This matrix is positive definite, so $c(\cdot)$ is strictly convex on Θ. The log-likelihood function is given by

$$l(\boldsymbol{\theta}) = \sum_i [s(Y_i;\boldsymbol{\theta}) - c(\boldsymbol{\theta})], \qquad \boldsymbol{\theta} \in \Theta.$$

The Hessian matrix of the log-likelihood function at $\boldsymbol{\theta}$ is $-\mathbf{I}(\boldsymbol{\theta}) = -n\mathbf{H}(\boldsymbol{\theta})$, where the information matrix $\mathbf{I}(\boldsymbol{\theta})$ is positive definite and hence its inverse $(\mathbf{I}(\boldsymbol{\theta}))^{-1}$ is also positive definite. Since the log-likelihood function is strictly concave, the maximum-likelihood estimate is unique if it exists.

Let τ be a real-valued parameter depending on f. Under the assumption that f belongs to the indicated exponential family, $\tau = g(\boldsymbol{\theta})$ for some function g on Θ. The maximum-likelihood estimate of τ is given by $\hat{\tau} = g(\hat{\boldsymbol{\theta}})$. Suppose that g is continuously differentiable on Θ. Let $\nabla g(\boldsymbol{\theta})$ denote the gradient of g at $\boldsymbol{\theta}$, which is the J-dimensional column vector whose jth entry is $\partial g(\boldsymbol{\theta})/\partial\theta_j$. The asymptotic standard deviation (ASD) and standard error (SE) of $\hat{\tau}$ are the nonnegative quantities defined by

$$\mathrm{ASD}(\hat{\tau}) = \sqrt{(\nabla g(\boldsymbol{\theta}))^t (\mathbf{I}(\boldsymbol{\theta}))^{-1} \nabla g(\boldsymbol{\theta})}$$

and

$$\mathrm{SE}(\hat{\tau}) = \sqrt{(\nabla g(\hat{\boldsymbol{\theta}}))^t (\mathbf{I}(\hat{\boldsymbol{\theta}}))^{-1} \nabla g(\hat{\boldsymbol{\theta}})}\,.$$

Suppose that $\nabla g(\boldsymbol{\theta}) \neq \mathbf{0}$, where $\boldsymbol{\theta}$ is the true parameter. Then

$$\mathrm{dist}\left[\frac{\hat{\tau} - \tau}{\mathrm{ASD}(\hat{\tau})}\right] \approx N(0,1) \quad \text{and} \quad \mathrm{dist}\left[\frac{\hat{\tau} - \tau}{\mathrm{SE}(\hat{\tau})}\right] \approx N(0,1), \qquad n \gg 1.$$

Consequently, $\hat{\tau} \pm z_{1-0.5\alpha}\mathrm{SE}(\hat{\tau})$ is an asymptotic $100(1-\alpha)\%$ confidence interval for $\hat{\tau}$. Here $\Phi(z_p) = p$, Φ being the standard normal distribution function.

In the present approach, we do not assume a fixed finite-parameter model for f, but we can still make use of finite-parameter models as approximations. In order for the corresponding maximum-likelihood estimates and asymptotic confidence bounds for parameters τ defined in terms of f to be reliable, we need the modeling error to tend to zero as $n \to \infty$; for this, it is necessary that the number of parameters tend to infinity as $n \to \infty$. Since the field of statistics containing this approach involves a blend of parametric inference and nonparametric inference, we refer to it as *functional inference*.

A thorough study of the use of finite-parameter models as approximations requires a combination of mathematically rigorous asymptotics and computer simulation. In the present paper we concentrate on the asymptotic approach. To carry it out, we require that $\mathscr{Y}$ be compact and that f be continuous and positive on $\mathscr{Y}$. Also, we restrict our attention to function spaces $\mathscr{S}$ consisting of splines of fixed order q (piecewise polynomials of degree less than q), to

bases consisting of B-splines and (for simplicity) to equally spaced knots. The constant function 1 is in $\mathscr{S}$, so we impose the constraint $\Sigma\theta_j = 0$ and thereby end up with a $(J-1)$-parameter identifiable exponential family. Since $\log(f(\cdot;\boldsymbol{\theta})) \in \mathscr{S}$, we refer to this family as a *log-spline model*.

The main results of this paper and their motivation and application will now be described in an informal manner.

Let $\|\ \|_2$ and $\|\ \|_\infty$ denote the usual L_2 and L_∞ norms of functions on $\mathscr{Y}$. Set

$$\delta = \inf_{s \in \mathscr{S}} \|s - \log(f)\|_\infty.$$

As pointed out in Section 2, we know that $\delta = o(1)$ as $n \to \infty$ (under the supposition that $J \to \infty$ as $n \to \infty$). Under appropriate smoothness conditions on f, it is also known that $\delta = O(J^{-p})$, where the positive number p depends on the smoothness condition. In particular, if f is m-times continuously differentiable, where $1 \le m \le q$, then $\delta = O(J^{-m})$.

Set $f^* = f(\cdot;\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ maximizes the expected value of the log-likelihood function. It is shown in Stone (1989) that $\|f - f^*\|_\infty = O(\delta)$, the proof being based on a similar result of de Boor (1976) involving L_2 projections. Such results are surprising, since minimizing the L_2 error of approximation and maximizing the expected log-likelihood do not appear to be closely related to minimizing the L_∞ error of approximation.

The functional viewpoint suggests looking at parameters that are values of the density function at individual points or of the corresponding distribution function or quantile function. We refer to $\hat{f} = f(\cdot;\hat{\boldsymbol{\theta}})$ as the *log-spline density estimate* of f. As a special case, if $\mathscr{S}$ is a collection of piecewise-constant functions on $\mathscr{Y}$, then $\hat{f}$ is a histogram density estimate.

For technical reasons, we require that $J = o(n^{0.5-\varepsilon})$ for some $\varepsilon > 0$. This is slightly stronger than the assumption $J = o(n^{0.5})$ that arises in Portnoy (1986, 1988).

We will show that $\|\hat{f} - f^*\|_2 = O_P(\sqrt{J/n})$ and $\|\hat{f} - f^*\|_\infty = O_P(\sqrt{J\log(J)/n})$. (The result for $\|\hat{f} - f^*\|_2$ is plausible: There are about n/J trials per unknown parameter, so the asymptotic standard deviation of the estimate of each parameter should be proportional to $\sqrt{J/n}$.) Suppose that $\delta = O(J^{-p})$, where $p > 0.5$. Set $\gamma = 1/(2p+1)$ and $r = p/(2p+1)$. By choosing $J \sim n^\gamma$, we get that $\|\hat{f} - f\|_2 = O_P(n^{-r})$. (Here $a_n \sim b_n$ means that a_n/b_n is bounded away from zero and infinity.) By choosing $J \sim [n/\log(n)]^\gamma$, we get that $\|\hat{f} - f\|_\infty = O_P([\log(n)/n]^r)$. Under suitable specifications, these are the well-known optimal rates of convergence for nonparametric density estimation; see Stone (1980, 1982, 1983).

Let F, F^* and $\hat{F}$ denote the distribution functions of f, f^* and $\hat{f}$, respectively. We will show that $\|F^* - F\|_\infty = O(\delta/J)$ and $\|\hat{F} - F^*\|_\infty = O_P(1/\sqrt{n})$. Thus, provided that $\delta/J = O(1/\sqrt{n})$. we get that $\|\hat{F} - F\|_\infty = O_P(1/\sqrt{n})$, which is well known to be the optimal rate of convergence for estimation of an unknown distribution function in both parametric and nonparametric settings.

Let $\hat{F}^{\text{emp}}$ denote the empirical distribution function, defined as usual by

$$\hat{F}^{\text{emp}}(y) = \frac{1}{n}\#\{i: 1 \le i \le n \text{ and } Y_i \le y\}, \qquad y \in \mathbb{R}.$$

Then $\text{SD}(\hat{F}^{\text{emp}}(y)) = \sqrt{F(y)(1 - F(y)/n}$, $y \in \mathscr{Y}$. It is well known that, for $n \gg 1$,

$$\text{dist}\left[\frac{\hat{F}^{\text{emp}}(y) - F(y)}{\text{SD}(\hat{F}^{\text{emp}}(y))}\right] \approx N(0,1) \quad \text{uniformly on compact subsets of int}(\mathscr{Y}),$$

where $\text{int}(\mathscr{Y})$ denotes the interior of $\mathscr{Y}$. We will show that, for $n \gg 1$,

$$\text{dist}\left[\frac{\hat{F}(y) - F^*(y)}{\text{ASD}(\hat{F}(y))}\right] \approx N(0,1) \quad \text{uniformly over compact subsets of int}(\mathscr{Y})$$

and

$$\frac{\text{ASD}(\hat{F}(y))}{\text{SD}(\hat{F}^{\text{emp}}(y))} \approx 1 \quad \text{uniformly over compact subsets of int}(\mathscr{Y}).$$

According to these results, the performance of the parametric estimate $\hat{F}(y)$ is similar to that of the nonparametric estimate $\hat{F}^{\text{emp}}(y)$ when $n \gg 1$ (recall that $J \to \infty$ as $n \to \infty$). The results can be used to obtain asymptotic confidence bounds for $F^*(y)$ or $F(y)$ based on the log-spline model. They also lead us to conjecture that

$$\|\hat{F} - \hat{F}^{\text{emp}}\|_\infty = o_P(1/\sqrt{n}) \quad \text{if } \delta/J = o(1/\sqrt{n}),$$

which should not be hard to verify.

Let $Q = F^{-1}$, $Q^* = (F^*)^{-1}$ and $\hat{Q} = \hat{F}^{-1}$ denote the quantile functions corresponding to f, f^* and $\hat{f}$, respectively. We will show that

$$\|Q^* - Q\|_\infty = O(\delta/J) \quad \text{and} \quad \|\hat{Q} - Q^*\|_\infty = O_P(1/\sqrt{n}).$$

Thus, provided that $\delta/J = O(1/\sqrt{n})$, we get that $\|\hat{Q} - Q\|_\infty = O_P(1/\sqrt{n})$, which again is the optimal rate of convergence.

Let $\hat{Q}^{\text{emp}}$ denote the empirical quantile function as usually defined [for example, $\hat{Q}^{\text{emp}}(0.5)$ is the sample median]. It is well known that, for $n \gg 1$,

$$\text{dist}\left[\frac{\hat{Q}^{\text{emp}}(p) - Q(p)}{\text{ASD}(\hat{Q}^{\text{emp}}(p))}\right] \approx N(0,1) \quad \text{uniformly over compact subsets of } (0,1),$$

where

$$\text{ASD}(\hat{Q}^{\text{emp}}(p)) = \sqrt{\frac{p(1-p)}{nf^2(Q(p))}}.$$

We will show that, for $n \gg 1$,

$$\text{dist}\left[\frac{\hat{Q}(p) - Q^*(p)}{\text{ASD}(\hat{Q}(p))}\right] \approx N(0,1) \quad \text{uniformly over compact subsets of } (0,1)$$

and

$$\frac{\text{ASD}(\hat{Q}(p))}{\text{ASD}(\hat{Q}^{\text{emp}}(p))} \approx 1 \quad \text{uniformly over compact subsets of } (0,1).$$

These results can be used to obtain asymptotic confidence bounds for $Q^*(p)$ or $Q(p)$ based on the log-spline model.

In parallel with the analytic approach and in continuation of the work of Stone and Koo (1986a, b), Charles Kooperberg and I are currently using computer simulation to determine the finite-sample performance of inference based on log-spline models. An important advantage of the computational approach is that attractive but mathematically unwieldy modifications can be studied. In our investigation, we have focused on positive random variables Y [$\mathscr{Y} = (0, \infty)$]. By using a suitable data-dependent transformation, which behaves like a power transformation at infinity and a logarithmic transformation at the origin, we first transform Y to a real-valued random variable [$\mathscr{Y} = (-\infty, \infty)$]. So far, we have used cubic splines. The first knot has been placed at the minimum value in the random sample $Y_1, \ldots, Y_n$, the last knot has been placed at the maximum value and a number of intermediate knots have been placed at selected order statistics of the random sample. Also, the selection of the number of intermediate knots and the order statistics used in placing them has been made in a data-independent manner. To reduce the standard errors of log-spline estimates of extreme quantiles, we have imposed linear restrictions on the fitted splines to the left of the first knot and to the right of the last knot. Thus the estimated distribution of the transformed random variable has exponential tails.

We have focused on confidence bounds for quantiles, especially on the 90% upper confidence bound for $Q(p)$ with $p \approx 1$. As applied to the transformed random variables, such confidence bounds are of the form $\hat{Q}(p) + t\text{SE}(\hat{Q}(p))$, where the indicated standard error is obtained by ignoring the data dependence of the preliminary transformation and the knot selection. In order to achieve a satisfactory approximation to the desired coverage probability over a broad range of n, p and underlying distributions, have chosen t by adaption to the exponential distribution (using computer simulation as in the use of the bootstrap) instead of using the value $t = z_{0.9} \doteq 1.28$ suggested by normal approximation.

The results obtained to date confirm the necessity of having $J \to \infty$ as $n \to \infty$. The performance of the nominal 90% log-spline upper confidence bounds for extreme quantiles ($1 - p = 1/n$ or $0.1/n$ with $n = 100$ or 500) is satisfactory, although not quite as good as that of the best procedure that emerged in the fairly extensive computer simulation study by Breiman, Stone and Kooperberg (1989). As expected, if there is an order statistic that serves as

a 90% upper confidence bound for $Q(p)$ (which, for $n \gg 1$, can only happen if $1 - p > 2/n$), then the performance of the upper confidence bound based on the log-spline model is very similar to that of the order statistic.

Some advantages of inference based on log-spline models over that based on kernel density estimates are as follows: Log-spline density estimates are automatically positive and can be chosen to achieve the optimal rate of convergence n^{-r} for values of r that are arbitrarily close to 0.5, log-spline estimates of extreme quantiles are sensible, the corresponding upper confidence bounds are reliable, and standard inferential tools for treating exponential families are applicable. A disadvantage of log-spline models based on quadratic and higher-order splines is that the evaluation of $c(\theta)$ and its partial derivatives requires computationally intensive numerical integration.

The log-spline approach could probably be extended to handle multidimensional distributions by using the tensor product B-splines or certain function spaces arising in connection with the finite-element method. In three or more dimensions, however, the approach may be impractical.

In Stone (1985), the goal was to achieve the optimal rate of convergence for nonparametric estimation of the best additive approximation to the regression function; it was realized by using the least-squares method to fit an additive spline. The setup of the follow-on paper, Stone (1986), included additive logistic regression and other nonparametric extensions of generalized linear models [see McCullagh and Nelder (1983)]. There the goal was to achieve the optimal rate of convergence for nonparametric estimation of the best additive approximation to the response function. This goal was realized by using maximum likelihood to fit an additive spline. Some of the results in Stone (1985, 1986) and the present paper are summarized in Stone (1987).

It would be worth while to include the setup of Stone (1986) and that of the present paper in a common framework. Consider, in particular, a random variable Y whose density depends on a real-valued variable x. It seems reasonable to use a log-spline model $f(\cdot; \boldsymbol{\theta}_x)$ based on linear splines to approximate the unknown density, where the dependence of each entry of $\boldsymbol{\theta}_x$ on x is approximated by a cubic spline. Tensor-product B-splines arise naturally in this manner. Stone (1989) contains a start at the mathematical analysis of this approach [namely, the extension of Theorem 1(i) in the next section].

Log-spline models are flexible exponential families. Previously, Neyman (1937) and Crain (1974, 1976a, b, 1977) considered other such families. [For recent results in the spirit of the present paper, written after the original versions of Stone (1989) and the present paper but independently of these papers, see Barron and Sheu (1988).] Leonard (1978) and Silverman (1982) considered various nonparametric estimates of $\log(f)$, an attractive feature being that the corresponding estimate of f itself is automatically positive.

Precise results will be stated in the next section and proven thereafter. The proofs use refinements of the known analytic properties of splines that have been developed in Stone (1985, 1986, 1989). They also use refinements of standard techniques for determining the asymptotic behavior of maximum-likelihood estimates of the unknown parameters of an exponential family that

are applicable when the number of parameters tends to infinity as $n \to \infty$. Some of these refinements were developed in Stone (1986). Portnoy (1988) has also studied the asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. It might be interesting to compare Portnoy's approach to that of Stone (1986) and the present paper.

2. Statement of results. Let $Y_1, Y_2 \ldots$ be independent and identically distributed random variables taking on values in a known compact interval $\mathscr{Y}$ having positive length; without loss of generality, let $\mathscr{Y} = [0, 1]$. These random variables are assumed to have a continuous and positive density f on $\mathscr{Y}$. Let F denote the distribution function of f and let $Q = F^{-1}$ denote its quantile function. Given the positive integer n, we refer to $Y_1, \ldots, Y_n$ as the random sample of size n.

For simplicity in notation, we suppress the dependence on n of various quantities after these quantities are defined.

Let q denote a positive integer. Given $n \geq 1$, let $K = K_n$ denote a positive integer. Let $\mathscr{Y}$ be partitioned into subintervals

$$\mathscr{Y}_k = [(k-1)/K, k/K), 1 \leq k < K \quad \text{and} \quad \mathscr{Y}_K = [(K-1)/K, 1].$$

Let $\mathscr{S} = \mathscr{S}_n$ denote the collection of functions s on $\mathscr{Y}$ satisfying the following two properties: s is a polynomial of order q (degree less than q) on each of the subintervals $\mathscr{Y}_1, \ldots, \mathscr{Y}_K$; if $q \geq 2$, s is $(q-2)$-times continuously differentiable on $\mathscr{Y}$. Then $\mathscr{S}$ is a vector space of dimension $J = J_n = q + K - 1$, which is a space of polynomial splines of order q with simple knots at k/K for $1 \leq k < K$. The functions in $\mathscr{S}$ are piecewise-constant, linear, quadratic or cubic splines according as $q = 1, 2, 3$ or 4. Consider the usual B-spline basis $B_j = B_{nj}$, $1 \leq j \leq J$, of $\mathscr{S}$ [see de Boor (1978)]. These functions are nonnegative and sum to 1 on $\mathscr{Y}$. Also, there is a fixed positive integer J_0, depending on q but not on n, such that (i) the support of each B_j is contained in the convex hull of J_0 consecutive knots and (ii) if $|k - j| > J_0$, the supports of B_j and B_k are disjoint. (The support of B_j is the subset of $\mathscr{Y}$ on which $B_j > 0$.)

Let $\Theta = \Theta_n$ denote the collection of all J-dimensional (column) vectors. Given $\boldsymbol{\theta} \in \Theta$, set

$$|\boldsymbol{\theta}| = \sqrt{\sum_j \theta_j^2},$$

$$s(\cdot\,; \boldsymbol{\theta}) = s_n(\cdot\,; \boldsymbol{\theta}) = \sum_j \theta_j B_j,$$

$$c(\boldsymbol{\theta}) = c_n(\boldsymbol{\theta}) = \log\left[\int \exp(s(\cdot\,; \boldsymbol{\theta}))\right],$$

and

$$f(\cdot\,; \boldsymbol{\theta}) = f_n(\cdot\,; \boldsymbol{\theta}) = \exp(s(\cdot\,; \boldsymbol{\theta}) - c(\boldsymbol{\theta})).$$

Then $\int f(\cdot;\boldsymbol{\theta}) = 1$ for $\boldsymbol{\theta} \in \Theta$. Let $F(\cdot;\boldsymbol{\theta}) = F_n(\cdot;\boldsymbol{\theta})$ and $Q(\cdot;\boldsymbol{\theta}) = Q_n(\cdot;\boldsymbol{\theta})$ denote the distribution function and quantile function corresponding to $f(\cdot;\theta)$. Set

$$\Lambda(\boldsymbol{\theta}) = \Lambda_n(\boldsymbol{\theta}) = E(\log(f(Y;\boldsymbol{\theta})))$$

$$= \int \log(f(\cdot;\boldsymbol{\theta}))\, f = \int s(\cdot;\boldsymbol{\theta})\, f - c(\boldsymbol{\theta}), \qquad \boldsymbol{\theta} \in \Theta,$$

where Y has density f.

It is assumed from now on that $J \geq 2$ for all n. The exponential family $f(\cdot;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is not identifiable, for if we add a constant to each element of $\boldsymbol{\theta}$, we do not change $f(\cdot;\boldsymbol{\theta})$. Let $\Theta_0 = \Theta_{n0}$ denote the $(J-1)$-dimensional subspace of Θ consisting of those vector $\boldsymbol{\theta} \in \Theta$ whose entries add up to zero.

Let $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{H}_n(\boldsymbol{\theta})$ denote the Hessian matrix of $c(\cdot)$ at $\boldsymbol{\theta}$, which is the symmetric $J \times J$ matrix having entry $\partial^2 c(\boldsymbol{\theta})/\partial\theta_j\,\partial\theta_k$ in row j and column k for $1 \leq j, k \leq J$. It is an elementary and well-known property of exponential families [see Lehmann (1983)] that if $\boldsymbol{\theta}, \boldsymbol{\tau} \in \Theta$, then

$$(1) \qquad \boldsymbol{\tau}^t \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\tau} = \int [s(\cdot;\boldsymbol{\tau}) - a]^2 f(\cdot;\boldsymbol{\theta}), \quad \text{where } a = \int s(\cdot;\boldsymbol{\tau})\, f(\cdot;\boldsymbol{\theta}).$$

Thus $\boldsymbol{\tau}^t \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\tau} > 0$ if $\boldsymbol{\tau}$ is a nonzero element of Θ_0. Consequently, $c(\cdot)$ is strictly convex on Θ_0. Since $-\mathbf{H}(\boldsymbol{\theta})$ is the Hessian matrix of $\Lambda(\cdot)$ at $\boldsymbol{\theta}$, $\Lambda(\cdot)$ is strictly concave on Θ_0. If $\boldsymbol{\theta} \in \Theta_0$ and $\boldsymbol{\theta} \neq 0$, then $\Lambda(t\boldsymbol{\theta}) = t\int s(\cdot;\boldsymbol{\theta}) f - \log(\int \exp(ts(\cdot;\boldsymbol{\theta})))$ and $s(\cdot;\boldsymbol{\theta})$ is not almost everywhere equal to a constant in $\mathscr{Y}$; thus $\Lambda(t\boldsymbol{\theta}) \to -\infty$ as $|t| \to \infty$ [consider the maximum value of $s(\cdot;\boldsymbol{\theta})$ on $\mathscr{Y}$]. It follows that for each $n \geq 1$, there is a unique $\boldsymbol{\theta}^* = \boldsymbol{\theta}_n^* \in \Theta_0$ that maximizes $\Lambda(\cdot)$ on Θ_0. Set $f^* = f_n^* = f(\cdot;\boldsymbol{\theta}^*)$, $F^* = F_n^* = F(\cdot;\boldsymbol{\theta}^*)$ and $Q^* = Q_n^* = Q(\cdot;\boldsymbol{\theta}^*)$.

It follows from the assumption on f that $\log(f)$ is continuous and hence bounded on $\mathscr{Y}$. Set

$$\delta = \delta_n = \inf_{s \in \mathscr{S}} \|s - \log(f)\|_\infty.$$

If $J \to \infty$ as $n \to \infty$, then $\delta = o(1)$ by (2) on page 167 of de Boor (1978). Let m be a nonnegative integer with $m \leq q$, let $0 < a \leq 1$ and set $p = m + a$. If f is m-times differentiable and its mth derivative satisfies a Hölder condition with index a, then $\delta = O(J^{-p})$ [see de Boor (1978)].

THEOREM 1.

(i) $$\|f^* - f\|_\infty = O(\delta);$$

(ii) $$\|F^* - F\|_\infty = O(\delta/J);$$

and

(iii) $$\|Q^* - Q\|_\infty = O(\delta/J).$$

Let $l(\cdot) = l_n(\cdot)$ be the log-likelihood function based on the log-spline model and the random sample of size n, which is given by

$$l(\boldsymbol{\theta}) = \sum_i \log(f(Y_i;\boldsymbol{\theta})) = \sum_i (s(Y_i;\boldsymbol{\theta}) - c(\boldsymbol{\theta})).$$

Then $l(\cdot)$ is a strictly concave function on Θ_0. Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_n \in \Theta_0$ denote the maximum-likelihood estimate of $\boldsymbol{\theta} \in \Theta_0$ based on the random sample of size n. The $\hat{\boldsymbol{\theta}}$ is unique if it exists. (The log-likelihood function has a maximum if and only if there is no nonconstant function $s \in \mathscr{S}$ such that $Y_1, \ldots, Y_n$ all maximize s.) Set $\hat{f} = \hat{f}_n = f(\cdot;\hat{\boldsymbol{\theta}})$, $\hat{F} = \hat{F}_n = F(\cdot;\hat{\boldsymbol{\theta}})$ and $\hat{Q} = \hat{Q}_n = Q(\cdot;\hat{\boldsymbol{\theta}})$.

From now on it is assumed that

$$J = o(n^{0.5-\varepsilon}) \quad \text{for some } \varepsilon > 0. \tag{2}$$

THEOREM 2. (i) $\hat{\boldsymbol{\theta}}$ *exists except on an event whose probability tends to zero with* n;

(ii) $$|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*| = O_P(J/\sqrt{n});$$

(iii) $$\max_{1\le j\le J} |\hat{\theta}_j - \theta_j^*| = O_P\left(\sqrt{J\log(J)/n}\right);$$

(iv) $$\|\hat{f} - f^*\|_2 = O_P\left(\sqrt{J/n}\right);$$

(v) $$\|\hat{f} - f^*\|_\infty = O_P\left(\sqrt{J\log(J)/n}\right);$$

(vi) $$\|\hat{F} - F^*\|_\infty = O_P(1/\sqrt{n});$$

and

(vii) $$\|\hat{Q} - Q^*\|_\infty = O_P(1/\sqrt{n}).$$

Let $\mathbf{I}(\boldsymbol{\theta}) = \mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{H}(\boldsymbol{\theta})$ denote the information matrix based on the random sample of size n. Then $\mathbf{I}(\boldsymbol{\theta})$ has range Θ_0; that is, $\mathbf{I}(\boldsymbol{\theta})\tau \in \Theta_0$ for $\tau \in \Theta$. Also, there is a positive semidefinite symmetric $J \times J$ matrix $(\mathbf{I}(\boldsymbol{\theta}))^-$ having range Θ_0 such that

$$\mathbf{I}(\boldsymbol{\theta})(\mathbf{I}(\boldsymbol{\theta}))^- \tau = (\mathbf{I}(\boldsymbol{\theta}))^- \mathbf{I}(\boldsymbol{\theta})\tau, \qquad \tau \in \Theta_0.$$

The matrix $(\mathbf{I}(\boldsymbol{\theta}))^-$ is referred to as the generalized inverse of $\mathbf{I}(\boldsymbol{\theta})$. Set $\mathbf{I}^* = \mathbf{I}_n^* = \mathbf{I}(\boldsymbol{\theta}^*)$, $(\mathbf{I}^*)^- = (\mathbf{I}_n^*)^- = (\mathbf{I}(\boldsymbol{\theta}^*))^-$, $\hat{\mathbf{I}} = \hat{\mathbf{I}}_n = \mathbf{I}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{I}}^- = \hat{\mathbf{I}}_n^- = (\mathbf{I}(\hat{\boldsymbol{\theta}}))^-$. Given $y \in \mathscr{Y}$, let

$$\mathbf{G}^*(y) = \mathbf{G}_n^*(y) \in \boldsymbol{\Theta}_0 \quad \text{and} \quad \hat{\mathbf{G}}(y) = \hat{\mathbf{G}}_n(y) \in \Theta_0$$

denote the J-dimensional vectors having elements

$$G_j^*(y) = B_j(y) - \frac{\partial c}{\partial \theta_j}(\boldsymbol{\theta}^*) \quad \text{and} \quad \hat{G}_j(y) = B_j(y) - \frac{\partial c}{\partial \theta_j}(\hat{\boldsymbol{\theta}}), \qquad 1 \le j \le J,$$

respectively. The asymptotic standard deviation and standard error of $\hat{f}(y)$ are

defined by

$$\mathrm{ASD}(\hat{f}(y)) = f^*(y)\sqrt{[\mathbf{G}^*(y)]^t(\mathbf{I}^*)^{-}\mathbf{G}^*(y)}$$

and

$$\mathrm{SE}(\hat{f}(y)) = \hat{f}(y)\sqrt{[\hat{\mathbf{G}}(y)]^t\hat{\mathbf{I}}^{-}\hat{\mathbf{G}}(y)}\,.$$

THEOREM 3. *Suppose that* $J \to \infty$ *as* $n \to \infty$. *Then uniformly in* $y \in \mathscr{Y}$,

$$\mathrm{ASD}(\hat{f}(y)) \sim \sqrt{J/n}\,,$$

$$\frac{\mathrm{SE}(\hat{f}(y))}{\mathrm{ASD}(\hat{f}(y))} = 1 + o_P(1),$$

and

$$\mathrm{dist}\left[\frac{\hat{f}(y) - f^*(y)}{\mathrm{ASD}(\hat{f}(y))}\right] \to N(0,1).$$

THEOREM 4. *Suppose that* $J \to \infty$ *as* $n \to \infty$. *Then*

$$\frac{\mathrm{ASD}(\hat{F}(y))}{\mathrm{SD}(\hat{F}^{\mathrm{emp}}(y))} \to 1 \qquad \textit{uniformly on compact subsets of } \mathrm{int}(\mathscr{Y}),$$

$$\mathrm{dist}\left[\frac{\hat{F}(y) - F^*(y)}{\mathrm{ASD}(\hat{F}(y))}\right] \to N(0,1) \qquad \textit{uniformly on compact subsets of } \mathrm{int}(\mathscr{Y}),$$

$$\frac{\mathrm{ASD}(\hat{Q}(p))}{\mathrm{ASD}(\hat{Q}^{\mathrm{emp}}(p))} \to 1 \qquad \textit{uniformly on compact subsets of } (0,1)$$

and

$$\mathrm{dist}\left[\frac{\hat{Q}(t) - Q^*(t)}{\mathrm{ASD}(\hat{Q}(t))}\right] \to N(0,1) \qquad \textit{uniformly on compact subsets of } (0,1).$$

The results in this paper can be extended in two directions with essentially no change in proof: The restriction that the functions in $\mathscr{S}$ be $(q-2)$-times continuously differentiable on $\mathscr{Y}$ can be weakened in an arbitrary manner and the knot locations $1/K, \ldots, (K-1)/K$ can be replaced by a sequence that is σ-quasiuniform in the sense of page 216 of Schumaker (1981) (that is, such that the ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in n).

3. Proof of Theorem 1.

LEMMA 1. $\int s(f^* - f) = 0$ *for* $s \in \mathscr{S}$.

PROOF. Choose $s \in \mathscr{S}$ and define g on $\mathbb{R}$ by $\int f^* e^{ts - g(t)} = 1$. Then $g'(0) = \int sf^*$. Also $\int(\log(f^*) + ts - g(t))f$ is maximized at $t = 0$ and hence $g'(0) = \int sf$. Thus the conclusion of the lemma is valid. □

The proof of Theorem 1(i) is contained in Stone (1989). We now verify Theorem 1(ii), from which Theorem 1(iii) follows easily. It can be assumed that $J \to \infty$ as $n \to \infty$. It will be shown below that there is a positive constant M satisfying the following condition: For $n \geq 1$ and $x \in [2M/J, 1 - M/J]$ there is an $s \in \mathscr{S}$ such that $-1 \leq s \leq 1$, $s = 1$ on $[M/J, x - M/J]$, and $s = 0$ on $[x + M/J, 1]$. According to Lemma 1,

$$0 = \int_0^1 s(f^* - f) = \int_0^x (f^* - f) + \int_0^{M/J} (s - 1)(f^* - f) + \int_{x-M/J}^{x} (s - 1)(f^* - f) + \int_x^{x+M/J} s(f^* - f).$$

The desired result now follows from Theorem 1(i).

In constructing the desired function s it can be assumed that $q > 1$ (since the result is obvious when $q = 1$). Let B_1 and B_2 be elements of the usual B-spline basis with q replaced by $q - 1$ such that B_1 vanishes outside $[0, M/J]$ and B_2 vanishes outside $[x - M/J, x + M/J]$. Then B_1 and B_2 are nonnegative functions. Let a_1 and a_2 be positive constants such that $\int a_1 B_1 = 1$ and $\int a_2 B_2 = 1$. Then the function s defined by

$$s(y) = \int_0^y [a_1 B_1(z) - a_2 B_2(z)]\, dz, \qquad 0 \leq y \leq 1,$$

has the desired properties. This completes the proof of Theorem 1. □

4. Proof of Theorem 2. The proof is broken up into a series of lemmas. In particular, the first conclusion is contained in Lemma 8, the second and fourth conclusions are contained in Lemma 12 and the third conclusion follows from (2) and Lemmas 19 and 20. For the proofs of the remaining three conclusions of Theorem 2, see the discussion following the proof of Lemma 20.

LEMMA 2. *For any positive number M there is a $\delta > 0$ such that if f and g are positive density functions on $\mathscr{Y}$, $\|g\|_\infty \leq M$ and $\|\log(f) - \log(g)\|_\infty \leq M$, then*

$$\min_a \int [\log(f) - \log(g) - a]^2 \geq \delta \int [\log(f) - \log(g)]^2.$$

PROOF. The minimum value of a is $a_0 = \int [\log(f) - \log(g)]$. Set

$$h = \log(f) - \log(g) - a_0.$$

Then $|h| \leq 2M$. Since f and g integrate to 1, $a_0 = -\log(1 + \int g(e^h - 1))$. Thus there is a positive constant M_1 depending on M such that $a_0^2 \leq M_1 \int h^2$. Consequently,

$$\int [\log(f) - \log(g)]^2 = a_0^2 + \int h^2 \leq (M_1 + 1) \int h^2,$$

which yields the desired result. □

LEMMA 3. *There is a positive number M such that*

$$|c(\boldsymbol{\theta})| \leq M\|\log(f(\cdot;\boldsymbol{\theta}))\|_\infty, \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta}_0.$$

PROOF. Since

$$\log(f(\cdot;\boldsymbol{\theta})) = \sum_j \left[\theta_j - c(\boldsymbol{\theta})\right] B_j, \qquad \boldsymbol{\theta} \in \Theta,$$

the desired result follows from (viii) on page 155 of de Boor (1978). □

LEMMA 4. *Let M be a positive constant. Then there are positive constants M_1 and δ such that if $n \geq 1$, $\boldsymbol{\theta} \in \Theta$ and $\|\log(f(\cdot;\boldsymbol{\theta})) - \log(f)\|_2 \leq M/\sqrt{J}$, then*

$$\|\log(f(\cdot;\boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*)))\|_\infty \leq M_1, \qquad 0 \leq t \leq 1,$$

and

$$\Lambda(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}^*) \leq -\delta\|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_2^2.$$

PROOF. Without loss of generality, it can be assumed that $\boldsymbol{\theta} \in \Theta_0$. It follows from Theorem 1(i) and from Lemma 7 of Stone (1986) that there is a positive constant M_2 depending on M such that if n and $\boldsymbol{\theta}$ are as in the statement of the present lemma, then $\|\log(f^*)\|_\infty \leq M_2$ and $\|\log(f(\cdot;\boldsymbol{\theta}))\|_\infty \leq M_2$. Thus by Lemma 3 there is a positive constant M_3 depending on M such that $\|s(\cdot;\boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*))\|_\infty \leq M_3$, $0 \leq t \leq 1$. This yields the first conclusion of the lemma. Observe that

$$\frac{d}{dt}\Lambda(\boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*))\bigg|_{t=0} = 0.$$

Thus it follows from (1) and Taylor's theorem with remainder that

$$\Lambda(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}^*) = -\tfrac{1}{2}\int \left[s(\cdot;\boldsymbol{\theta} - \boldsymbol{\theta}^*) - a\right]^2 f(\cdot;\boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*))$$

for some $t \in (0, 1)$ and $a \in \mathbb{R}$. The second conclusion of the lemma now follows from the first conclusion and Lemma 2. □

LEMMA 5. *Given $M > 0$, there is a $\delta > 0$ such that*

$$P\left(\left|\frac{l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}^*)}{n} - [\Lambda(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}^*)]\right| \geq b\|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_2\right) \leq 2e^{-\delta n b^2}$$

for $n \geq 1$, $\boldsymbol{\theta} \in \Theta$ and $0 < b \leq M\sqrt{J}$.

PROOF. Write

$$l(\boldsymbol{\theta}) - l(\boldsymbol{\theta}^*) - n[\Lambda(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}^*)] = \sum_i Z_i,$$

where

$$Z_i = \log(f(Y_i;\boldsymbol{\theta})) - \log(f^*(Y_i)) - E\left[\log(f(Y_i,\boldsymbol{\theta})) - \log(f^*(Y_i))\right].$$

Set $a = \|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_2$. By Lemma 7 of Stone (1986) there is a positive number M_1 such that $\|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_\infty \le M_1 a\sqrt{J}$. Thus there is a positive constant M_2 such that $|Z_i| \le M_2 a\sqrt{J}$ and $\operatorname{var}(Z_i) \le M_2 a^2$. The desired result now follows from Bernstein's inequality [see (2.13) of Hoeffding (1963)]. □

The next result is an immediate consequence of the definitions of the various terms.

LEMMA 6. *If* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, *then*

$$\left| \frac{l(\boldsymbol{\theta}_2) - l(\boldsymbol{\theta}_1)}{n} - (\Lambda(\boldsymbol{\theta}_2) - \Lambda(\boldsymbol{\theta}_1)) \right| \le 2\|\log(f(\cdot;\boldsymbol{\theta}_2)) - \log(f(\cdot;\boldsymbol{\theta}_1))\|_\infty.$$

Set $\mathscr{F} = \mathscr{F}_n = \{f(\cdot;\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta\}$. It is convenient to define the "diameter" of a subset A of $\mathscr{F}$ as $\sup\{\|\log(f_2) - \log(f_1)\|_\infty: f_1, f_2 \in A\}$. The next result, essentially Lemma 12 of Stone (1986), is a consequence of Lemma 7 of that paper and (viii) on page 155 of de Boor (1978).

LEMMA 7. *Given* $\varepsilon > 0$ *and* $\delta > 0$, *there is an* $M > 0$ *such that the following is valid*:

$$\left\{f(\cdot;\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta \text{ and } \|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_2 \le n^\varepsilon\sqrt{J/n}\right\}$$

can be covered by $O(\exp(MJ\log(n)))$ *subsets each having diameter at most* $\delta n^{2\varepsilon} J/n$.

LEMMA 8. $\hat{f}$ *exists and is unique except on an event whose probability tends to zero as* $n \to \infty$. *Moreover,* $\|\log(\hat{f}) - \log(f^*)\|_2 = O_P(n^\varepsilon\sqrt{J/n})$ *for* $\varepsilon > 0$.

PROOF. Set $b = b_n = n^\varepsilon\sqrt{J/n}$ and

$$\Theta_1 = \Theta_{n1} = \{\boldsymbol{\theta} \in \Theta_0: \|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_2 \le b\}.$$

Then Θ_1 is a compact set whose boundary, relative to Θ_0, is contained in

$$\Theta_2 = \Theta_{n2} = \{\boldsymbol{\theta} \in \Theta_0: \|\log(f(\cdot;\boldsymbol{\theta})) - \log(f^*)\|_2 = b\}.$$

In light of (2), it can be assumed that there is a positive constant M such that $b \le M/\sqrt{J}$ for $n \ge 1$. By Lemma 4 there is a $\delta > 0$ such that $\Lambda(\boldsymbol{\theta}) - \Lambda(\boldsymbol{\theta}^*) \le -\delta b^2$, $\boldsymbol{\theta} \in \Theta_2$. Thus by Lemmas 5 through 7, except on an event whose probability tends to zero as $n \to \infty$,

$$l(\boldsymbol{\theta}) < l(\boldsymbol{\theta}^*), \qquad \boldsymbol{\theta} \in \Theta_2,$$

and hence $l(\cdot)$ has a local maximum in the interior of Θ_1 relative to Θ_0. The desired conclusions now follow from the strict concavity of $l(\boldsymbol{\theta})$ on Θ_0. (The first conclusion of the lemma can also be obtained from the necessary and

sufficient condition for the log-likelihood function to have a maximum that was mentioned in Section 2.) □

Let $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}_n(\boldsymbol{\theta}) \in \Theta_0$ denote the gradient of $l(\cdot)$ at $\boldsymbol{\theta}$; that is, the J-dimensional vector whose jth entry is

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} = \sum_i \left(B_j(Y_i) - \frac{\partial c}{\partial \theta_j}(\boldsymbol{\theta}) \right).$$

Set $\mathbf{S}^* = \mathbf{S}_n^* = \mathbf{S}(\boldsymbol{\theta}^*)$. Then $E\mathbf{S}^* = \mathbf{0}$ and

$$E(|\mathbf{S}^*|^2) = n \sum_j \operatorname{var}\big(B_j(Y)\big) \le n \sum_j EB_j^2(Y) = nE \sum_j B_j^2(Y) \le n.$$

Consequently, the following result is valid.

LEMMA 9. $|\mathbf{S}^*|^2 = O_P(n)$.

The maximum-likelihood equation $\mathbf{S}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ for $\hat{\boldsymbol{\theta}}$ can be rewritten as $\mathbf{D}(\hat{\boldsymbol{\theta}}\boldsymbol{\theta}^*) = \mathbf{S}^*$, where $\mathbf{D} = \mathbf{D}_n$ is the $J \times J$ matrix defined by $\mathbf{D} = n\int_0^1 \mathbf{H}(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))\, dt$.

LEMMA 10. *There is a $\delta > 0$ such that, except on an event whose probability tends to zero as $n \to \infty$, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \ge \delta n \|\log(\hat{f}) - \log(f^*)\|_2^2$.*

PROOF. It follows from Lemmas 4 and 8 and (2) that

$$\max_{0 \le t \le 1} \|\log\big(f(\cdot; \boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))\big)\|_\infty = O_P(1).$$

The desired result now follows from (1), Theorem 1(i) and Lemma 2. □

LEMMA 11. $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{S}^* = O_P(\sqrt{nJ})\|\log(\hat{f}) - \log(f^*)\|_2$.

PROOF. Let $\mathbf{1}$ denote the J-dimensional vector each of whose coordinates is 1. Then $\mathbf{1}^t\mathbf{S}^* = 0$ since $\mathbf{S}^* \in \Theta_0$. Now

$$\log(\hat{f}) - \log(f^*) = \sum_j \left[\hat{\theta}_j - \theta_j^* - c(\hat{\boldsymbol{\theta}}) + c(\boldsymbol{\theta}^*)\right] B_j, \tag{3}$$

so it follows form Lemma 9 together with (12) of Stone (1986) that

$$\begin{aligned} \left|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{S}^*\right|^2 &= \left|\left[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \big(c(\hat{\boldsymbol{\theta}}) - c(\boldsymbol{\theta}^*)\big)\mathbf{1}\right]^t \mathbf{S}^*\right|^2 \\ &\le \left|(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \big(c(\hat{\boldsymbol{\theta}}) - c(\boldsymbol{\theta}^*)\big)\mathbf{1}\right|^2 |\mathbf{S}^*|^2 \\ &= O_P(nJ)\|\log(\hat{f}) - \log(f^*)\|_2^2 \end{aligned}$$

as desired. □

LEMMA 12. (i) $\|\log(\hat{f}) - \log(f^*)\|_2 = O_P(\sqrt{J/n})$;

(ii) $$\|\log(\hat{f}) - \log(f^*)\|_\infty = O_P(J/\sqrt{n}) = o_P(1);$$

(iii) $$\|\hat{f} - f^*\|_2 = O_P(\sqrt{J/n});$$

and

(iv) $$|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*| = O_P(J/\sqrt{n}).$$

PROOF. According to the maximum-likelihood equation for $\hat{\boldsymbol{\theta}}$,

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{D}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{S}^*.$$

Thus the first result follows from Lemmas 10 and 11. The second result now follows from (2) and Lemma 7 of Stone (1986). The third result follows from the first two results and Theorem 1(i). Since $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{1} = 0$, it follows from (3), the first result, and (12) of Stone (1986) that

$$|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2 + J\big[c(\hat{\boldsymbol{\theta}}) - c(\boldsymbol{\theta}^*)\big]^2 = \big|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* + \big[c(\hat{\boldsymbol{\theta}}) - c(\boldsymbol{\theta}_n^*)\big]\mathbf{1}\big|^2$$
$$= O\big(J\|\log(\hat{f}) - \log(f^*)\|_2^2\big) = O_P(J^2/n)$$

and hence that the last result is valid. □

The next result follows from (2) and Lemmas 4 and 12(i).

LEMMA 13. *There are positive constants M_1 and M_2 such that, except on an event whose probability tends to zero as $n \to \infty$,*

$$M_1 \le f(\cdot; \boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)) \le M_2, \qquad 0 \le t \le 1.$$

Let $\nabla c(\boldsymbol{\theta})$ denote the gradient of $c(\cdot)$ at $\boldsymbol{\theta}$; that is, the J-dimensional vector having entries $\partial c(\theta)/\partial\theta_j$, $1 \le j \le J$.

LEMMA 14. $c(\hat{\boldsymbol{\theta}}) - c(\boldsymbol{\theta}^*) = [\nabla c(\boldsymbol{\theta}^*)]^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + O_P(J/n)$.

PROOF. Observe that $c(\hat{\boldsymbol{\theta}}) - c(\boldsymbol{\theta}^*) = [\nabla c(\boldsymbol{\theta}^*)]^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^t \mathbf{R}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, where $\mathbf{R} = \mathbf{R}_n$ is the $J \times J$ matrix defined by $\mathbf{R} = \int_0^1 (1-t)\,\mathbf{H}(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))\,dt$. The desired result now follows from (1), Lemmas 12(iv) and 13 and (12) of Stone (1986). □

LEMMA 15. *There is a positive constant M such that, except on an event whose probability tends to zero as $n \to \infty$,*

$$\sum_j \left(\sum_k \sum_m \max_{0 \le t \le 1} \left| \frac{\partial^3 c}{\partial\theta_j\,\partial\theta_k\,\partial\theta_m}(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)) \right| |\tau_k| \right)^2 \le MJ^{-2}|\boldsymbol{\tau}|^2, \qquad \boldsymbol{\tau} \in \Theta.$$

PROOF. It is easily seen that

$$\frac{\partial^3 c(\boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_m} = \int B_j B_k B_m f(\cdot;\boldsymbol{\theta}) - \int B_j B_k f(\cdot;\boldsymbol{\theta}) \int B_m f(\cdot;\boldsymbol{\theta}) - \int B_j B_m f(\cdot;\boldsymbol{\theta}) \int B_k f(\cdot;\boldsymbol{\theta}) - \int B_k B_m f(\cdot;\boldsymbol{\theta}) \int B_j f(\cdot;\boldsymbol{\theta}) + \int B_j f(\cdot;\boldsymbol{\theta}) \int B_k f(\cdot;\boldsymbol{\theta}) \int B_m f(\cdot;\boldsymbol{\theta}).$$

The desired result now follows from Lemma 13 and the basic properties of B-splines. □

It follows from Theorem 1(i) that

$$\max_{1\le j\le J} \left| \frac{\partial c}{\partial\theta_j}(\boldsymbol{\theta}^*) \right| = O(J^{-1}) \tag{4}$$

and hence that

$$|\mathbf{G}^*(y)| \sim 1 \quad \text{uniformly for } y \in \mathscr{Y}. \tag{5}$$

The next result follows from (3) and Lemma 14.

LEMMA 16. $\|\log(\hat{f}) - \log(f^*) - [\mathbf{G}^*(\cdot)]^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_\infty = O_P(J/n)$.

For $\boldsymbol{\theta} \in \Theta$ let $\min f(\cdot;\boldsymbol{\theta})$ and $\max f(\cdot;\boldsymbol{\theta})$, respectively, denote the minimum and maximum values of $f(\cdot;\boldsymbol{\theta})$.

LEMMA 17. *There are positive constants M_1 and M_2 such that*

$$M_1 J^{-1}|\boldsymbol{\tau}|^2 \min f(\cdot;\boldsymbol{\theta}) \le \boldsymbol{\tau}^t \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\tau} \le M_2 J^{-1}|\boldsymbol{\tau}|^2 \max f(\cdot;\boldsymbol{\theta})$$

for $n \ge 1$, $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\tau} \in \Theta_0$.

PROOF. By (12) of Stone (1986) there are positive constants M_1 and M_2 such that

$$M_1 J^{-1}|\boldsymbol{\tau}|^2 \le \|s(\cdot;\boldsymbol{\tau})\|_2^2 \le M_2 J^{-1}|\boldsymbol{\tau}|^2, \qquad \boldsymbol{\tau} \in \Theta.$$

It follows from (1) that

$$\boldsymbol{\tau}^t \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\tau} \le \|s(\cdot;\boldsymbol{\tau})\|_2^2 \max f(\cdot;\boldsymbol{\theta}) \le M_2 J^{-1}|\boldsymbol{\tau}|^2 \max f(\cdot;\boldsymbol{\theta})$$

for $\tau \in \Theta$. It also follows from (1) that if $\tau \in \Theta_0$, then

$$\begin{aligned}\tau^t \mathbf{H}(\theta)\tau &\geq \|s(\cdot;\tau - a\mathbf{1})\|_2^2 \min f(\cdot;\theta) \\ &\geq M_1 J^{-1}|\tau - a\mathbf{1}|^2 \min f(\cdot;\theta) \\ &\geq M_1 J^{-1}|\tau|^2 \min f(\cdot;\theta).\end{aligned}$$

Thus the conclusion of the lemma is valid. □

Let VC($\mathbf{S}^*$) denote the variance–covariance matrix of $\mathbf{S}^*$.

LEMMA 18. *There are positive constants M_1 and M_2 such that*

$$M_1 n J^{-1}|\tau|^2 \leq \tau^t \mathrm{VC}(\mathbf{S}^*)\tau \leq M_2 n J^{-1}|\tau|^2, \qquad n \geq 1, \tau \in \Theta_0.$$

PROOF. Since $\tau^t \mathrm{VC}(\mathbf{S}^*)\tau = n\int[s(\cdot;\tau) - a]^2 f$, where $a = \int s(\cdot;\tau)f$, the result follows from the argument used to prove Lemma 17. □

Consider the approximation $\hat{\varphi} = \hat{\varphi}_n \in \Theta_0$ to $\hat{\theta} - \theta^*$ defined by $\mathbf{I}^*\hat{\varphi} = \mathbf{S}^*$. Then $\hat{\varphi} = (\mathbf{I}^*)^- \mathbf{S}^*$ and hence $[\mathbf{G}^*(y)]^t\hat{\varphi} = [\mathbf{G}^*(y)]^t(\mathbf{I}^*)^- \mathbf{S}^*$. It follows easily from (5) and Lemma 17 that

(6) $|\tau^t(\mathbf{I}^*)^{-1}\mathbf{G}^*(y)| = O(n^{-1}J|\tau|)$ uniformly in $n \geq 1$, $\tau \in \Theta_0$ and $y \in \mathscr{Y}$.

LEMMA 19. $\max_{1 \leq j \leq J}|\hat{\varphi}_j| = O_P\left(\sqrt{J \log(J)/n}\right)$.

PROOF. Since $(\mathbf{I}^*)^- \mathrm{VC}(\mathbf{S}^*)(\mathbf{I}^*)^-$ is the variance–covariance matrix of $\hat{\varphi}$, it follows from Lemmas 17 and 18 that $\max_j \mathrm{var}(\hat{\varphi}_j) = O(J/n)$. Observe that

$$\hat{\varphi}_j = \sum_i \left((\mathbf{I}^*)^- \mathbf{G}^*(Y_i)\right)_j.$$

According to (6),

$$\sup_{y \in \mathscr{Y}} \max_{1 \leq j \leq J} \left|\left((\mathbf{I}^*)^- \mathbf{G}^*(y)\right)_j\right| = O(J/n).$$

The desired result now follows from (2) and Bernstein's inequality. □

LEMMA 20. $|\hat{\theta} - \theta^* - \hat{\varphi}|^2 = O_P(n^{-2}J^3 \log(J))$.

PROOF. It follows from the maximum-likelihood equation that

$$\hat{\theta} - \theta^* = \hat{\varphi} - (\mathbf{I}^*)^- (\mathbf{D} - \mathbf{I}^*)(\hat{\theta} - \theta^*).$$

According to Lemmas 13 and 17,

$$\left|(\mathbf{I}^*)^- (\mathbf{D} - \mathbf{I}^*)(\hat{\theta} - \theta^*)\right|^2 = O_P\left(n^{-2}J^2\left|(\mathbf{D} - \mathbf{I}^*)(\hat{\theta} - \theta^*)\right|^2\right).$$

The entry in row j and column k of $\mathbf{D} - \mathbf{I}^*$ can be written as

$$n\sum_m A_{jkm}(\hat{\theta}_m - \theta_m^*),$$

where

$$A_{jkm} = A_{njkm} = \int_0^1 (1-t) \frac{\partial^3 c}{\partial\theta_j\,\partial\theta_k\,\partial\theta_m}\left(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right) dt.$$

Thus the jth entry of $(\mathbf{D} - \mathbf{I}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is

$$n \sum_k \sum_m A_{jkm}(\hat{\theta}_k - \theta_k^*)(\hat{\theta}_m - \theta_m^*).$$

Hence by Lemmas 12 and 15

$$|(\mathbf{D} - \mathbf{I}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)|^2 = O_P\left(n \max_{1 \le j \le J} \left(\hat{\theta}_j - \theta_j^*\right)^2\right)$$

and therefore

$$|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \hat{\boldsymbol{\varphi}}|^2 = O_P\left(n^{-1}J^2 \max_{1 \le j \le J} \left(\hat{\theta}_j - \theta_j^*\right)^2\right).$$

Consequently, by Lemma 19,

$$\max_{1 \le j \le J} \left(\hat{\theta}_j - \theta_j^*\right)^2 = O_P\left(n^{-1}J \log(J) + n^{-1}J^2 \max_{1 \le j \le J} \left(\hat{\theta}_j - \theta_j^*\right)^2\right).$$

Thus by (2)

$$\max_{1 \le j \le J} \left(\hat{\theta}_j - \theta_j^*\right)^2 = O_P(n^{-1}J \log(J)),$$

which yields the desired result. □

This completes the proof of the first four conclusions of Theorem 2 (see the first paragraph of this section). We will now verify the fifty conclusion. It follows from (3) and Lemma 14 that

$$\|\log(\hat{f}) - \log(f^*) - [\mathbf{G}^*(\cdot)]^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_\infty = O_P(J/n). \tag{7}$$

Since

$$[\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}} = \sum_j B_j(y)\hat{\varphi}_j - \sum_j \frac{\partial c}{\partial\theta_j}(\boldsymbol{\theta}^*)\hat{\varphi}_j,$$

it follows from (4) and Lemma 19 that

$$\left\|[\mathbf{G}^*(\cdot)]^t\hat{\boldsymbol{\varphi}}\right\|_\infty = O_P\left(\sqrt{J \log(J)/n}\right). \tag{8}$$

The fifth conclusion follows from (2), (5), (7), (8), Theorem 1(i) and Lemmas 12 (ii) and 20.

We will now verify the sixth conclusion. It follows from (4) and Lemma 20 that

$$(\nabla c(\boldsymbol{\theta}^*))^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \hat{\boldsymbol{\varphi}}) = O_P\left(n^{-1}J\sqrt{\log(J)}\right). \tag{9}$$

It follows from Lemma 20 (and basic properties of B-splines) that

$$\sum_j |\hat{\theta}_j - \theta_j - \hat{\varphi}_j| \int B_j = O_P\left(n^{-1}J\sqrt{\log(J)}\right). \tag{10}$$

By (9) and (10),

$$\int \left|[\mathbf{G}^*(\cdot)]^t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* - \hat{\boldsymbol{\varphi}})\right| = O_P\left(n^{-1}J\sqrt{\log(J)}\right). \tag{11}$$

The sixth conclusion is a consequence of (2), (5), (7), (8), (11), Theorem 1(i), Lemma 20 and the following result; the seventh conclusion follows from the sixth conclusion and Theorem 1.

LEMMA 21. $\max_{0 \le x \le 1} |\int_0^x f^*(y)[\mathbf{G}^*(y)]^t \hat{\boldsymbol{\varphi}}\, dy| = O_P(1/\sqrt{n})$.

PROOF. Observe that

$$\begin{aligned}\operatorname{var}\left([\nabla c(\boldsymbol{\theta}^*)]^t \hat{\boldsymbol{\varphi}}\right) &= \operatorname{var}\left([\nabla c(\boldsymbol{\theta}^*)]^t (\mathbf{I}^*)^- \mathbf{S}^*\right) \\ &= \left[(\mathbf{I}^*)^- \nabla c(\boldsymbol{\theta}^*)\right]^t \mathrm{VC}(\mathbf{S}^*)(\mathbf{I}^*)^- \nabla c(\boldsymbol{\theta}^*).\end{aligned}$$

Thus it follows from (4) and Lemmas 17 and 18 that

$$\operatorname{var}\left([\nabla c(\boldsymbol{\theta}^*)]^t \hat{\boldsymbol{\varphi}}\right) = O\left(nJ^{-1}|(\mathbf{I}^*)^- \nabla c(\boldsymbol{\theta}^*)|^2\right) = O\left(n^{-1}J|\nabla c(\boldsymbol{\theta}^*)|^2\right) = O(n^{-1})$$

and hence that $[\nabla c(\boldsymbol{\theta}^*)]^t \hat{\boldsymbol{\varphi}} = O_P(1/\sqrt{n})$. Consequently, to prove the desired result, it suffices to verify that

$$\max_{0 \le x \le 1} \left| \sum_j \hat{\varphi}_j \int_0^x f^*(y)\, B_j(y)\, dy \right| = O_P\left(1/\sqrt{n}\right). \tag{12}$$

For any particular value of x, all but a bounded number of terms $\int_0^x f^*(y)B_j(y)\,dy$ are equal to $\int_0^1 f^*(y)B_j(y)\,dy$ or to zero. By (2) and Lemma 19, the total contribution of the bounded number of exceptional terms is

$$O_P\left(J^{-1}\sqrt{J\log(J)/n}\right) = o_P\left(1/\sqrt{n}\right).$$

Thus, by the form of the supports of the B-splines B_j, $1 \le j \le J$ [see de Boor (1978)], to verify (12), it suffices to show that

$$\max_{1 \le k \le J} \left| \sum_{j=1}^{k} \hat{\varphi}_j \int f^* B_j \right| = O_P\left(1/\sqrt{n}\right). \tag{13}$$

Let $\mathcal{J}$ be a subset of consecutive integers in $\{1, \ldots, J\}$ and let J' denote the number of integers in $\mathcal{J}$. Let $\boldsymbol{\tau}$ denote the J-dimensional vector having elements $\int f^* B_j$ for $j \in \mathcal{J}$ and zero otherwise. Then

$$|\boldsymbol{\tau}|^2 = O(J'/J^2) \quad \text{uniformly in } \mathcal{J} \text{ and } n. \tag{14}$$

Since

$$\operatorname{var}\left(\sum_{j\in\mathscr{J}}\hat{\varphi}_j\int f^*B_j\right) = \left((\mathbf{I}^*)^{-}\boldsymbol{\tau}\right)^t\mathrm{VC}(\mathbf{S}^*)(\mathbf{I}^*)^{-}\boldsymbol{\tau}.$$

it follows from (14) and Lemmas 17 and 18 that

$$\operatorname{var}\left(\sum_{j\in\mathscr{J}}\hat{\varphi}_j\int f^*B_j\right) = O\left[\frac{J'}{nJ}\right] \quad \text{uniformly in } \mathscr{J} \text{ and } n. \tag{15}$$

Observe next that

$$\sum_{j\in\mathscr{J}}\hat{\varphi}_j\int f^*B_j = \sum_i\left(\sum_{j\in\mathscr{J}}\left((\mathbf{I}^*)^{-}\mathbf{G}^*(Y_i)\right)_j\int f^*B_j\right) = \sum_i X_{\mathscr{J}i}. \tag{16}$$

Here $X_{\mathscr{J}i} = X_{n\mathscr{J}i}$, $1 \le i \le n$, are independent random variables having mean 0; by (5), Lemma 17 and the basic properties of B-splines,

$$|X_{\mathscr{J}i}| \le b \quad \text{with } b = b_n = O(n^{-1}\sqrt{J}). \tag{17}$$

It follows from (15) through (17) and Bernstein's inequality that there is a $\beta > 0$, which does not depend on n or $\mathscr{J}$, such that

$$\begin{aligned} &P\left(\left|\sum_{j\in\mathscr{J}}\hat{\varphi}_j\int f^*B_j\right| \ge An^{-1/2}(J'/J)^a\right) \\ &\quad\le 2\left[\exp\left(-\beta A(n/J)^{1/2}(J'/J)^a\right) + \exp\left(-\beta A^2(J/J')^{1-2a}\right)\right], \end{aligned} \tag{18}$$

for $A > 0$ and $0 < a < 0.5$.

Set $R = R_n = \min[r\colon 2^r \ge J]$. For $0 \le r \le R$, let $\mathscr{M}_r = \mathscr{M}_{nr}$ denote the collection of all sets of integers of the form $\{(m-1)2^r + 1, \ldots, m2^r\}$, where $1 \le m \le J/2^r$. It follows from (2) and (18) that, for any $\varepsilon \sim 0$, A can be chosen sufficiently large so that

$$P\left(\left|\sum_{j\in\mathscr{J}}\hat{\varphi}_j\int f^*B_j\right| \ge An^{-1/2}(J'/J)^a \text{ for some } \mathscr{J}\in\bigcup_0^R\mathscr{M}_r\right) \le \varepsilon, \qquad n \ge 1. \tag{19}$$

For $1 \le k \le J$, $\{1, \ldots, k\}$ can be written as a disjoint union of sets $\mathscr{J}\in\bigcup_0^R\mathscr{M}_r$ such that for $0 \le r \le R$, there is at most one such $\mathscr{J}\in\mathscr{M}_r$. Thus it follows from (19) that (13) holds, as desired. □

5. Proof of Theorem 3. Now

$$\operatorname{Var}\left([\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}}\right) = [\mathbf{G}^*(y)]^t(\mathbf{I}^*)^{-}\mathrm{VC}(\mathbf{S}^*)(\mathbf{I}^*)^{-}\mathbf{G}^*(y), \tag{20}$$

so it follows from (5), Theorem 1(i) and Lemmas 17 and 18 that

$$\operatorname{Var}\left([\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}}\right) \sim J/n. \tag{21}$$

LEMMA 22. *Uniformly in* $y \in \mathrm{I}$,

$$\operatorname{dist}\left(\frac{[\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}}}{\mathrm{SD}\big([\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}}\big)}\right) \to N(0,1) \quad \text{as } n \to \infty.$$

PROOF. Observe that $[\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}} = \sum Z_i$, where $Z_i = Z_{ni} = [\mathbf{G}^*(y)]^t\mathbf{I}^{-}\mathbf{G}^*(Y_i)$ for $1 \le i \le n$. For each n, the random variables $Z_1, \ldots, Z_n$ have mean 0 and are independent and identically distributed. Moreover,

$$\begin{aligned}&\big|[\mathbf{G}^*(y)]^t(\mathbf{I}^*)^{-}\mathbf{G}^*(Y_i)\big|^2\\ &\quad\le \big|[\mathbf{G}^*(y)]^t(\mathbf{I}^*)^{-}\mathbf{G}^*(y)\big|\,\big|[\mathbf{G}^*(Y_i)]^t(\mathbf{I}^*)^{-}\mathbf{G}^*(Y_i)\big| = O(J^2/n^2).\end{aligned}$$

The desired result now follows from (2), (21) and the central limit theorem [see the corollary on page 201 of Chung (1974)]. □

LEMMA 23. *There is a positive constant* M *such that*

$$\big|\boldsymbol{\tau}^t\mathrm{VC}(\mathbf{S}^*)\boldsymbol{\tau} - \boldsymbol{\tau}^t\mathbf{I}^*\boldsymbol{\tau}\big| \le MnJ^{-1}\delta|\boldsymbol{\tau}|^2, \qquad n \ge 1, \boldsymbol{\tau} \in \Theta.$$

PROOF. Set $a = a_n = \int s(\cdot;\boldsymbol{\tau})f$ and $a^* = a_n^* = \int s(\cdot;\boldsymbol{\tau})f^*$. Then

$$\boldsymbol{\tau}^t\mathrm{VC}(\mathbf{S}^*)\boldsymbol{\tau} = n\int[s(\cdot;\boldsymbol{\tau}) - a]^2 f$$

and

$$\boldsymbol{\tau}^t\mathbf{I}^*\boldsymbol{\tau} = n\int[s(\cdot;\boldsymbol{\tau}) - a^*]^2 f^*.$$

The desired result now follows easily from Theorem 1(i) and (12) of Stone (1986). □

It follows from (5), Theorem 1(i) and Lemmas 17 and 23 that there is a positive constant M such that, for $n \ge 1$ and $y \in \mathscr{Y}$,

$$(22)\quad \big|[\mathbf{G}^*(y)]^t(\mathbf{I}^*)^{-}\mathrm{VC}(\mathbf{S}^*)(\mathbf{I}^*)^{-}\mathbf{G}^*(y) - [\mathbf{G}^*(y)]^t(\mathbf{I}^*)^{-}\mathbf{G}^*(y)\big| \le Mn^{-1}J\delta.$$

It follows from (2), (5), (7) and Lemma 20 that

$$\big|\log(\hat{f}) - \log(f^*) - [\mathbf{G}^*(\cdot)]^t\hat{\boldsymbol{\varphi}}\big|_\infty = o_P\big(\sqrt{J/n}\big).$$

Thus by (21) and Lemma 22,

$$\operatorname{dist}\left(\frac{\log(\hat{f}) - \log(f^*)}{\mathrm{SD}\big([\mathbf{G}^*(y)]^t\hat{\boldsymbol{\varphi}}\big)}\right) \to N(0,1) \quad \text{uniformly in } y \text{ as } n \to \infty.$$

It follows easily from this together with (2), (20) through (22) and Theorem 1(i) that the first and third conclusions of Theorem 3 are valid.

LEMMA 24. *Uniformly in* $\tau \in \Theta$,

$$\left|(\hat{\mathbf{I}} - \mathbf{I}^*)\tau\right|^2 = O_P\big(nJ^{-1}\log(J)\big)|\tau|^2.$$

PROOF. Observe that

$$\frac{\partial^2 c}{\partial\theta_j\,\partial\theta_k}(\hat{\theta}) - \frac{\partial^2 c}{\partial\theta_j\,\partial\theta_k}(\theta^*) = \sum_m \int_0^1 \frac{\partial^3 c}{\partial\theta_j\,\partial\theta_k\,\partial\theta_m}\big(\theta^* + t(\hat{\theta} - \theta^*)\big)\big(\hat{\theta}_m - \theta_m^*\big)\,dt.$$

Thus the desired result follows form Lemmas 15, 19 and 20. □

Since $\hat{\mathbf{I}}^- - (\mathbf{I}^*)^- = (\mathbf{I}^*)^-(\mathbf{I}^* - \hat{\mathbf{I}})\hat{\mathbf{I}}^-$, the next result follows from Lemmas 13, 17 and 24.

LEMMA 25. *Uniformly in* $\tau \in \Theta_0$,

$$\left|(\hat{\mathbf{I}}^- - (\mathbf{I}^*)^-)\tau\right|^2 = O_P\big(n^{-3}J^3\log(J)\big)|\tau|^2.$$

LEMMA 26. $|\hat{\mathbf{G}}(y) - \mathbf{G}^*(y)|^2 = O_P(1/n)$ *uniformly in* y.

PROOF. Observe that $\hat{\mathbf{G}}(y) - \mathbf{G}^*(y) = -[\nabla c(\hat{\theta}) - \nabla c(\theta^*)]$. Since

$$\nabla c(\hat{\theta}) - \nabla c(\theta^*) = \left(\int_0^1 \mathbf{H}\big(\theta^* + t(\theta^* - \hat{\theta})\big)\,dt\right)(\hat{\theta} - \theta^*),$$

the desired result follows from Theorem 2(i) and Lemmas 13 and 17. □

The next result follows from (5), (22), Theorem 1(i) and Lemmas 17, 25 and 26.

LEMMA 27. *Uniformly in* y,

$$\begin{aligned}&\big[\hat{\mathbf{G}}(y)\big]^t\hat{\mathbf{I}}^-\hat{\mathbf{G}}(y) - [\mathbf{G}^*(y)]^t(\mathbf{I}^*)^-\mathrm{VC}(\mathbf{S}^*)(\mathbf{I}^*)^-\mathbf{G}^*(y)\\ &\quad = O_P\Big(\sqrt{J^3\log(J)/n^3} + J\,\delta/n\Big).\end{aligned}$$

The second conclusion of Theorem 3 follows from (2), (20), (21), Theorems 1(i) and 2(v) and Lemma 27. □

6. Proof of Theorem 4. For a given value of $x \in [0, 1]$, set

$$g(\theta) = g_n(\theta) = \int_0^x f(y{:}\,\theta)\,dy, \qquad \theta \in \Theta.$$

It is easily seen that

$$\nabla g(\boldsymbol{\theta}^*) = \int_0^x \mathbf{G}^*(y)\, f^*(y)\, dy. \tag{23}$$

The next result follows from (2), (7), (8), (11), (23) and Theorem 1(i).

LEMMA 28. $\hat{F}(x) - F^*(x) = [\nabla g(\boldsymbol{\theta}^*)]^t \hat{\boldsymbol{\varphi}} + o_P(1/\sqrt{n})$ *uniformly in* x.

Observe that

$$\operatorname{var}\bigl([\nabla g(\boldsymbol{\theta}^*)]^t \hat{\boldsymbol{\varphi}}\bigr) = [\nabla g(\boldsymbol{\theta}^*)]^t (\mathbf{I}^*)^- \operatorname{VC}(\mathbf{S}^*)(\mathbf{I}^*)^- \nabla g(\boldsymbol{\theta}^*). \tag{24}$$

By (4) and (23),

$$|\nabla g(\boldsymbol{\theta}^*)|^2 = O(J^{-1}) \quad \text{uniformly in } x. \tag{25}$$

It follows from (24), (25), Theorem 1(i) and Lemmas 17 and 23 that

$$\operatorname{var}\bigl([\nabla g(\boldsymbol{\theta}^*)]^t \hat{\boldsymbol{\varphi}}\bigr) = [\nabla g(\boldsymbol{\theta}^*)]^t (\mathbf{I}^*)^- \nabla g(\boldsymbol{\theta}^*) + O(\delta/n) \quad \text{uniformly in } x. \tag{26}$$

Let E_θ and $\operatorname{var}_\theta$ correspond to the assumption that Y has density $f(\cdot;\boldsymbol{\theta})$. According to the Cramér–Rao inequality,

$$[\nabla g(\boldsymbol{\theta}^*)]^t (\mathbf{I}^*)^- \nabla g(\boldsymbol{\theta}^*) \le n^{-1} \operatorname{var}_{\boldsymbol{\theta}^*}(\operatorname{ind}_{[0,x]}(Y)) = n^{-1} F_n(x)[1 - F_n(x)]. \tag{27}$$

LEMMA 29. *Suppose that* $J \to \infty$ *as* $n \to \infty$. *Then*

$$\lim_n \frac{[\nabla g(\boldsymbol{\theta}^*)]^t (\mathbf{I}^*)^- \nabla g(\boldsymbol{\theta}^*)}{n^{-1} F(x)[1 - F(x)]} = 1 \quad \textit{uniformly for } x \textit{ in compact subsets of } \operatorname{int}(\mathscr{Y}).$$

PROOF. Choose $\boldsymbol{\varphi} = \boldsymbol{\varphi}_n \in \Theta_0$. By Schwarz's inequality

$$[\nabla g(\boldsymbol{\theta}^*)]^t (\mathbf{I}^*)^- \nabla g(\boldsymbol{\theta}^*) \ge \frac{\bigl([\nabla g(\boldsymbol{\theta}^*)]^t \boldsymbol{\varphi}\bigr)^2}{\boldsymbol{\varphi}^t \mathbf{I}^* \boldsymbol{\varphi}}. \tag{28}$$

By (1)

$$\boldsymbol{\varphi}^t \mathbf{I}^* \boldsymbol{\varphi} = n \operatorname{var}_{\boldsymbol{\theta}^*}(s(Y;\boldsymbol{\varphi})). \tag{29}$$

It follows from (23) that

$$[\nabla g(\boldsymbol{\theta}^*)]^t \boldsymbol{\varphi} = E_{\boldsymbol{\theta}^*}\bigl(\operatorname{ind}_{[0,x]}(Y)\bigl[s(Y;\boldsymbol{\varphi}) - E_{\boldsymbol{\theta}^*} s(Y;\boldsymbol{\varphi})\bigr]\bigr). \tag{30}$$

The desired result follows from (27) through (30), Theorem 1 (ii), the construction of $s \in \mathscr{S}_n$ used in the proof of that result and the corresponding choice of $\boldsymbol{\varphi}_n$. □

The proof of the next result is similar to that of Lemma 22.

LEMMA 30. *Uniformly for x in compact subsets of* $\operatorname{int}(\mathscr{Y})$,

$$\operatorname{dist}\left(\frac{[\nabla g(\boldsymbol{\theta}^*)]^t\hat{\boldsymbol{\varphi}}}{\operatorname{SD}([\nabla g(\boldsymbol{\theta}^*)]^t\hat{\boldsymbol{\varphi}})}\right) \to N(0,1) \quad \text{as } n \to \infty.$$

The first conclusion of Theorem 4 follows from (26) and Lemmas 28 through 30. The second conclusion follows from the first conclusion and Theorem 1. □

REFERENCES

BARRON, A. R. and CHEU, C.-H. (1988). Approximation of density functions by sequences of exponential families. Technical Report 8, Dept. Statist., Univ. Illinois at Urbana-Champaign.

DE BOOR, C. (1976). A bound on the L_∞-norm of the L_2-approximation by splines in terms of a global mesh ratio. *Math. Comp.* **30** 765–771.

DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.

BREIMAN, L., STONE, C. J. and KOOPERBERG, C. (1988). Confidence bounds for extreme quantiles. Technical Report 167, Dept. Statist., Univ. California at Berkeley.

CHUNG, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic, New York.

CRAIN, B. R. (1974). Estimation of distributions using orthogonal expansions. *Ann. Statist.* **2** 454–463.

CRAIN, B. R. (1976a). Exponential models, maximum likelihood estimation, and the Haar conditions. *J. Amer. Statist. Assoc.* **71** 737–740.

CRAIN, B. R. (1976b). More on estimation of distributions using orthogonal expansions. *J. Amer. Statist. Assoc.* **71** 741–745.

CRAIN, B. R. (1977). An information theoretic approach to approximating a probability distribution. *SIAM J. Appl. Math.* **32** 339–346.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

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

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

NEYMAN, J. (1937). "Smooth" test for goodness of fit. *Scand. Actuar. J.* **20** 149–199.

PORTNOY, S. (1986). On the central limit in R^p when $p \to \infty$. *Probab. Theory Related Fields* **73** 571–583.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *Ann. Statist.* **16** 356–366.

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

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

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics: Papers in Honor*

of Herman Chernoff on His Sixtieth Birthday (M. H. Rezvi, J. S. Rustagi and D. Siegmund, eds.) 393–406. Academic, New York.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

STONE, C. J. (1987). A nonparametric framework for statistical modelling. In *Proc. Internat. Congress of Mathematicians, Berkeley, California, 1986* 1052–1056. Amer. Math. Soc., Providence.

STONE, C. J. (1989). Uniform error bounds involving logspline models. In *Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin* (T. W. Anderson, K. B. Athreya and D. L. Iglehart, eds.) 335–355. Academic, Boston.

STONE, C. J. and KOO, C.-Y. (1986a). Additive splines in statistics. In *1985 Statistical Computing Section Proc. Amer. Statist. Assoc.* 45–48. Amer. Statist. Assoc., Washington, D.C.

STONE, C. J. and KOO, C.-Y. (1986b). Logspline density estimation. In *AMS Contemporary Math. Ser.* **29** 1–15. Amer. Math. Soc., Providence.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720