**Approximation of Density Functions by Sequences of Exponential Families**

Andrew R. Barron; Chyong-Hwa Sheu

*The Annals of Statistics*, Vol. 19, No. 3 (Sep., 1991), 1347-1369.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199109%2919%3A3%3C1347%3AAODFBS%3E2.0.CO%3B2-T

*The Annals of Statistics* is currently published by Institute of Mathematical Statistics.

*The Annals of Statistics*
1991, Vol. 19, No. 3, 1347–1369

# APPROXIMATION OF DENSITY FUNCTIONS BY SEQUENCES OF EXPONENTIAL FAMILIES[1]

BY ANDREW R. BARRON AND CHYONG-HWA SHEU

*University of Illinois at Urbana-Champaign*

Probability density functions are estimated by the method of maximum likelihood in sequences of regular exponential families. This method is also familiar as entropy maximization subject to empirical constraints. The approximating families of log-densities that we consider are polynomials, splines and trigonometric series. Bounds on the relative entropy (Kullback–Leibler distance) between the true density and the estimator are obtained and rates of convergence are established for log-density functions assumed to have square integrable derivatives.

**1. Introduction.** Consider the estimation of a probability density function $p(x)$ defined on a bounded interval. We approximate the logarithm of the density by a basis function expansion consisting of polynomials, splines or trigonometric series. The expansion yields a regular exponential family within which we estimate the density by the method of maximum likelihood. This method of density estimation arises by application of the principle of maximum entropy or minimum relative entropy subject to empirical constraints. We show that if the logarithm of the density has $r$ square-integrable derivatives, $\int |D^r \log p|^2 < \infty$, then the sequence of density estimators $\hat{p}_n$ converges to $p$ in the sense of relative entropy (Kullback–Leibler distance) $\int p \log(p/\hat{p}_n)$ at rate $O_{pr}(1/m^{2r} + m/n)$ as $m \to \infty$ and $m^2/n \to 0$ in the spline and trigonometric cases and $m^3/n \to 0$ in the polynomial case, where $m$ is the dimension of the family and $n$ is the sample size. Boundary conditions are assumed for the density in the trigonometric case. This convergence rate specializes to $O_{pr}(n^{-2r/(2r+1)})$ by setting $m = n^{1/(2r+1)}$ when the log-density is known to have degree of smoothness at least $r$. Analogous convergence results for the relative entropy are shown to hold in general, for any class of log-density functions and sequence of finite-dimensional linear spaces having $L_2$ and $L_\infty$ approximation properties.

The approximation of log-densities using polynomials has previously been considered by Neyman (1937) to define alternatives for goodness-of-fit tests, by Good (1963) as an application of the method of maximum entropy or minimum relative entropy, by Crain (1974, 1976a, b 1977) who demonstrates existence and consistency of the maximum likelihood estimator and by Mead and

Received April 1988; revised July 1989.

[1]Work supported in part by Office of Naval Research Grants N00014-86-K-0670 and N00014-89-J-1811 and by an NSF Postdoctoral Research Fellowship.

*AMS* 1980 *subject classifications*. Primary 62G05; secondary 41A17, 62B10, 62F12.

*Key words and phrases*. Log-density estimation, exponential families, minimum relative entropy estimation, Kullback–Leibler number, $L_2$ approximation.

Papanicolaou (1984) who demonstrate the usefulness of the method in some physics contexts and discuss some of the computational issues. Log-spline estimation was previously considered by Stone and Koo (1986) who address the issues of asymptotic normality, confidence intervals for the density and the selection of the knots. In work independent of ours, Stone (1989, 1990) obtains rates of convergence specifically for the spline case, though it may be possible to extend his technique to other exponential families. Some general theory on sequences of exponential families is developed in Cencov (1982) and Portnoy (1988). Of course, regular exponential family models for probability densities are extensively utilized in statistical practice and their finite-dimensional properties have been thoroughly studied; see, for example, Brown (1986). Other nonparametric estimators of the log-density are examined in Leonard (1978) and Silverman (1982). The method of sieves due to Grenander (1981) includes the estimators considered here as a special case. Consistency properties of sieves are established in Geman and Hwang (1982).

The use of exponential family density estimation is natural with an entropy based loss function. These densities are discovered to have a maximum entropy property in Shannon (1948) and Jaynes (1957), are shown more generally to have a minimum relative entropy (information projection) property in Kullback (1959) and Csiszár (1975), are identified as limits of conditional densities by Van Campenhout and Cover (1981) and Csiszár (1984) and are given axiomatic justification in Shore and Johnson (1980), Jones (1989) and Csiszár (1989). We mention two applications of density estimation which require accuracy in the sense of relative entropy, denoted $D(p\|\hat{p})$. In a stock market setup, $D(p\|\hat{p})$ bounds the difference between the optimal exponential growth rate of wealth and the actual growth rate when investment portfolios are based on the estimated density instead of the true density [Barron and Cover (1988)]. For a data compression problem, $D(p\|\hat{p})$ determines the redundancy (excess average length) of a code based on the estimated density instead of the true density [see Davisson (1973)]. Indeed, using results developed here, bounds on the redundancy of universal codes can be obtained for some nonparametric classes of densities as in Barron and Cover (1991).

Other traditional methods for nonparametric density estimation, such as kernel estimators and orthogonal series expansions (of the density rather than the log-density), have received detailed theoretical treatment of their asymptotic properties [see, e.g., Prakasa Rao (1983), Devroye and Györfi (1985) and Devroye (1987)]. For instance, it is known that for the class of densities with $r$ square integrable derivatives, an optimal convergence of the integrated squared error at rate $n^{-2r/(2r+1)}$ is achieved by kernel and orthogonal series methods [Nadaraya (1974), Bretagnolle and Huber (1979) and Efroimovich and Pinsker (1983)]. However, for $r > 2$ the kernel and orthogonal series estimators which achieve this rate have the disconcerting property that they are not necessarily strictly positive (indeed they are sometimes negative), so that these estimators are not suitable for applications which require accuracy in the Kullback–Leibler sense. Density estimators can be modified to force positivity and in some cases to permit consistency and convergence rates for the Kullback–

Leibler distance. See Barron, Györti and van der Meulen (1991) for convergence properties of the Kullback–Leibler distance for modified histogram estimators. Hall (1987) gives a detailed examination of the Kullback–Leibler risk of estimators based on positive kernels. However, no positive kernel estimator can have a faster rate of convergence than $n^{-4/5}$. In this paper we avoid these difficulties by using estimators which are natural for the information-theoretic loss function.

For probability density functions having support on the whole real line, the methods developed here are not directly applicable, because of the boundedness requirement of the log-density implicit in the assumption of integrability of the derivative. One could map the problem into the unit interval, for instance by a transformation based on a cumulative distribution. However, the transformed density will have an unbounded logarithm at the boundaries, unless the tail behavior of the true density is known and incorporated in the choice of the transformation. Nevertheless, exponential family density estimation on the whole line is plausible using bounded basis functions and a reference density $p_0$ with infinite support. It should be possible to obtain consistency for densities for which the relative entropy $\int p \log p/p_0$ is finite. It is anticipated that the rate of convergence would depend in part on the tail behavior of this integral.

In practice, the dimension $m$ of the exponential family should be chosen automatically from the data. The analysis in this paper does not directly address this issue. However, the selection of the dimension for expotential family models is examined in Barron and Cover (1991) as a special case of general model selection theory developed there. It is shown that if the dimension is chosen by an information criterion similar to those proposed by Schwarz (1978) or Rissanen (1983), then the density estimator converges in squared Hellinger distance at rate bounded by an index of resolvability. This index is of order $(n^{-1} \log n)^{2r/(2r+1)}$ for log-densities with $r$ square integrable derivatives; whereas it is of order $n^{-1} \log n$ for densities $p$ in one of the countably many exponential families. So whether the true density is in a finite- or infinite-dimensional family, we converge at a rate within a logarithmic factor of the rate obtainable with true knowledge of the family. Haughton (1988) shows that for a bounded number of exponential families, the Schwarz criterion chooses the correct family with probability tending to 1. In related contexts of regression, Shibata (1981) shows that a criterion proposed by Akaike leads to optimal convergence rate properties provided the true regression is not finite dimensional.

Multivariate density estimation on a bounded cube in $\mathbb{R}^d$ can be directly handled by the present theory using the usual product basis functions for polynomials and splines and the multi-indexed trigonometric functions. However, the use of such expansions in high dimensions is precluded by the exponential growth of the number of basis functions as a function of $d$. Other traditional density estimators, such as kernels, suffer from a similar curse of dimensionality. Methods of surface estimation in high dimensions which are based on composing lower-dimensional relationships into a network have

experienced some success; see, for example, Barron and Barron (1988) and Barron (1991).

The outline of the paper is as follows. In section 2 we state the results and discuss some of the implications. Some useful tools are developed in Sections 3, 4 and 5, followed by the proof of the general result in Section 6. Conditions are checked in Section 7 for the polynomial, spline and trigonometric cases. In Section 8 the estimator is illustrated with a practical example.

**2. Formulation and discussion of results.** Let $X_1, X_2, \ldots, X_n$ be independent random variables with an unknown probability density function $p(x)$ defined on a bounded interval, which for simplicity is taken to be the unit interval $[0, 1]$. The relative entropy (Kullback–Leibler distance) between probability densities is denoted by

$$D(p\|\hat{p}) = \int p(x)\log\frac{p(x)}{\hat{p}(x)}\,dx.$$

Throughout this paper logarithms are taken with base $e$. It is well known that $D$ is nonnegative and equals 0 if and only if $p = \hat{p}$ a.e. Also $D(p\|\hat{p}) \geq (1/2)(\int|p - \hat{p}|)^2$ [Csiszár (1967) and Kullback (1967)]. Inequalities in Section 3 show that $D$ behaves like a squared $L_2$ norm between the logarithms of the densities.

The density estimator $\hat{p}_{n,m}(x) = p_{\hat{\theta}}(x)$ is defined to maximize the likelihood in the exponential family

$$p_\theta(x) = p_0(x)\exp\left\{\sum_{k=1}^{m}\theta_k\phi_k(x) - \psi_m(\theta)\right\} \tag{2.1}$$

where $\psi_m(\theta) = \log\int p_0(x)\exp\{\sum\theta_k\phi_k(x)\}\,dx$, $\theta \in \mathbb{R}^m$. Here we are given a reference probability density function $p_0(x)$ on $[0, 1]$ and a linear space $S_m$ of functions spanned by bounded and linearly independent functions $1, \phi_1(x), \ldots, \phi_m(x)$. Three choices for the space $S_m$ are polynomials, trigonometric series and splines of order $s$ with equally spaced knots: where the degree $m$ of the polynomials, the maximum frequency $m/2$ of the trigonometric functions and the number of interior knots $m - s + 1$ in the spline case are set so as to make the dimension of the family (2.1) be equal to $m$. For simplicity, we assume $m$ is even in the trigonometric case. The reference density $p_0(x)$ is often taken to be the uniform; nevertheless, the results we obtain permit it to be any density satisfying the same smoothness assumptions as are required of $p$.

We recall several characterizations of the estimator. From the likelihood equations, $p_{\hat{\theta}}$ is the density in the family (2.1) that satisfies

$$\int\phi_k(x)\hat{p}(x)\,dx = \hat{\alpha}_k \tag{2.2}$$

for $k = 1, 2, \ldots, m$ where $\hat{\alpha}_k = (1/n)\sum_{i=1}^{n}\phi_k(X_i)$. [The maximum likelihood solution exists with high probability as shown below; uniqueness is a familiar

consequence of the strict convexity of the log-likelihood; see, for example, Brown (1986).] Equation (2.2) entails that expectations with respect to $\hat{p}$ agree with empirical expectations for all functions in the linear space $S_m$. The maximum entropy characterization, valid when $p_0$ is the uniform density, states that the estimator $\hat{p}_{n,m}$ is the unique maximizer of the entropy $-\int \hat{p} \log \hat{p}$ among all density functions which satisfy (2.2). More generally, the minimum relative entropy characterization states that given $p_0(x)$, the estimator minimizes $D(\hat{p}\|p_0)$ among all density functions which satisfy the constraint (2.2) [see Kullback (1959) and Csiszár (1975)]. The conditional limit characterization of Van Campenhout and Cover (1981) and Csiszár (1984) establishes that for large $n$, $\hat{p}_{n,m}$ is the asymptotic conditional probability density function for $X_1$ given $(1/n)\sum_{i=1}^{n}\phi_k(X_i) = \alpha_k$ when the unconditional density is taken to be $p_0$. Thus given an initial guess $p_0$, the estimator $\hat{p}_{n,m}$ is a natural update based on the sample expectations.

The parameterization of the family requires a choice of basis functions $1, \phi_1(x), \ldots, \phi_m(x)$ for the given linear space $S_m$. The maximum likelihood estimator of the density does not depend on which basis is used for the given space. Traditional basis functions are $1, x, \ldots, x^m$ in the polynomial case; $1, \cos(2\pi x), \sin(2\pi x), \ldots, \cos(2\pi(m/2)x), \sin(2\pi(m/2)x)$ in the trigonometric case; and $1, x, \ldots, x^{s-1}, ((x-\Delta)_+)^{s-1}, \ldots, ((x-k\Delta)_+)^{s-1}$ in the spline case, where $(\cdot)_+$ denotes the positive part, $\Delta = 1/(m+2-s)$ is the width between the knots and $k = m+1-s$ is the number of knots. In each case the dimension of $S_m$ is $m+1$. Parameterizations based on the Legendre polynomials as in Crain (1974, 1977) and the $B$-spline basis as in Stone and Koo (1986) are believed to have superior numerical properties in the polynomial and spline cases, respectively.

Let $W_2^r$ for $r \geq 1$ be the Sobolev space of functions $f$ on $[0,1]$ for which $f^{(r-1)}$ is absolutely continuous and $\int (f^{(r)}(x))^2\,dx$ is finite. The log-density function $f = \log p$ is assumed to be a member of this Sobolev space. This assumption forces the density to be strictly positive and finite on $[0,1]$.

The main result on the asymptotics for the exponential family density estimator in the polynomial, spline and trigonometric case is as follows.

THEOREM 1. *If $m \to \infty$, $m^2/n \to 0$ in the spline and trigonometric cases and $m \to \infty$, $m^3/n \to 0$ in the polynomial case, then the Kullback–Leibler distance for the sequence of exponential family estimators satisfies*

$$D(p\|\hat{p}_{n,m}) = O_{pr}\left(\left(\frac{1}{m}\right)^{2r} + \frac{m}{n}\right). \tag{2.3}$$

*In particular, if $m$ is proportional to $n^{1/(2r+1)}$ then*

$$D(p\|\hat{p}_n) = O_{pr}(n^{-2r/(2r+1)}). \tag{2.4}$$

*The density function $p$ is assumed to satisfy $\log p \in W_2^r$, with $r \geq 2$ in the polynomial case, $1 \leq r \leq s$ in the spline case and $r \geq 1$ in the trigonometric case. In the trigonometric case the boundary conditions $f^{(j)}(0) = f^{(j)}(1)$ for*

*$0 \le j < r$ are also required for $f = \log p$. The same requirements are assumed for the reference density $p_0$.*

REMARK 1. The convergence in probability is uniform for any set $B$ of log-densities having bounded Sobolev norm. In particular, it is seen that

$$\lim_{\mathscr{K}\to\infty} \lim_{n\to\infty} \sup_{\log p \in B} P\left\{D(p\|\hat{p}_n) \ge \left((1/m_n)^{2r} + m_n/n\right)\mathscr{K}\right\} = 0 \tag{2.5}$$

for any sequence $m_n$ satisfying $m_n \to \infty$ and $m_n^2/n \to 0$ ($m_n^3/n \to 0$ in the polynomial case) as $n \to \infty$. The requirement on the set $B$ is that there is a constant $c$ such that $\|f^{(r)}\|_2$ and $\|f\|_\infty$ are less than $c$ for all $f \in B$. (For the trigonometric case, $B$ is restricted to functions which also satisfy the indicated boundary conditions.)

REMARK 2. It is anticipated that $n^{-2r/(2r+1)}$ is the optimal minimax rate for the Kullback–Leibler distance for the class of log-densities with bounded Sobolev norm, in which case the estimators given above possess optimal rate properties. In support of this conjecture is the optimality of the same rates $n^{-2r/(2r+1)}$ for the integrated squared error for density functions with bounded Sobolev norm [Bretognolle and Huber (1979) and Efroimovich and Pinsker (1983)]. For densities which have a bounded logarithm the Kullback–Leibler number is related to the integrated squared error (see Lemma 2). Moreover, when the density is bounded away from 0, Sobolev assumptions on the density are not too different from Sobolev assumptions on the log-density. See Yu and Speed (1990) for a derivation of the minimax rate in a closely related setting.

REMARK 3. As part of the proof of the theorem, it is shown that the maximum likelihood estimate $\hat{p}_{n,m} = p_{\hat{\theta}}$ exists except in a set of probability tending to 0 as $n \to \infty$. By other methods, Crain (1976a, b) has shown that for $n > m$, the maximum likelihood estimator exists with probability 1 in the polynomial and trigonometric cases (and more generally when a Haar condition is satisfied by a basis for the space $S_m$). However, in the spline case there is a small positive probability that $\hat{\theta}$ in $\mathbb{R}^m$ does not exist. Indeed, considering nonnegative spline basis functions which are 0 except in part of the unit interval, it is seen that if there are no observations in the nonzero part of a basis function, then (2.2) cannot be satisfied by a density in the family. To illustrate, consider the case of splines of order $s = 1$ (piecewise constants). In this case, the maximum likelihood estimator of the density is the histogram with $m + 1$ equally spaced bins. If at least one of the bins is empty, then $\hat{\theta}$ in $\mathbb{R}^m$ does not exist and the relative entropy distance for the histogram is infinite. As noted by a referee, the probability that at least one of the bins is empty is bounded by $(m + 1)e^{-\varepsilon n/(m+1)}$ where $\varepsilon = \inf\{p(x)\colon 0 \le x \le 1\}$.

REMARK 4. For the histogram estimator (the spline case with $s = 1$), the result of the theorem is that $D(p\|\hat{p}_n)$ converges to 0 in probability at rate

$n^{-2/3}$ when $\log p$ has a square integrable derivative and $m$ is proportional to $n^{1/3}$.

REMARK 5. The spline methods saturate at rate $1/m^{2s} + m/n$, so that even if the log-density is infinitely differentiable, no faster rate of convergence than $n^{-2s/(2s+1)}$ can be obtained by choice of $m$. The rate $n^{-2r/(2r+1)}$ is achieved only with $s \geq r$. In contrast, the polynomial method does not have such saturation properties and convergence at rates close to $n^{-1}$ is possible. In particular, if the norm of the derivative of order $m$ grows no faster than a factorial, that is, $\|(\log p)^{(m)}\|_2 \leq cm!$ for some constant $c$, then with a choice of $m_n$ proportional to $\log n$, it is seen that $D(p\|\hat{p}_n) = O_{pr}(\log n)/n$ (see Section 7).

REMARK 6. Basic to our analysis is a decomposition of the relative entropy $D(p\|\hat{p})$ into the sum of two terms which correspond to approximation error and estimation error, respectively (analogous to the familiar bias and variance decomposition of mean squared error), and bounds are provided for both terms. The density $p^*$ in the exponential family which is closest to $p$ in the relative entropy sense is called the information projection [Csiszár (1975)]. It is characterized as the unique density in the family for which $\int \phi_k p^* = \alpha_k$ for $k = 1, \ldots, m$ (where $\alpha_k = \int \phi_k p$ denotes the expectation of the basis functions with respect to the true density) and it is also characterized by the Pythagorean-like relation $D(p\|p_\theta) = D(p\|p^*) + D(p^*\|p_\theta)$ valid for all densities $p_\theta$ in the exponential family. In particular, we have the decomposition

$$D(p\|\hat{p}) = D(p\|p^*) + D(p^*\|\hat{p}). \tag{2.6}$$

The first term $D(p\|p^*)$ is the approximation error: It converges to 0 at rate $m^{-2r}$ as $m \to \infty$ for log-densities in $W_2^r$. The second term $D(p^*\|\hat{p})$ is the estimation error for densities in the family: Under the right conditions, it converges to 0 in probability at rate $m/n$.

Now we state the general result on sequences of exponential families for which Theorem 1 is obtained as a special case. For $m \geq 1$, let $S_m$ be a linear space spanned by bounded and linearly independent functions $1, \phi_1(x), \ldots, \phi_m(x)$ on a measurable space $(\mathbf{X}, \mathbf{B})$. A random sample $X_1, \ldots, X_n$ is drawn from a distribution $P$ which has a density $p(x)$ with respect to a finite measure $\nu(dx)$. Let $\hat{p}_{n,m} = p_{\hat{\theta}}$ be the maximum likelihood density estimate in the regular exponential family $p_\theta(x) = \exp\{\sum_{k=1}^m \theta_k \phi_k(x) - \psi_m(\theta)\}$, where $\psi_m(\theta) = \log \int \exp\{\sum \theta_k \phi_k(x)\} \nu(dx)$, $\theta \in \mathbb{R}^m$. Let $\|\cdot\|_\infty$ and $\|\cdot\|_2$, respectively, denote the $L_\infty$ and $L_2$ norms with respect to $\nu$. The relative entropy is $D(p\|\hat{p}) = \int p(x) \log(p(x)/\hat{p}(x)) \nu(dx)$.

THEOREM 2. *For $S_m$, suppose there exists positive numbers $A_m$ such that $\|f_m\|_\infty \leq A_m \|f_m\|_2$ for all $f_m \in S_m$. For $f = \log p$ let*

$$\Delta_m = \|f - f_m\|_2 \tag{2.7}$$

*and*

$$\gamma_m = \| f - f_m \|_\infty \tag{2.8}$$

*be $L_2$ and $L_\infty$ degrees of approximation of $f$ by some $f_m \in S_m$.*

*If the sequence $\gamma_m$ is bounded and if $A_m \Delta_m \to 0$ as $m \to \infty$, then for all large $m$ the information projection $p_m^*$ exists, achieving the minimum $D(p\|p_m^*)$ for log-densities in $S_m$, and satisfies*

$$D(p\|p_m^*) = O(\Delta_m^2). \tag{2.9}$$

*Moreover, if $A_m\sqrt{m/n} \to 0$, then with probability tending to 1 as $n \to \infty$, the maximum likelihood estimator in the exponential family exists and satisfies*

$$D(p_m^*\|\hat{p}_{n,m}) \le O_{pr}\left(\frac{m}{n}\right), \tag{2.10}$$

$$D(p\|\hat{p}_{n,m}) \le O_{pr}\left(\Delta_m^2 + \frac{m}{n}\right). \tag{2.11}$$

REMARK 7. For the specialization of Theorem 2 to the context of Theorem 1, it is verified that for $\log p$ in $W_2^r$, $\Delta_m \doteq O(m^{-r})$ and $\gamma_m$ is bounded, in fact $\gamma_m \to 0$, by appropriate choice of $f_m$ in the polynomial, spline and trigonometric cases. The condition on $S_m$ is satisfied with $A_m = O(m)$ in the polynomial case and $A_m = O(\sqrt{m})$ in the spline and trigonometric cases (see Section 7). In this specialization, we have $\mathbf{X} = [0, 1]$, $\nu(dx) = p_0(x)\,dx$ and the density with respect to $\nu$ is $p(x)/p_0(x)$. If $\log p$ and $\log p_0$ are both in $W_2^r$, then so is $\log p/p_0$. Also, since $p_0$ is bounded away from 0 and $\infty$, the rates of approximation in $L_2(p_0)$ are the same as for $L_2$ with respect to Lebesgue measure.

REMARK 8. We note the relationship of our method for general exponential families to those developed by Cencov (1982), Portnoy (1988) and Stone (1989, 1990). The book by Cencov (1982) has a substantial treatment of sequences of exponential families. Cencov (1982), Section 28, examines compact subfamilies of the exponential families and shows that the maximum likelihood estimator of the density converges at a rate determined by the degree of approximation in the relative entropy sense. The compact subfamilies are assumed to satisfy a property of quasihomogeneity, meaning that uniformly for densities in the sequence of subfamilies, the relative entropy is bounded above and below by a constant times the $L_2$ distance between the logarithms of the densities. In contrast, we do not restrict the estimation to compact subfamilies and the full exponential family is not quasihomogeneous, so the results of Cencov do not directly apply to our setting.

Portnoy (1988) examined the asymptotics in exponential families of the Euclidean distance $\|\hat{\theta} - \theta\|$ and the log-likelihood ratio test statistic $D(p_{\hat{\theta}}\|p_\theta)$ under the assumption that the number of parameters tends to $\infty$. However, Portnoy assumed that the distribution for the random variables $X_i$ has a density function $p_\theta$ in the parametric family, that is, the bias term referred to above is 0. We prefer to not make such an assumption, since in that case the

distribution for the random variables would mysteriously hop from one exponential family to the next whenever we change $m$. Nevertheless, a key step in the proof, in particular Lemma 5 in Section 4, is based in part on an idea from Portnoy (1988), Theorem 2.1.

In independent work, Stone (1989, 1990) examines log-spline density estimation and determines rates of convergence of the density in $L_2$ and in $L_\infty$. The relative entropy and the information projection also play a key role in his analysis and some of the same inequalities are obtained. A difference is that much of his analysis is specific to splines and it is not clear to what extent his methods would extend to other linear spaces $S_m$.

In the following sections we develop some basic tools needed for the proof of the results.

**3. $L_2$ bounds on relative entropy.** Let $p(x)$ and $q(x)$ be two probability density functions with respect to a dominating measure $\nu(dx)$. Some quadratic bounds on the relative entropy are easily derived, e.g., $\int(\sqrt{p} - \sqrt{q})^2 \le D(p\|q) \le \int(p-q)^2/q$ [which follow from the slightly tighter bounds $-2\log \int\sqrt{pq} \le D(p\|q) \le \log \int p^2/q$ based on Jensen's inequality]. All integrals are understood to be with respect to the dominating measure. We require quadratic bounds in terms of the log-density. Such bounds are obtained for the case that $\|\log p/q\|_\infty$ is finite.

LEMMA 1.

$$D(p\|q) \ge \frac{1}{2}e^{-\|\log p/q\|_\infty}\int p\left(\log\frac{p}{q}\right)^2 \tag{3.1}$$

*and*

$$D(p\|q) \le \frac{1}{2}e^{\|\log p/q - c\|_\infty}\int p\left(\log\frac{p}{q} - c\right)^2, \tag{3.2}$$

*where c is any constant.*

REMARK. Since $D$ is an expected value of $\log p/q$, the fact that the bound is proportional to a *squared* norm of $\log p/q$ is surprising. The more obvious inequality only gives $D \le \sqrt{\int p(\log p/q)^2}$.

PROOF OF LEMMA 1. From the Taylor expansion of $e^z$ we have

$$\frac{z^2}{2}e^{-z_-} \le e^z - 1 - z \le \frac{z^2}{2}e^{z_+} \tag{3.3}$$

for $-\infty < z < \infty$, where $z_+ = \max\{z, 0\}$ and $z_- = \max\{-z, 0\}$.

To obtain the lower bound, let $f(x) = \log p(x)/q(x)$, then

$$\begin{aligned}\int p \log \frac{p}{q} &= \int \left( p \log \frac{p}{q} + q - p \right) \\ &= \int p(e^{-f} - 1 + f) \\ &\geq \int p \frac{f^2 e^{(-f)_-}}{2} \\ &\geq \frac{1}{2} e^{-\|f_+\|_\infty} \int p f^2,\end{aligned} \tag{3.4}$$

which yields inequality (3.1).

Now to obtain the upper bound, let $f(x) = \log p(x)/q(x) - c$, then

$$\begin{aligned}\int p \log \frac{p}{q} &= \int p(e^{-f} - 1 + f) + 1 + c - e^c \\ &\leq \int p \frac{f^2 e^{(-f)_+}}{2} \\ &\leq \frac{1}{2} e^{\|f_-\|_\infty} \int p f^2,\end{aligned} \tag{3.5}$$

which yields the desired inequality. □

We also need the following lemma.

LEMMA 2.

$$\int \frac{(p-q)^2}{p} \leq e^{2(\|f\|_\infty - c)} \int p \left( \log \frac{p}{q} - c \right)^2$$

*for any* $c$, *where* $f = \log p/q - c$.

PROOF. Use the fact that $|e^z - 1| \leq |z| e^{z_+}$ for $-\infty < z < \infty$ to get

$$\begin{aligned}\int (p-q)^2/p &= \int (q/p - 1)^2 p \\ &= e^{-2c} \int (e^{-f} - 1)^2 p - (e^{-c} - 1)^2 \\ &\leq e^{-2c} \int f^2 e^{2f_-} p \\ &\leq e^{2(\|f_-\|_\infty - c)} \int p f^2.\end{aligned}$$

□

**4. Information projection.** We adopt the framework given in the paragraph preceding the statement of Theorem 2. Thus the exponential family

takes the form $p_\theta(x) = e^{\theta\cdot\phi(x)-\psi(\theta)}$, where $\phi(x) = (\phi_1(x), \ldots, \phi_m(x))$, $\theta \cdot \phi = \sum_{k=1}^m \theta_k\phi_k$ and $\psi(\theta) = \log \int e^{\theta\cdot\phi(x)}\nu(dx)$, for $\theta \in \mathbb{R}^m$. The function $\psi(\theta)$ is clearly finite for all $\theta \in \mathbb{R}^m$, since the $\phi_k(x)$ are assumed to be bounded and $\nu$ is assumed to be a finite measure. Thus in the terminology of Brown (1986), page 2, the natural parameter space is $\mathbb{R}^m$ and the exponential family is regular. The linear independence of the functions $1, \phi_1, \ldots, \phi_m$ means that if $\sum\theta_k\phi_k - \sum\theta'_k\phi_k$ is constant almost everywhere then $\theta' = \theta$.

Let $C = \{p: \int\phi p = \alpha\}$ be the hyperplane of all density functions for which the expected value of $\phi(X)$ is equal to $\alpha$, where $\alpha \in \mathbb{R}^m$. It turns out that the set $C$ and the family $\{p_\theta: \theta \in \mathbb{R}^m\}$ are orthogonal in the sense that all members of the family have the same information projection onto $C$ denoted by $p^*$: that is, $p^*$ achieves $\min_{p\in C} D(p\|p_\theta)$ for each $\theta$ in $\mathbb{R}^m$. The following lemma recalls for convenience some of the projection properties [see also Csiszár (1975)]. We let $\Omega = \{\int\phi p_\theta: \theta \in \mathbb{R}^m\}$ and consider the equation

$$\int\phi p_\theta = \alpha. \tag{4.1}$$

LEMMA 3. *Suppose $\alpha \in \Omega$. Then the solution $\theta^* = \theta(\alpha)$ to (4.1) is unique. Moreover, for all $p \in C$ and $\theta \in \mathbb{R}^m$, a Pythagorean-like identity holds*

$$D(p\|p_\theta) = D(p\|p^*) + D(p^*\|p_\theta), \tag{4.2}$$

*where $p^* = p_{\theta^*}$. Consequently, $p^*$ is characterized as achieving $\min_p D(p\|p_0)$ subject to $p \in C$. Also, the parameter $\theta^*$ uniquely achieves $\min_\theta D(p\|p_\theta)$ for any $p \in C$ for which $D(p\|p_0)$ is finite. Also $F(\theta) = \theta\cdot\alpha - \psi(\theta)$ has a unique maximum at $\theta(\alpha)$.*

PROOF. Since the densities $p_\theta$ are positive we may write

$$\log\frac{p(x)}{p_\theta(x)} = \log\frac{p(x)}{p_{\theta^*}(x)} + \log\frac{p_{\theta^*}(x)}{p_\theta(x)}, \tag{4.3}$$

where $\theta^*$ is any solution to (4.1). Taking the expected value with respect to $p$ establishes the Pythagorean-like identity because the second term on the right side of (4.3) has the same expectation with respect to $p$ or $p^*$ (indeed this term is simply a linear combination of the $\phi_k$ so the expectation is the same for all densities in $C$). The remaining facts all immediately follow from this identity, since $D(p\|q)$ is strictly greater than 0, unless $p = q$ almost everywhere, and since maximization of $F(\theta)$ is the same as the minimization of $D(p^*\|p_\theta) = F(\theta^*) - F(\theta)$, so the proof is complete. □

Note that no derivatives need be taken to prove these facts. Also note that when $\alpha$ is replaced by the empirical average $\hat{\alpha}$, then $nF(\theta)$ is the log-likelihood function and $\hat{\theta} = \theta(\hat{\alpha})$ is the maximum likelihood estimator.

**5. Bounds within exponential families.** Here we give bounds on $D(p_{\theta_0}\|p_\theta)$ in terms of the Euclidean distance $\|\theta_0 - \theta\|$ for any $\theta_0, \theta$ in $\mathbb{R}^m$, and

we give bounds on $\|\theta(\alpha_0) - \theta(\alpha)\|$ in terms of $\|\alpha_0 - \alpha\|$. Since our ultimate interest is in the densities rather than the parameters, we are free to choose any convenient basis for the space $S_m$. In particular, for this section it is assumed that the functions $1, \phi_1, \ldots, \phi_m$ are chosen to be an *orthonormal* basis for $S_m$ with respect to a probability density function $q$. Here $q$ may be any density function for which $\log q$ is bounded.

Let $A_m = A_m(q) < \infty$ be such that for all $f_m \in S_m$:

$$\|f_m\|_\infty \le A_m \|f_m\|_{L_2(q)}. \tag{5.1}$$

First we relate distances between the densities in the parametric family to distances between the parameters. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^m$.

LEMMA 4. *For* $\theta_0, \theta \in \mathbb{R}^m$,

$$\|\log p_{\theta_0}/p_\theta\|_\infty \le 2A_m\|\theta_0 - \theta\|, \tag{5.2}$$

$$D(p_{\theta_0}\|p_\theta) \le \frac{b}{2} e^{A_m\|\theta_0-\theta\|}\|\theta_0 - \theta\|^2 \tag{5.3}$$

*and*

$$D(p_{\theta_0}\|p_\theta) \ge \frac{1}{2b} e^{-2A_m\|\theta_0-\theta\|}\|\theta_0 - \theta\|^2, \tag{5.4}$$

*where* $b = e^{\|\log q/p_{\theta_0}\|_\infty}$.

PROOF. Observe that

$$\psi(\theta) - \psi(\theta_0) = \log \int \exp\{(\theta - \theta_0)\cdot\phi(x)\} P_{\theta_0}(dx)$$

from which it follows that $|\psi(\theta) - \psi(\theta_0)| \le \|(\theta - \theta_0)\cdot\phi\|_\infty$. Now $\log p_{\theta_0}/p_\theta = (\theta_0 - \theta)\cdot\phi + \psi(\theta) - \psi(\theta_0)$ so it follows that $\|\log p_{\theta_0}/p_\theta\|_\infty \le 2\|(\theta - \theta_0)\cdot\phi\|_\infty \le 2A_m\|\theta_0 - \theta\|$ which gives (5.2). Using the assumed orthonormality of the $\phi_k$, the inequalities (5.3) and (5.4) follow from Lemma 1 with $c = \psi(\theta) - \psi(\theta_0)$, to complete the proof. □

Now for a key lemma. Recall that $\theta(\alpha)$ denotes the unique solution to $E_{p_\theta}\phi(X) = \alpha$ (whenever such a solution exists). We relate distances between the parameters $\theta$ to distances between the corresponding parameters $\alpha$.

LEMMA 5. *Let* $\theta_0 \in \mathbb{R}^m$, $\alpha_0 = \int \phi p_{\theta_0}$ *and* $\alpha \in \mathbb{R}^m$ *be given. Let* $b = e^{\|\log q/p_{\theta_0}\|_\infty}$ *and assume that* (5.1) *holds. If*

$$\|\alpha - \alpha_0\|_2 \le \frac{1}{4ebA_m}, \tag{5.5}$$

*then the solution* $\theta(\alpha)$ *to* $\int \phi p_\theta = \alpha$ *exists and satisfies*

$$\|\theta(\alpha) - \theta(\alpha_0)\| \le 2be^\tau \|\alpha - \alpha_0\|, \tag{5.6}$$

$$\|\log p_{\theta(\alpha_0)}/p_{\theta(\alpha)}\|_\infty \le 4be^\tau A_m \|\alpha - \alpha_0\|_2 \le \tau \tag{5.7}$$

*and*

$$D(p_{\theta(\alpha_0)} \| p_{\theta(\alpha)}) \le 2be^\tau \|\alpha - \alpha_0\|^2, \tag{5.8}$$

*for* $\tau$ *satisfying* $4ebA_m\|\alpha - \alpha_0\| \le \tau \le 1$.

In our application of this lemma, bounds which are adequate for identifying asymptotic rates may be obtained with $\tau = 1$; however, the smallest choice $\tau = 4ebA_m\|\alpha - \alpha_0\|$ yields tighter bounds for each $m$ as well as improved constants for the asymptotics.

PROOF OF LEMMA 5. Suppose $\alpha \neq \alpha_0$, since if $\alpha = \alpha_0$ the inequalities are trivial. Let $F(\theta) = \theta \cdot \alpha - \psi(\theta)$ as in Section 4. Then since $D(p_{\theta_0}\|p_\theta) = (\theta_0 - \theta) \cdot \alpha_0 + \psi(\theta) - \psi(\theta_0)$, we have that for all $\theta \in \mathbb{R}^m$:

$$\begin{aligned} F(\theta_0) - F(\theta) &= (\theta_0 - \theta) \cdot \alpha + \psi(\theta) - \psi(\theta_0) \\ &= D(p_{\theta_0}\|p_\theta) - (\theta_0 - \theta) \cdot (\alpha_0 - \alpha). \end{aligned} \tag{5.9}$$

It follows by Lemma 4 and the Cauchy–Schwarz inequality that for all $\theta \in \mathbb{R}^m$,

$$F(\theta_0) - F(\theta) \ge \frac{1}{2b} e^{-2A_m\|\theta_0 - \theta\|}\|\theta_0 - \theta\|^2 - \|\theta_0 - \theta\|\|\alpha_0 - \alpha\|. \tag{5.10}$$

This inequality is seen to be strict for $\theta \neq \theta_0$. Consider $\theta$ on the sphere $\{\theta: \|\theta - \theta_0\| = r\}$ where $r = 2e^\tau b\|\alpha - \alpha_0\|$. For all $\theta$ on this sphere

$$F(\theta_0) - F(\theta) > (e^{\tau - 4A_m e^\tau b\|\alpha - \alpha_0\|} - 1)2e^\tau b\|\alpha - \alpha_0\|^2. \tag{5.11}$$

The right side is nonnegative when $4ebA_m\|\alpha - \alpha_0\|_2 \le \tau \le 1$. Thus the value of $F$ at $\theta_0$ (inside the sphere) is larger than all the values $F(\theta)$ on the sphere. Consequently, $F$ has an extreme point $\theta^*$ which is inside the sphere, that is, $\|\theta^* - \theta_0\| < r$. The gradient of $F$ at $\theta^*$ must be zero which means that $\alpha - \int \phi p_{\theta^*} = 0$, that is, $\theta^* = \theta(\alpha)$. Therefore $\|\theta(\alpha) - \theta(\alpha_0)\|_2 < r$ which verifies (5.6) . Inequality (5.7) follows by applying Lemma 4. To verify (5.8), since $F(\theta(\alpha)) \ge F(\theta_0)$ it follows from (5.9) and (5.6) that

$$\begin{aligned} D(p_{\theta(\alpha_0)}\|p_{\theta(\alpha)}) &\le (\theta(\alpha_0) - \theta(\alpha)) \cdot (\alpha_0 - \alpha) \\ &\le \|\theta_0 - \theta\|\|\alpha_0 - \alpha\| \\ &\le 2be^\tau \|\alpha - \alpha_0\|^2. \end{aligned} \tag{5.12}$$

This completes the proof of Lemma 5. □

**6. Proof of the main result.** Here we give our result in terms of bounds for each $m$ and $n$ from which Theorem 2 is easily shown to follow. To yield

simpler expressions for the bounds, the result is stated in terms of the $L_2(p)$ norm instead of the $L_2(\nu)$ norm. The asymptotic equivalence follows from the assumed boundedness of $\log p$.

THEOREM 3. *Let $A_m = A_m(p)$ be such that $\|f_m\|_\infty \le A_m \|f_m\|_{L_2(p)}$ for all $f_m \in S_m$. For $f = \log p$ let*

$$\Delta_m = \|f - f_m\|_{L_2(p)} \tag{6.1}$$

*and*

$$\gamma_m = \|f - f_m\|_\infty \tag{6.2}$$

*be $L_2$ and $L_\infty$ degrees of approximation of $f$ by some $f_m \in S_m$. Set $\varepsilon_m = 4e^{4\gamma_m+1}A_m\Delta_m$ and $\delta_{m,n} = 4e^{2\gamma_m+2}A_m\sqrt{m/n}$.*

*If $\varepsilon_m \le 1$, the information projection $p_m^*$ exists [achieving the minimum $D(p\|p_m^*)$ for log-densities in $S_m$] and satisfies*

$$D(p\|p_m^*) \le C_1\Delta_m^2, \tag{6.3}$$

*where $C_1 = \frac{1}{2}e^{\gamma_m}$. Moreover, if $\delta_{m,n} \le 1$, then for every $\mathcal{K} \le \delta_{m,n}^{-2}$, there is a set of probability less than $1/\mathcal{K}$, such that outside this set, the maximum likelihood estimator in the exponential family exists and satisfies*

$$\begin{aligned} D(p_m^*\|\hat{p}_{n,m}) &\le C_2\frac{m}{n}\mathcal{K}, \\ D(p\|\hat{p}_{n,m}) &\le C_1\Delta_m^2 + C_2\frac{m}{n}\mathcal{K}, \end{aligned} \tag{6.4}$$

*where $C_2 = 2e^{2\gamma_m+\varepsilon_m+\tau}$ and $\tau = \delta_{m,n}\mathcal{K}^{1/2} \le 1$.*

Taking $\gamma_m$ to be a bounded sequence and assuming $A_m\Delta_m \to 0$ and $A_m\sqrt{m/n} \to 0$ (so that $\varepsilon_m$ and $\delta_{m,n}$ tend to 0), Theorem 2 is readily seen to follow from Theorem 3. If also $\gamma_m \to 0$, then asymptotically $C_1$ and $C_2$ approach $1/2$ and 2, respectively.

PROOF OF THEOREM 3. Choose $\phi(x) = (\phi_1(x), \ldots, \phi_m(x))$ so that $1, \phi_1, \phi_2, \ldots, \phi_m$ is a basis for $S_m$ which is orthonormal with respect to $p$. We divide the proof into two main tasks. The first task is to show that $\theta^*$ exists with $\int \phi p_{\theta^*} = \int \phi p$ and that $\log p/p_{\theta^*}$ is bounded by a constant. This $p_{\theta^*}$ is the information projection achieving the minimum $D(p\|p_\theta)$ for densities in the exponential family. The second task involves the examination of the terms $D(p_{\theta^*}\|p_{\hat{\theta}})$ and $D(p\|p_{\theta^*})$.

For the first task, let $f_m(x) = \sum_{k=0}^m \beta_k\phi_k(x)$ be the approximation of $f$ which is assumed to satisfy the given $L_2$ and $L_\infty$ bounds on the error $f - f_m$. Set $\alpha_0 = \int \phi p_\beta$, where $\beta = (\beta_1, \ldots, \beta_m)$ and set $\alpha = \int \phi p$. Then the entries in the vector $\alpha - \alpha_0$ are given by $\int((p - p_\beta)/p)\phi_k\, dP$ for $k = 1, \ldots, m$. These entries are seen to be coefficients in the $L_2(p)$ orthonormal projection of

$(p - p_\beta)/p$ onto $S_m$, so by Bessel's inequality and Lemma 3,

$$\begin{aligned} \|\alpha - \alpha_0\| &\le \|(p - p_\beta)/p\|_{L_2(p)} \\ &\le e^{\|f - f_m\|_\infty - (\beta_0 + \psi(\beta))} \|f - f_m\|_{L_2(p)} \\ &\le e^{2\gamma_m} \Delta_m, \end{aligned} \tag{6.5}$$

where we have used the fact that $|\psi(\beta) + \beta_0|$ is not greater than $\|f - f_m\|_\infty$. [Indeed $\psi(\beta) + \beta_0$ is seen to equal $\log \int e^{f_m(x) - f(x)} P(dx)$ from which the fact follows.] From this same fact it is seen that $\|\log p/p_\beta\|_\infty$ is not greater than $2\|f - f_m\|_\infty = 2\gamma_m$. Now apply Lemma 5 with $\theta_0 = \beta$, $q = p$, $\alpha = \int \phi p$ and $b = e^{\|\log p/p_\beta\|_\infty} \le e^{2\gamma_m}$. The condition (5.5) is satisfied if $e^{2\gamma_m}\Delta_m \le 1/(4ebA_m)$, that is, if $\varepsilon_m \le 1$. In which case we may conclude that $\theta^* = \theta(\alpha)$ exists and that $\|\log p_{\theta^*}/p_\beta\|_\infty \le \varepsilon_m$. So by the triangle inequality

$$\|\log p/p_{\theta^*}\|_\infty \le 2\gamma_m + \varepsilon_m. \tag{6.6}$$

Now for the second task, we show that $D(p_{\theta^*}\|p_{\hat\theta})$ is small with high probability. Lemma 5 is applied once more with different choices of the parameters. In particular, take $\theta_0$ to be $\theta^*$: The corresponding $\alpha_0$ is $\int \phi p^*$ (which is the same as $\int \phi p$). For $\alpha$ take $\bar\phi_n = (1/n)\sum_{i=1}^n \phi(X_i)$. [Whenever a solution to $\int \phi p_\theta = \bar\phi_n$ exists, we recognize this solution $\hat\theta = \theta(\bar\phi_n)$ as the maximum likelihood estimate.] With these choices $\|\alpha - \alpha_0\|^2 = \sum_{k=1}^m (\bar\phi_{n,k} - E_p\phi_k)^2$. Lemma 5 requires that this distance between $\alpha$ and $\alpha_0$ be not too large. By Chebyshev's inequality $\|\alpha - \alpha_0\|^2 \le \mathscr{K}m/n$ except in a set of probability which satisfies

$$\begin{aligned} P\left\{\sum_{k=1}^m (\bar\phi_{n,k} - E_P\phi_k)^2 > \frac{m}{n}\mathscr{K}\right\} &\le \frac{n}{m\mathscr{K}} E_p\left[\sum_{k=1}^m (\bar\phi_{n,k} - E_P\phi_k)^2\right] \\ &= 1/\mathscr{K}, \end{aligned} \tag{6.7}$$

where the last identity is due to the fact that $X_1, \ldots, X_n$ are independent with density $p$ and the functions $\phi_k(X)$ are normalized to have zero mean and unit variance with respect to $p$. Now apply Lemma 5 with $q = p$ and $b = e^{\|\log p/p_{\theta^*}\|_\infty} \le e^{2\gamma_m + \varepsilon_m}$. If $(\mathscr{K}m/n)^{1/2} \le 1/(4ebA_m)$, that is, if $\delta_{m,n}^2 \le 1/\mathscr{K}$, then except in the set above (which has probability less than $1/\mathscr{K}$), the conditions of the lemma are satisfied, whence the MLE $\hat\theta$ exists and

$$D(p_{\theta^*}\|p_{\hat\theta}) \le 2be^\tau \frac{m}{n}\mathscr{K} \le 2e^{2\gamma_m + \varepsilon_m + \tau}\frac{m}{n}\mathscr{K}. \tag{6.8}$$

Finally, by Lemma 3, the Kullback–Leibler loss decomposes into a sum of approximation error and estimation error terms:

$$D(p\|\hat p) = D(p\|p^*) + D(p^*\|\hat p). \tag{6.9}$$

The estimation error $D(p^*\|\hat p)$ has just been shown to be less than $C_2(m/n)\mathscr{K}$ except in a set of probability less than $1/\mathscr{K}$. By Lemmas 1 and 3, the

approximation error satisfies

$$(6.10)\qquad D(p\|p^*) \le D(p\|p_\beta) \le \frac{1}{2}e^{\|f-f_m\|_\infty}\|f-f_m\|_2^2 \le \frac{1}{2}e^{\gamma_m}\Delta_m^2.$$

This completes the proof of the theorem. □

**7. Verification of the details.** In this section, it is shown how the conditions on $A_m$, $\Delta_m$ and $\gamma_m$ are satisfied in the polynomial, spline and trigonometric cases. For the approximations here the $L_2$ space is taken with respect to Lebesgue measure on $[0, 1]$. Bounds for the $L_2(p)$ formulation, as needed for Theorem 3, then follow using the assumption that $p$ is bounded away from 0 and $\infty$.

Given a class of functions $S_m$ and a density $q$, we denote $A_m(q) = \sup\{\|g\|_\infty/\|g\|_{L_2(q)}\colon g \in S_m\}$. Note that $A_m(p) \le \|q/p\|_\infty^{1/2}A_m(q)$. In this section, when $A_m$ is written without an argument $q$, it is with respect to Lesbesgue measure on $[0, 1]$. For polynomials and splines, the following lemma is used to bound $A_m$.

LEMMA 6. *If $g(x)$ is a polynomial of degree less than or equal to $d$ on $[a, b]$, then*

$$(7.1)\qquad \sup_{x\in[a,b]}|g(x)| \le (d+1)\left(\frac{1}{b-a}\right)^{1/2}\left(\int_a^b g^2\right)^{1/2},$$

*and there exist polynomials of degree $d$ on $[a, b]$ for which equality is achieved.*

REMARK 1. In the polynomial case, the lemma applies with $d = m$ and $[a, b] = [0, 1]$ to show that $A_m = m + 1$ and hence $A_m(p) \le (m + 1)\|1/p\|_\infty^{1/2}$.

REMARK 2. In the case of splines $g$ of order $s$ with knots at $\Delta, 2\Delta, \ldots, 1 - \Delta$, the lemma applies with $d = s - 1$ to each of the polynomial pieces to yield

$$(7.2)\qquad \begin{aligned}\sup_{x\in[0,1]}|g(x)| &\le \max_{j=1,2,\ldots,1/\Delta} s\left(\frac{1}{\Delta}\right)^{1/2}\left(\int_{(j-1)\Delta}^{j\Delta} g^2(x)\,dx\right)^{1/2}\\ &\le s\left(\frac{1}{\Delta}\right)^{1/2}\left(\int_0^1 g^2(x)\,dx\right)^{1/2}.\end{aligned}$$

Setting $\Delta = 1/(m - s + 2)$, this shows that $A_m \le s\sqrt{m - s + 2}$ and $A_m(p) \le s\sqrt{m - s + 2}\,\|1/p\|_\infty^{1/2}$ in the spline case.

PROOF OF LEMMA 6. First note that by scaling the polynomials it suffices to prove the result for $[a, b] = [0, 1]$. Let $\phi_k(x)$, $k = 0, 1, \ldots, d$, be the orthonormal Legendre polynomials which are bounded in absolute value by $\sqrt{2k + 1}$. This bound is achieved for each $k$ at $x = 0$ and $x = 1$ [see Jackson (1930),

page 25]. Summing the squares of the bounds yields $\max_x \sum_{k=0}^{d} (\phi_k(x))^2 = (d+1)^2$. If $g$ is a polynomial of degree $d$, then $g(x) = \sum_{k=0}^{d} \beta_k \phi_k(x)$ for some coefficients $\beta_k$. By the Cauchy–Schwarz inequality

$$|g(x)| \le \left( \sum_{k=0}^{d} \phi_k^2(x) \right)^{1/2} \left( \sum_{k=0}^{d} \beta_k^2 \right)^{1/2} \le (d+1) \left( \int_0^1 g^2 \right)^{1/2} \tag{7.3}$$

uniformly for $x$ in $[0, 1]$. Equality is achieved at $x = 0$ and $x = 1$ for polynomials with coefficients $\beta_k$ proportional to $\sqrt{2k+1}$. This completes the proof of Lemma 6. □

Now we examine the $L_2$ and $L_\infty$ approximation properties of polynomials, splines and trigonometric series. Approximation rate results are available in the literature [e.g., Schumaker (1981)], giving the best $L_2$ and $L_\infty$ rates of approximation for functions in the Sobolev spaces. In particular, the Sobolev space $W_2^r$ readily yields $L_2$ bounds on the best $L_2$ approximation. Our requirements are slightly complicated by the fact that we also need to bound the $L_\infty$ error of the $L_2$ approximation (rather than the best uniform approximation) assuming only that the function is in $W_2^r$ (rather than $W_\infty^r$). Also, in the polynomial case, we desire accurate bounds for very smooth functions, which permit us to let $r = m$ grow with the dimension of the approximation, and thereby obtain faster rates of convergence in this case.

*Polynomials.* It is convenient to use a recent result of Cox (1988) which we briefly summarize. First we fix $r \ge 1$. Let $\phi_k(x)$, $k = 0, 1, \ldots$, denote the normalized Legendre polynomials which are orthonormal with respect to the uniform weight function on $[0, 1]$. The system of derivatives $\{\phi_k^{(r)}: k \ge r\}$ is orthogonal with respect to the weight function $(x(1-x))^r$ on $[0, 1]$ and has normalizing constants

$$c_k^2 = \int_0^1 \left(\phi_k^{(r)}(x)\right)^2 (x(1-x))^r \, dx = (k+r)!/(k-r)!.$$

Consequently, if $f$ is in $W_2^r$ with Legendre coefficients $\beta_k$, then the sum $\sum_{k \ge r} c_k^2 \beta_k^2$ is equal to the squared norm $\int_0^1 (f^{(r)}(x))^2 (x(1-x))^r \, dx$ which is not greater than $(1/4)^r \int (f^{(r)}(x))^2 \, dx$. Let $f_m(x) = \sum_{k=0}^{m} \beta_k \phi_k(x)$. Then for $m \ge r$,

$$\begin{aligned} \| f - f_m \|_2^2 &= \sum_{k=m+1}^{\infty} \beta_k^2 \\ &\le \frac{1}{c_{m+1}^2} \sum_{k=m+1}^{\infty} c_k^2 \beta_k^2 \\ &\le \frac{1}{(m+r+1) \cdots (m-r+2)} \left( \frac{1}{4} \right)^r \| f^{(r)} \|_2^2. \end{aligned} \tag{7.4}$$

The first inequality in (7.4) follows from the monotonicity of the sequence

$c_k^2 = (k + r) \cdots (k - r + 1)$ with increasing $k$. Thus $\| f - f_m \|_2 = O(1/m)^r$ for $f \in W_2^r$ and explicit constants are identified. Note that since $\lim_m \sum_{k>m} c_k^2 \beta_k^2 = 0$, we in fact have that $\| f - f_m \|_2 = o(1/m)^r$; however, this improved rate is not uniform for densities in a Sobolev ball.

To bound the $L_\infty$ error for the Legendre approximation, assuming only that $\| f^{(r)} \|_2 < \infty$, we apply the Cauchy–Schwarz inequality to the series $\sum \beta_k \phi_k(x)$ and use the bound $|\phi(x)| \le \sqrt{2k + 1}$. For $r > 1$, it is seen that the Legendre series is absolutely convergent with error bounded for $m \ge r$ by

$$
\begin{aligned}
|f(x) - f_m(x)| &\le \left( \sum_{k=m+1}^{\infty} \frac{2k+1}{c_k^2} \right)^{1/2} \left( \sum_{k=m}^{\infty} c_k^2 \beta_k^2 \right)^{1/2} \\
&\le \left( \sum_{k=m+1}^{\infty} \frac{2e^{2r}}{(k+r)^{2r-1}} \right)^{1/2} \left( \frac{1}{2} \right)^r \| f^{(r)} \|_2 \qquad (7.5) \\
&\le \frac{e^r}{(r-1)^{1/2}(m+r)^{r-1}} \left( \frac{1}{2} \right)^r \| f^{(r)} \|_2 .
\end{aligned}
$$

Here we have used the inequality $c_k^2 \ge (k + r)^{2r} e^{-2r}$ (which may be deduced by comparing the sum $\sum_{j=k-r+1}^{k+r} \log j$ to the integral $\int_{k-r}^{k+r} \log x \, dx$) as well as the inequality for the sum $\sum_{k>m} (k + r)^{-2r+1} \le (2(r - 1))^{-1} (m + r)^{-2(r-1)}$ (which is also deduced by comparing the sum to an integral). Consequently, $\gamma_m = \| f - f_m \|_\infty = O(1/m)^{r-1}$ for $f \in W_2^r$. [An alternative proof of this rate can be obtained by deriving that $f^{(r-1)}$ has modulus of continuity $w(\delta) \le \delta^{1/2} \| f^{(r)} \|_2$ and then applying bounds from Jackson (1930), page 31, with the refinement that Jackson credits to Gronwall (1913).] This completes the details for the polynomial case.

*Splines.* Let $S_m$ be the space of splines of order $s$ on $[0, 1]$ with knots spaced with equal widths $\Delta = 1/(m - s + 2)$. Fix $s$ and consider $m \ge s$. We use the results of De Boor and Fix (1973), where the same approximating spline function $f_m$ is used for both the $L_2$ and $L_\infty$ approximation. It is assumed that $f$ is in $W_2^r$ for some $1 \le r \le s$. By De Boor and Fix (1973), Theorem 5.2, $\| f - f_m \|_2 \le K \Delta^r \| f^{(r)} \|_2$, where $K$ is an absolute constant. Thus $\| f - f_m \|_2 = O(1/m^r)$. Now $f^{(r-1)}$ is continuous with modulus of continuity not greater than $\Delta^{1/2} \| f^{(r)} \|_2$. [Indeed if $|x - y| \le \Delta$, then $|f^{(r-1)}(x) - f^{(r-1)}(y)| = |\int_x^y f^{(r)}(z) \, dz|$ which is less than $\Delta^{1/2} \| f^{(r)} \|_2$ by the Cauchy–Schwarz inequality.] So by De Boor and Fix (1973), Theorem 2.1, $\| f - f_m \|_\infty \le K' \Delta^{r-1/2} \| f^{(r)} \|_2$, where $K'$ is an absolute constant. Thus $\gamma_m = O(1/m)^{r-1/2}$. This completes the details for the spline case.

*Trigonometric series.* The $m + 1$ term truncated Fourier series represents functions of the form

$$
f_m(x) = \beta_0 + \sum_{k=1}^{m/2} (\beta_{2k} \phi_{2k}(x) + \beta_{2k+1} \phi_{2k+1}(x)),
$$

where $\phi_0 = 1$, $\phi_{2k}(x) = \sqrt{2}\cos(2\pi kx)$ and $\phi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$ for $0 \le x \le 1$. (For simplicity we focus on the case that $m$ is even.) For functions $f \in W_2^r$ which satisfy the boundary conditions, a familiar calculation shows that $\|f^{(r)}\|_2^2 = \sum_{k=1}^{\infty}(2\pi k)^{2r}(\beta_{2k}^2 + \beta_{2k+1}^2)$, where the $\beta_k$ are the Fourier coefficients of $f$. Consequently, the Fourier series approximation has $L_2$ error $\|f - f_m\|_2 \le (\pi(m+2))^{-r}\|f^{(r)}\|_2$. Similarly, applying the Cauchy–Schwarz inequality, it is seen that the Fourier series is absolutely convergent, with error $|f(x) - f_m(x)|$ bounded by $(\sum_{k>m/2}(2\pi k)^{-2r})^{1/2}(\sum_{k>m/2}2(2\pi k)^{2r}(\beta_{2k}^2 + \beta_{2k+1}^2))^{1/2}$ which is not greater than $(2r-1)^{-1/2}m^{-(r-1/2)}\pi^{-r}\|f^{(r)}\|_2$. Thus $\|f - f_m\|_\infty \le O(m^{-(r-1/2)})$ for $f$ in $W_2^r$. [An alternative method of bounding the $L_\infty$ error, using the modulus of continuity of $f^{(r-1)}$ and applying the theorem of Jackson (1930), page 22, Corollary 4, yields the slightly worse but also satisfactory rate $\|f - f_m\|_\infty \le O(m^{-(r-1/2)}\log m)$.]

To determine $A_m$ for the trigonometric case, we see by the Cauchy–Schwarz inequality and the identity $\cos^2 + \sin^2 = 1$, if $f_m = \sum_{k=0}^{m}\beta_k\phi_k(x)$, then

$$|f_m(x)| \le \left(\sum_{k=0}^{m}\phi_k^2(x)\right)^{1/2}\left(\sum_{k=0}^{m}\beta_k^2\right)^{1/2} = (m+1)^{1/2}\|f_m\|_2 \tag{7.6}$$

uniformly in $[0,1]$. Given any $x_0$ in $[0,1]$, equality in (7.6) is achieved at $x = x_0$ when the coefficients $\beta_k$ are proportional to $\phi_k(x_0)$. It follows that $A_m = \sqrt{m+1}$.

This completes the approximation details needed for the asymptotics stated in Theorem 1. Note that by using Theorem 3 and assuming bounds on $\|\log p\|_\infty$ and $\|(\log p)^{(r)}\|_2$, explicit bounds are obtained which are applicable for each finite value of $m$ and $n$, subject to $\varepsilon_m$ and $\delta_{m,n} \le 1$.

*Approximation of very smooth functions.* We return to the polynomial case and deduce bounds in the case that $f \in W_2^r$. By (7.4) and (7.5) with $r = m$,

$$\|f - f_m\|_2 \le \left(\frac{1}{(2m+1)!}\right)^{1/2}\left(\frac{1}{2}\right)^m\|f^{(m)}\|_2 \tag{7.7}$$

and

$$\|f - f_m\|_\infty \le \frac{e}{2(m-1)^{1/2}(4m/e)^{m-1}}\|f^{(m)}\|_2. \tag{7.8}$$

For instance, if $m = 4$ and if the fourth derivative of $f$ has $L_2$ norm bounded by $10m! = 240$, then the Legendre approximation has $L_2$ error not greater than $240/(16\sqrt{9!}) = 0.0000413$.

Suppose $f = \log p$ is an infinitely differentiable on $[0,1]$ and that the sequence of derivatives $f^{(m)}$ have $L_2$ norms which do not grow faster than a factorial: that is, $\|f^{(m)}\|_2 \le cm!$ for some constant $c$. From Stirling's formula

it is seen that $m!/\sqrt{(2m+1)!} \le (1/2)^m$ and, for $m > 1$,

$$\|f - f_m\|_2 \le c\left(\frac{1}{4}\right)^m, \tag{7.9}$$

$$\|f - f_m\|_\infty \le 4\sqrt{\pi c}\, m\left(\frac{1}{4}\right)^m. \tag{7.10}$$

In this case, a consequence of Theorem 2 is that if $m_n = (\log n)/(\log 4)$ then the relative entropy distance converges to 0 in probability at rate

$$D(p\|\hat{p}_n) = O_{pr}\left(\frac{\log n}{n}\right). \tag{7.11}$$

This verifies the claim in Remark 5 of Section 2.

The practical implication is that the order of the polynomial need not be chosen very large to get an accurate approximation whenever the log-density is sufficiently smooth.

**8. Example.** The density estimator is illustrated using data on the eruption lengths (in minutes) of 107 eruptions of the Old Faithful geyser as tabulated in Silverman (1986), page 8. Using an exponential family with a polynomial of degree 4 on $[1, 5]$, we obtain the density estimate plotted in Figure 1. The reference density $p_0$ is taken to be uniform on $[1, 5]$. The computations were obtained using a program by Gayle Nygaard which per-

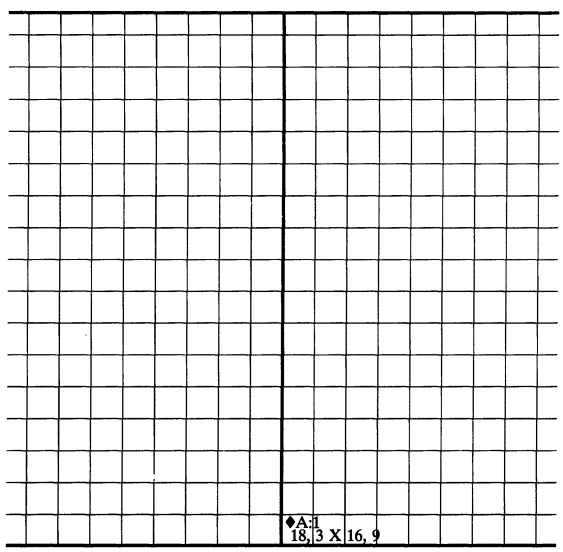


FIG. 1. *Exponential family estimate for Old Faithful Geyser data using a polynomial of degree* 4.

forms Newton's algorithm to maximize the likelihood. To avoid numerical overflow problems in the parameter search, we found it advisable to scale the data to the interval $[-1, 1]$ and to use the Legendre polynomial basis. The answer is then scaled back to the original interval.

The degree 4 of the polynomial is chosen to capture the bimodal shape of the density. Visually, our estimate is somewhat comparable to the kernel estimate shown in Silverman (1986), page 17. [For other estimates based on the same data see pages 9, 13 and 20 of Silverman (1986).] A difference is that the kernel estimate has noticeably broader peaks, due to the spreading of the empirical distribution caused by the convolution with a kernel of width $h = 0.25$. In contrast, our estimate agrees with the empirical distribution in mean, variance, skew and kurtosis. Other plots illustrating the polynomial and spline cases are in Mead and Papanicolaou (1984) and Stone and Koo (1986).

## REFERENCES


BARRON, A. R. (1991). Universal approximation bounds for superpositions of a sigmoidal function. Technical Report 58, Dept. Statistics, Univ. Illinois.

BARRON, A. R. and BARRON, R. L. (1988). Statistical learning networks: A unifying view. In *Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface* 192–203. Amer. Statist. Assoc., Alexandria, Va.

BARRON, A. R. and COVER, T. M. (1988). A bound on the financial value of information. *IEEE Trans. Inform. Theory* **34** 1097–1100.

BARRON, A. R. and COVER, T. M. (1991). Minimum complexity density estimation. *IEEE Trans. Inform. Theory*. To appear.

BARRON, A. R., GYÖRFI, L., and VAN DER MEULEN, E. C. (1991). Distribution estimation convergent in total variation and in information divergence. *IEEE Trans. Inform. Theory*. To appear.

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densités: Risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137.

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families*. IMS, Hayward, Calif.

CENCOV, N. N. (1982). *Statistical Decision Rules and Optimal Inference*. Amer. Math. Soc. Transl. **53**. Providence, R.I.

COX, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

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 condition. *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.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318.

CSISZÁR, I. (1975). $I$-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.

CSISZÁR, I. (1984). Sanov property, generalized $I$-projection and a conditional limit theorem. *Ann. Probab.* **12** 768–793.

CSISZÁR, I. (1989). Why least squares and maximum entropy? An axiomatic approach to inverse problems. Preprint 19, Mathematics Inst. Hungarian Academy of Sciences.

DAVISSON, L. D. (1973). Universal noiseless coding. *IEEE Trans. Inform. Theory* **19** 783–795.

DE BOOR, C. and FIX, G. J. (1973). Spline approximation by quasiinterpolants. *J. Approximation Theory* **8** 19–45.

DEVROYE, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The $L_1$ View*. Wiley, New York.

EFROIMOVICH, S. Y. and PINSKER, M. S. (1983). Estimation of square-integrable probability density of a random variable. *Problems Inform. Transmission* **18** 175–189.

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of series. *Ann. Statist.* **10** 401–414.

GOOD, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

GRONWALL, T. H. (1913). On the degree of convergence of Laplace's series. *Trans. Ann. Math. Soc.* **1** 1–30.

HALL, P. (1987). On Kullback–Leibler loss and density estimation. *Ann. Statist.* **15** 1491–1519.

HAUGHTON, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16** 342–355.

JACKSON, D. (1930). *The Theory of Approximation*. Amer. Math. Soc., New York.

JAYNES, E. T. (1957). Information theory and statistical mechanics. I. *Phys. Rev.* **106** 620–630.

JONES, L. K. (1989). Approximation-theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge. *SIAM J. Appl. Math.* **49** 650–661.

KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

KULLBACK, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Inform. Theory* **13** 126–127.

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

MEAD, J. R. and PAPANICOLAOU, N. (1984). Maximum entropy in the problem of moments. *J. Math. Phys.* **25** 2404–2417.

NADARAYA, E. A. (1974). On the integral mean square error of some nonparametric estimates for the density function. *Theory Probab. Appl.* **19** 133–141.

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

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

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic, Orlando, Fla.

RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

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

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHANNON, C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

SHORE, J. E. and JOHNSON, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory.* **26** 26–37.

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

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

STONE, C. J. (1989). Uniform error bounds involving log-spline 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. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18** 717–741.

STONE, C. J. and KOO, C.-Y. (1986). Logspline density estimation. In *Contemporary Mathematics* **59** 1–15. Amer. Math. Soc., Providence, R.I.

YU, B. and SPEED, T. P. (1990). Stochastic complexity and model selection II. Histograms. Technical Report 241, Dept. Statistics, Univ. California, Berkeley.

VAN CAMPENHOUT, J. M. and COVER, T. M. (1981). Maximum entropy and conditional probability. *IEEE Trans. Inform. Theory* **27** 483–489.

DEPARTMENTS OF STATISTICS AND
ELECTRICAL AND COMPUTER ENGINEERING
UNIVERSITY OF ILLINOIS
725 SOUTH WRIGHT STREET
CHAMPAIGN, ILLINIOS 61820