# DENSITY ESTIMATION WITH KULLBACK-LEIBLER LOSS

BY

CHYONG-HWA SHEU

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THE GRADUATE COLLEGE

July 27, 1989

WE HEREBY RECOMMEND THAT THE THESIS BY

CHYONG-HWA SHEU

ENTITLED DENSITY ESTIMATION WITH KULLBACK-LEIBLER LOSS

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF PHILOSOPHY

Andrew R Barron
Director of Thesis Research

Jerome Sacks
Head of Department

Committee on Final Examination†

Andrew R Barron
Chairperson

John I Marden

Douglas G. Simpson

† Required for doctor's degree but not for master's.

O-517

# DENSITY ESTIMATION WITH KULLBACK-LEIBLER LOSS

BY

CHYONG-HWA SHEU

B.S., National Kaohsiung Teacher's College, 1979
M.A., Eastern Illinois University, 1984

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1990.

Urbana, Illinois

## Abstract

Probability density functions are estimated by the method of maximum likelihood in sequences of regular exponential families. The approximation families of log-densities that we consider are polynomials, splines, and trigonometric series. Bounds on the relative entropy (Kullback-Leibler number) 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.

The relative entropy risk between true probability density function and the estimator is shown to converge to zero at a desired rate. The idea is to select n samples from the true distribution and choose the estimator which is the maximum posterior likelihood estimator in certain regular m-parameter exponential families, given that a Gaussian distribution is the prior on the parameter space. The implications for universal source coding and portfolio selection are discussed.

## Acknowledgements

I owe a special debt of gratitude to my thesis advisor Professor Andrew R. Barron. Without his guidance, I would not have been able to go beyond observation to learn the beauty of systematic reasonings. I have appreciated his stimulating enthusiasm and his support. Professor Barron was not only my academic advisor but more importantly my teacher in every aspect of learning.

I would like to express my sincere appreciation to my committee members, Professor John Marden, Professor Douglas Simpson, and Professor Zhiliang Ying, for the comments on the draft of this thesis.

Finally, I thank our Father in heaven for leading me to Champaign-Urbana. Being here has had a great impact in my life. I especially thank my wife. She is my best friend and closest discipleship partner.

Table of Contents

# CHAPTER 1 OVERVIEW

## 1.1. Introduction

This thesis describes some developments of density estimation in Kullback-Leibler loss. We give some background and informal statements of the main results in Chapter 1. The exponential family method is discussed in the following chapter. We consider some essential properties of information projection in Section 2.2 and $L_2$ bounds on the Kullback-Leibler number are calculated in Section 2.3. In Section 2.4, it is shown that the Kullback-Leibler number between the true density function and the estimator converges to zero in probability at rate $n^{-2r/(2r+1)}$. The verification of the details is presented in Section 2.5.

In Chapter 3, the main theorem there shows that the expected value of the Kullback-Leibler number between the true probability density function $p(x)$, which is assumed to satisfy some smoothness condition, and the estimator $\hat{p}_n(x)$ is shown to converge to zero at rate $n^{-2r/(2r+1)}$. We choose the estimator to be the maximum posterior likelihood estimator in certain regular m-parameter exponential families, given that a Gaussian distribution is the prior on the parameter space, with $m = n^{1/(2r+1)}$ where $n$ is the sample size. This result has direct impact on applications which are discussed in Chapter 5.

We concentrate on the estimation of a density underlying a set of univariate observations in the first three chapters. However, many of the important applications of density estimation involve the analysis of multivariate data. We will discuss the estimation of multivariate densities in Chapter 4. Some $L_2$ and $L_\infty$ assumptions will also be examined there.

In Chapter 5, we apply the results of the previous chapters to universal coding theory and portfolio selection in investment theory. We first introduce some basic ideas of coding theory and portfolio selection there and then some authors' works will be discussed in detail. We also show how these results can be applied to those areas. Some ideas for future work is introduced there.

During the past several years parametric modeling has been a subject of investigation by various workers. A disadvantage of parametric modeling is that it may not be robust in the sense that slight contamination of the data by observations not following the particular parametric family might lead to erroneous conclusions. Further, the data might be of such a type that there is no suitable parametric family that gives a good fit. Under these circumstances, one might take recourse to nonparametric modeling.

Rosenblatt (1956) and Parzen (1962) introduced the concept of non-parametric density estimation with their kernel density estimator. This estimate is obtained by convolving the sample distribution with a kernel function. Since the appearance of these papers several methods have been developed for the nonparametric estimation of density functions.

Kullback, S. and Leibler, R. A. (1951) pointed out the Kullback-Leibler information number and its relationship with statistics. The Kullback-Leibler number has several properties which make it a natural choice as a loss function in the decision theory framework. Let $X_1, X_2, \ldots, X_n$ be independent random variables with unknown probability density function $p(x)$ with respect to a known dominating measure $\lambda(dx)$. The asymptotics of density estimators is considered in terms of the Kullback-Leibler number (relative entropy)

$$D(p \| \hat{p}) = \int p(x) \log \frac{p(x)}{\hat{p}(x)} \lambda(dx). \tag{1.1}$$

It is well known that $D$ is non-negative and equals zero if and only if $p=\hat{p}$ a.e. Also $D(p \| \hat{p}) \geq (1/2)(\int |p-\hat{p}|)^2$. Inequalities in Section 2.3 show that $D$ behaves like a squared $L_2$ norm between the logarithms of the densities.

Popular methods such as kernel density estimation and orthogonal series density estimation have advantageous (in some cases optimal) asymptotic properties when measured by either $L_1$ or $L_2$ distances between density functions, see e.g. Devroye and Gyorfi (1983), Nadaraya (1974), Efroimovich and Pinsker (1983). Wahba (1975) establishes that spline, kernel, and orthogonal series estimators posses optimal rates of convergence for the mean squared error at an arbitrary point $x$, assuming that the density satisfies a Sobolev condition. However, for these loss functions, the kernel and series density estimators which achieve the fastest rates of convergence are not necessarily strictly positive (indeed they are sometimes negative, even in the support of $p$) in which case $D(p \| \hat{p}) = \infty$. This is the motivation to find a new approach to overcome the problem. We will roughly introduce some methods in the next section.

## 1.2 Histogram and Kernel Smoothing

The oldest and most widely used density estimator is the histogram. Let $p(x)$ be an unknown probability density function on the unit interval [0,1] which is to be estimated from a sample of independent random variables $X_1, X_2, \cdots, X_n$ each drawn from $P$.( In practice these random variables on [0,1] might be obtained by transformation $X = G(Z)$ of random variables $Z_i$ which have arbitrary support on the line. One reasonable transformation is to let $G$ be the cumulative distribution function corresponding to an initial guess $g(z)$ of the true density function $f(z) = p(G(z))g(z)$.) The usual histogram density estimator of $p(x)$ is

$$\hat{p}_n(x) = m\frac{N_j}{n} \quad \text{for } \frac{j-1}{m} < x \leq \frac{j}{m}$$

where $N_j$ is the number of $X_i$ in the interval $((j-1)/m, j/m]$, for $j = 1,2,\ldots,m$. Let $\pi_m$ denote the partition of $(0,1]$ into these intervals. The number of cells $m$ is allowed to depend on the sample size $n$. We also call $1/m$ the bin width.

Abou-Jaoude (1976) has shown that the histogram is consistent in $L_1$ distance for any density if and only if $m \to \infty$ and $m/n \to 0$ as the sample size tends to infinity ( i.e. $\lim_{n\to\infty} E\int |p(x)-\hat{p}_n(x)|dx = 0$ ). Under appropriate restrictions on the density function other modes of convergence such as $L_2$ are also well known. However, for some applications in information theory it is necessary to measure the accuracy of the density estimate by the relative entropy. Unfortunately, the histogram estimator $\hat{p}_n(x)$ is zero in cells which are empty. Moreover, there is positive probability that at least one cell is empty. Therefore, the expected relative entropy for the histogram is infinite.

To rectify this deficiency of histograms, Barron (1987) found the following modification

$$\hat{p}_n(x) = m\frac{N_j+1}{n+m} \quad \text{for } \frac{j-1}{m} < x \leq \frac{j}{m}.$$

This modified histogram estimator is the predictive density $P^{Bayes}(x_{n+1} \mid x_1, \ldots, x_n)$ ( evaluated at $x_{n+1} = x$ ) for a Bayesian who presumes that the cell probabilities have a Dirichlet $(1,1,\ldots,1)$ prior distribution. It will be shown that the expected Kullback-Leibler number is bounded by two terms. In the derivation the two terms $D(p \| p_m)$ and $(m-1)/(n+1)$ are analogous to the usual squared-bias and variance terms in a traditional mean squared error analysis of

density estimates. It will be shown that as $m \to \infty$ the first term $D(p \| p_m)$ tends to zero if and only if the entropy $-\int p \log p$ is finite. We state the results as follows.

**Theorem 1.1:** *Let $\hat{p}_n$ be the modified histogram estimator, for all $n, m \geq 1$*

$$ED(p \| \hat{p}_n) \leq D(p \| p_m) + \frac{m-1}{n+1} \tag{1.2}$$

*where $p_m(x) = m \int_{j-1/m}^{j/m} p(y)dy$ for $\frac{j-1}{m} < x \leq \frac{j}{m}$ is the " theoretical " equivalent of histogram. Moreover, as $m \to \infty$ and $m/n \to 0$ $ED(p \| \hat{p}_n) \to 0$ if and only if $-\int p \log p$ is finite.*

**Proof:**

$$ED(p \| \hat{p}_n) = E \int p \log \frac{p}{\hat{p}_n} \tag{1.3}$$

$$= \int p \, E \log \frac{p}{\hat{p}_n}$$

$$\leq \int p \log p E \frac{1}{\hat{p}_n} \tag{1.4}$$

$$\leq \int p \log \left(\frac{p}{p_m} \frac{n+m}{n+1}\right) \tag{1.5}$$

$$= \int p \log \frac{p}{p_m} + \log \left(1 + \frac{m-1}{n+1}\right)$$

$$\leq D(p \| p_m) + \frac{m-1}{n-1}. \tag{1.6}$$

The exchange of the expectation and integration is by the Fubini-Tonelli Theorem applied to $(p \log p/\hat{p}_n) + \hat{p}_n - p$, which is nonnegative and has the

same

integral as (1.3). The inequality (1.4) is by the concavity of the logarithm. Finally, (1.5) follows from an inequality for the binomial distribution.

Let $u$ be the uniform [0,1] density. Then $\int_0^1 p \log p = D(p \| u)$. Wo show that $D(p \| p_m) \to 0$ if and only if $D(\ \| u)$ is finite. This follows from the identity

$$D(p \| p_m) = D(p \| u) - D(p_m \| u)$$

$$= D(p \| u) - D_{\pi_m}(p \| u),$$

where $D_\pi$ denotes the divergence restricted to a partition. It is known that $D_{\pi_m}$ converges to $D(p \| u)$(Pinsker 1964). So if $D(p \| unif)$ is finite, then $D(p \| p_m) \to 0$. This completes the proof. □

**Remark:** $p_m$ is the information projection. In the other words, the density closest to $p$ in information divergence among all histogram shaped densities.

Barron proved this theorem in the class which is nonparametric density estimation and pattern recognition in 1987. As for the rate of convergence, suppose the derivative of the logarithm of the density $p(x)$ is bounded. It is shown that in this case $D(p \| p_m) = O(1/m)^2$. It follows from (1.2) that if $m$ is chosen to be proportional to $n^{1/3}$ then the expected relative entropy converges at rate $n^{-2/3}$. Other traditional density estimators suffer similar deficiencies which sometimes can be repaired by similar modifications. For instance, let $K(t) = 1_{\{|t| \le 1/2\}}$ be the uniform kernel on $[-1/2,1/2]$. The kernel density estimator $(1/nh)\sum K((x-X_i)/h)$ may be modified by setting

$$\hat{p}_n(x) = \frac{(\sum_{i=1}^{n} K((x-X_i)/h)) + 1}{nh + 1} \quad \text{for } 0 \leq x \leq 1.$$

In this case it can be shown (by an argument similar to the one which is used to prove Theorem 1.1) that the expected Kullback-Leibler number is bounded by

$$ED(p \| \hat{p}_n) \leq D(p \| p_h) + \frac{1}{nh}$$

where $p_h$ is the convolution of $p$ with the kernel of width $h$. We will not discuss details in here.

### 1.3 Exponential Family Method

Hall (1987) gives a detailed examination of the Kullback-Leibler risk of estimators based on positive kernels and shows the necessity of using a sufficiently heavy-tailed kernel. However, no positive kernel estimator can have a faster rate of convergence than $n^{-4/5}$ in the Kullback-Leibler sense, even if the true density has a high degree of smoothness. Barron and Sheu (1988) avoid these difficulties by using estimators which are natural for the information-theoretic loss function. Rates of convergence of the Kullback-Leibler number of order $n^{-2r/(2r+1)}$ are obtained when the log-density function is assumed to have $r$ derivatives which are square integrable. The material from Barron and Sheu (1988) is included as part of this thesis.

For a given set of functions $\phi_1(x),...,\phi_m(x)$ and a density function $p_0(x)$, the probability density $\hat{p}(x)$ which minimizes the relative entropy $D(\hat{p} \| p_0)$ subject to the constraint that the expected values of $\phi_k(x)$ with respect to $\hat{p}$ match the sample expected values $(1/n)\sum_{i=1}^{n}\phi_k(X_i)$ for $k=1,...,m$, is known to be the density $\hat{p}(x) = p_{\hat{\theta}}(x)$ which maximizes the likelihood in the exponential

family

$$p_\theta(x) = p_0(x)\exp\{\theta_1\phi_1(x) + ... + \theta_m\phi_m(x) - \psi(\theta)\} \qquad (1.7)$$

where $\psi(\theta) = \log \int p_0 \exp\{\theta_1\phi_1 + ... + \theta_m\phi_m\}$. (Throughout this thesis logarithms are taken with base $e$.) Here $p_0$ may be thought of as an initial guess of the density function. Having observed the sample expectations, the estimator $\hat{p}$ updates the initial choice in an optimal way. Indeed, if $p_0$ were the (unconditional) probability density function for $X_1$, then as $n \to \infty$, $\hat{p}$ would be the asymptotic conditional probability density function for $X_1$ given that $(1/n)\sum_{i=1}^{n}\phi_k(X_i) = \alpha_k$ (see Zabell 1980, Van Campenhout and Cover 1981, Csiszár 1984). Of course, regular exponential family models for probability densities are extensively utilized in statistical practice. We refer to Brown (1986) for a thorough treatment of the fundamental properties of these models.

Here sequences of such exponential families are considered with $m = 1,2,...$ and an approach is developed to examine the asymptotics of $D(p \| p_{\hat{\theta}})$ as $m \to \infty$ and $n/m \to \infty$. It is not assumed that the true density $p$ is in the parametric family for any finite $m$. However, it is assumed that $\log p/p_0$ satisfies conditions which ensure that there exists a sequence of approximations $\sum_{k=0}^{m}\beta_k\phi_k(x)$ which converges to $\log p/p_0$ in $L_2$ as $m \to \infty$. We are particularly interested in the cases that the $\phi_k$ are basis functions for polynomials, splines, or trigonometric series. The relative entropy $D(p \| \hat{p})$ is shown to decompose 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. Exponential family method is also useful to multivariate density estimation.

An important reference for asymptotics of exponential families when the number of parameters tends to infinity is Portnoy (1988). He obtained asymptotics for $\|\hat{\theta}-\theta\|$ and $D(p_{\hat{\theta}} \| p_{\theta})$ under the assumption that the distribution for the random variables $X_i$ has a density function $p_{\theta}$ in the parametric family, i.e. the bias or approximation error term referred to above is zero. We prefer to not making 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 $m$.

As the analysis reveals, the Kullback-Leibler number is mathematically convenient when examining asymptotics of exponential families. Indeed, $D(p_{\hat{\theta}} \| p_{\theta})$ arises naturally as $1/n$ times the decrement in log-likelihood at $\theta$ from the maximum at $\hat{\theta}$. Nevertheless, the most compelling motivation for examining the Kullback-Leibler number is not its mathematical appropriateness but rather its suitability for application. In Barron and Cover (1988) it is shown that $D(P \| \hat{P})$ bounds the decrement in exponential growth of wealth when investment portfolios are based on an estimate $\hat{P}$ of the stock market distribution rather than based on the (unknown) actual distribution $P$. For a data compression problem, $D(P \| \hat{P})$ determines the redundancy (excess average length) of a code based on $\hat{P}$ compared to the optimal code based on the unknown $P$.

# CHAPTER 2 CONVERGENCE OF $D(P \| \hat{P}_{n,m})$ IN PROBABILITY

## 2.1 Statements

In this chapter the exponential family method of the density estimator in the univariate case will be discussed in more detail. Our concentration on the exponential family method is not intended to imply that the method is the best to use to overcome the problem in which case $D(p \| \hat{p}) = \infty$, but there are several reasons for introducing this method first of all. In this section a general theorem is given and then its implications are developed for polynomial, spline, and trigonometric cases.

Let $(\mathbf{X},\mathbf{B})$ be a measurable space and let $\nu$ be a fixed probability measure on this space. For $m \geq 1$, let $S_m$ be a linear space of dimension $m$ spanned by bounded measurable functions $\phi_{0,m}(x) = 1, \phi_{1,m}(x),\ldots,\phi_{m,m}(x)$. It is assumed that there exist positive numbers $a_m$ such that for all $f_m \in S_m$

$$\| f_m \|_\infty \leq a_m \| f_m \|_2. \tag{2.1}$$

Here $\| f \|_\infty$ is the essential supremum of $| f |$ and $\| f \|_2 = (\int f^2 d\nu)^{1/2}$. For the particular cases of interest, the sequences numbers $a_m$ are seen to be proportional to $\sqrt{m}$ or proportional to $m$ (see Section 2.5).

The $L_2$ approximation of the log-density function plays a key role in determining the asymptotics of the Kullback-Leibler number. In particular, the rate of convergence (of the bias component) is primarily determined by the degree of approximation $\delta_m(f) = \min\{\| f - f_m \|_2 : f_m \in S_m\}$ with $f = \log p$. This minimum is known to be achieved by a unique $f_m$: the orthogonal projection of $f$ onto $S_m$. A condition on the $L_\infty$ approximation error is also required

($\| f - f_m \|_\infty \leq \gamma_m$ for some bounded sequence $\gamma_m$), but the $L_\infty$ bounds have only a secondary influence on the asymptotics.

Let $\gamma_0$ and $\gamma$ be arbitrary positive constants. Consider functions $f$ for which there exists $f_m \in S_m$ and $\delta_m > 0$ such that the following approximation properties are satisfied

$$\| f \|_\infty \leq \gamma_0 \tag{2.2}$$

$$\| f - f_m \|_\infty \leq \gamma \tag{2.3}$$

and

$$\| f - f_m \|_2 \leq \delta_m . \tag{2.4}$$

It is assumed that the numbers $\delta_m$ and $a_m$ satisfy

$$\delta_m a_m \leq c_0 \tag{2.5}$$

where $c_0$ is a positive constant which depends only on $\gamma_0$ and $\gamma$: namely, $c_0 = 1/(4e^{\gamma_0 + 4\gamma + 1})$. Often it is the case that the sequence $\delta_m a_m$ tends to zero as $m \to \infty$.

Consider the family of probability density functions with respect to $\nu$ for which the logarithm of the density is in $S_m$. A parameterization of this family $\{ p_\theta(x) : \theta \in \mathbf{R}^m \}$ is

$$p_\theta(x) = \exp\{ \sum_{k=1}^{m} \theta_k \phi_{k,m}(x) - \psi_m(\theta) \} \tag{2.6}$$

where $\psi_m(\theta) = \log \int \exp\{\sum \theta_k \phi_{k,m}(x)\} \nu(dx)$, $\theta \in \mathbf{R}^m$. (In this definition of the family, the factor $p_0(x)$ from equation (1.2) has been incorporated into the dominating measure $\nu(dx) = p_0(x)\lambda(dx)$.)

Let $\hat{p}_{n,m} = p_{\hat{\theta}}$ where $\hat{\theta}$ is the maximum likelihood estimator which is characterized as the solution $\hat{\theta}$, when it exists, to the system of equations

$$\int \phi_k(x) p_{\hat{\theta}}(x) \nu(dx) = \frac{1}{n} \sum_{i=1}^{n} \phi_k(X_i) \tag{2.7}$$

for $k = 1,2,\ldots,m$, $\hat{\theta} \in \mathbf{R}^m$. The random sample $X_1,\ldots,X_n$ is drawn from a probability distribution $P$ which has a density function $p$ with respect to $\nu$.

**Proposition:** *If conditions (2.1) through (2.5) are satisfied with* $f = \log p$, *then there exist positive constants* $c_1,c_2,c_3$ *depending only on* $\gamma_0$ *and* $\gamma$ *such that for every* $1 \le K \le c_3 n/(m a_m^2)$, *there is a set of probability less than* $1/K$, *such that outside this set, the solution* $\hat{\theta}$ *exists and*

$$D(p \,\|\, \hat{p}_{n,m}) \le c_1 \delta_m^2 + c_2 \frac{m}{n} K. \tag{2.8}$$

*The bound on the probability,* $P\{ D(p \,\|\, \hat{p}_{n,m}) > c_1 \delta_m^2 + c_2 (m/n) K \} \le 1/K$, *holds uniformly for all density functions which satisfy the indicated conditions.*

**Remark 1:** Suppose $m = m_n$ is chosen such that $m_n \to \infty$ and $a_{m_n}^2 m_n / n \to 0$. If $\log p$ satisfies the indicated conditions for all large $m$ and if $\lim \delta_m = 0$, then a consequence of the proposition is that the Kullback-Leibler distance $D(p \,\|\, \hat{p}_{n,m_n})$ converges to zero in probability at rate $\delta_{m_n}^2 + m_n/n$.

**Remark 2:** Explicit constants are $c_1 = (1/2)e^{\gamma_0+\gamma}$, $c_2 = 2e^{3\gamma_0 + 6\gamma + 6}$ and $c_3 = (1/16)e^{-\gamma_0-4\gamma-4}$. Improvement of these constants is possible as shown in Section 2.4.

**Remark 3:** One consequence of the Proposition is that the maximum likelihood estimator exists except perhaps in a set of probability less than $(m a_m^2)/(c_3 n)$.

For an example in which a finite-valued maximum likelihood estimate can fail to exist in a set of small probability, consider the regular exponential family of densities on [0,1] with $\phi_k(x)$ equal to the indicator of the interval $(k/(m+1),(k+1)/(m+1)]$ for $k = 1,2,\ldots,m$. The ordinary histogram estimator $\hat{p}_n$ is the maximum likelihood density in this case and it is seen below (as a special case of a corollary to the Proposition) that under reasonable assumptions $D(p \| \hat{p}_n)$ converges to zero in probability. Nevertheless, whenever the unlikely event occurs that one or more of the intervals have no observations, then $\hat{\theta}$ in $\mathbf{R}^m$ does not exist and $D(p \| \hat{p}_n) = \infty$.

**Remark 4**: For probability density functions having support on the whole real line, application of the Proposition would require a choice of a density function $p_0(x)$ for which the ratio with the true density has a bounded logarithm. This is a rather severe requirement on prior knowledge of the tail behavior of the unknown density. Attempts to map the problem into the unit interval (e.g. by transforming using a cumulative distribution function) are going to suffer from a similar difficulty: the transformed density will have an unbounded logarithm near 0 and 1, unless knowledge of the tail behavior is incorporated in the choice of transformation. For these reasons, the result is perhaps best suited for problems where a bounded support set is known and the density function is bounded away from zero and infinity on this set. In the examples below we focus on the case that $\mathbf{X} = [0,1]$ and the dominating measure is the uniform (Lebesgue measure).

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 $\|f^{(r)}\|_2 < \infty$.

We now consider three special cases. The verification of the details for

these cases is in Section 2.5. In each of these cases we assume that the reference density $p_0(x)$ is a density function with respect to Lebesgue measure, and that $\log p_0(x)$ satisfies the same smoothness requirements that are assumed for the true density. In particular, we require that $\log p_0(x) \in W_2^r$.

**Polynomials:** Let $S_m$ be the space of polynomials of degree $\leq m$ on $[0,1]$. The family of density functions may be parameterized as

$$p_\theta(x) = p_0(x) \exp\{\theta_1 x + \theta_2 x^2 + \ldots + \theta_m x^m - \psi_m(\theta)\}.$$

It is shown that (2.1) holds with $a_m$ proportional to $m$. Also it is shown that if $f$ is in $W_2^r$ with $r \geq 1$, then $L_2$ and $L_\infty$ approximation bounds hold for $m > r$ with $\delta_m = (1/(m-r))^r \| f^{(r)} \|_2$ and $\gamma_m$ proportional to $(1/m)^{r-1} \| f^{(r)} \|_2$. Then $\delta_m a_m$ tends to zero for $r \geq 2$, so (2.5) is satisfied for all large $m$. Thus we have the following result.

**Corollary** (*Polynomial Case*): *If* $\log p \in W_2^r$ *for some* $r \geq 2$ *and* $m \to \infty$, $m^3/n \to 0$, *then the Kullback-Leibler distance* $D(p \| \hat{p}_n)$ *converges to zero in probability at rate*

$$(\frac{1}{m})^{2r} + \frac{m}{n}. \tag{2.9}$$

In particular if $m$ is chosen to be proportional to $n^{1/(2r+1)}$ then the convergence rate is $n^{-2r/(2r+1)}$.

**Splines:** Fix $s \geq 1$ and let $S_m$ be the space of splines of order $s$ on $[0,1]$ with interior knots at the points $k\Delta$ for $k = 1,2,\ldots,m+1-s$ and $\Delta = 1/(m+2-s)$. This is the linear space of functions $f$ which are piecewise polynomials of degree less than $s$ for which $f^{(j)}$ is continuous at the knots for $0 \leq j < s-1$. One basis for this space consists of the functions $\{1, x, \ldots, x^{s-1},$

$((x-k\Delta)_+)^{s-1}, k=1,2,\ldots,m+1-s\}$ where $(\cdot)_+$ denotes the positive part.

It is shown that (2.1) holds with $a_m$ proportional to $\sqrt{m}$. This smaller value of $a_m$ permits approximation results to be applied for all $r \geq 1$. (Again, see Section 2.5 for the details.)

**Corollary** (*Spline Case*): *If* $\log p \in W_2^r$, $r \geq 1$, *and* $m \to \infty$, $m^2/n \to 0$, *then the Kullback-Leibler distance* $D(p \| \hat{p}_n)$ *converges to zero in probability at rate*

$$(\frac{1}{m})^{2\min\{r,s\}} + \frac{m}{n}.$$

In particular if $s=r$ and $m$ is chosen to be proportional to $n^{1/(2r+1)}$ then the convergence rate is $n^{-2r/(2r+1)}$.

When $s=1$, $S_m$ consists of piecewise constant functions and $\hat{p}_n$ is simply the histogram estimator. In this case we have that $D(p \| \hat{p}_n)$ converges to zero in probability at rate $n^{-2/3}$ when $\log p$ has a bounded derivative and $m$ is proportional to $n^{1/3}$.

**Trigonometric Series:** Let $\phi_0(x) = 1$, $\phi_{2k-1} = \sqrt{2}\sin(2\pi kx)$, and $\phi_{2k} = \sqrt{2}\cos(2\pi kx)$ for $k \leq m/2$ and $0 \leq x \leq 1$. It is seen that (2.1) holds with $a_m = \sqrt{m+1}$. Using standard approximation theorems for periodic functions yields the following.

**Corollary** (*Trigonometric Case*): *Suppose* $f = \log p$ *is in* $W_2^r$, $r \geq 1$, *and satisfies the boundary conditions* $f^{(j)}(0) = f^{(j)}(1)$ *for* $0 \leq j < r$. *If* $m \to \infty$, $m^2/n \to 0$, *then* $D(p \| \hat{p}_n)$ *converges to zero in probability at rate*

$$(\frac{1}{m})^{2r} + \frac{m}{n}.$$

In particular if $m$ is chosen to be proportional to $n^{1/(2r+1)}$ then the convergence rate is $n^{-2r/(2r+1)}$.

**Remark 5**: In each of the three cases considered above, the convergence in probability is uniform for any set $B$ of log densities having bounded Sobolev norm. In particular it is seen that

$$\lim_{K\to\infty} \lim_{n\to\infty} \sup_{\log p \in B} P\{ D(p \| \hat{p}_n) \geq ((1/m_n)^{2r^*} + m_n/n)K\} = 0 \qquad (2.10)$$

for any sequence $m_n$ with $m_n \to \infty$ and $m_n^2/n$ as $n \to \infty$. (Here $r^* = r$ in the polynomial and trig cases and $r^* = \min\{r,s\}$ in the spline case. For the trig case $B$ is restricted to log densities which satisfy the indicated boundary conditions.) To verify this, note that for any Sobolev ball $B \subset W_2^r$, there is a constant $c$ such that for all $f \in B$, $\| f^{(r)} \|_2$ and $\| f \|_\infty$ are less than $c$.

**Remark 6**: The same rates $n^{-2r/(2r+1)}$ are known to be the optimal minimax rates for convergence of the integrated squared error for sets of density functions having bounded Sobolev norm (see e.g. Efroimovich and Pinsker 1983). For densities which have a bounded logarithm the Kullback-Leibler number is closely related to the integrated squared error (see Lemma 3). Moreover, when the density is bounded away from zero, Sobolev assumptions on the density are not too different from Sobolev assumptions on the log-density. This suggests that $n^{-2r/(2r+1)}$ should be a lower bound on the minimax rate for Kullback-Leibler error as well as for integrated squared error. If so then each of the three estimators discussed above possess *optimal* rate properties. The determination of lower bounds on minimax risk for the Kullback-Leibler number is left as a topic for later study.

**Remark 7**: A refinement of the approximation bounds given in section 2.4

shows that for any fixed $f$ in $W_2^r$, a rate of convergence of $\delta_m(f)$ which is of smaller order than $(1/m)^r$ is obtained in the polynomial and trigonometric cases (although the improvement is not uniform for balls in $W_2^r$), whence for any fixed log-density which satisfies the indicated assumptions, $D(p \| \hat{p}_n)$ converges at a rate slightly faster than indicated in the corollaries.

## 2.2 Information Projection

Let $1, \phi_1(x), \phi_2(x), \ldots, \phi_m(x)$ be linearly independent measurable functions (so that $\sum \theta_k \phi_k - \sum \theta'_k \phi_k$ = constant a.e. implies $\theta' = \theta$). Let $\{P_\theta : \theta \in \Theta\}$ be the $m$-dimensional exponential family of probability measures with density functions $p_\theta = dP_\theta/d\nu$ of the form

$$p_\theta(x) = e^{\theta \cdot \phi(x) - \psi(\theta)} \tag{2.11}$$

where $\phi(x) = (\phi_1(x), \ldots, \phi_m(x))$, $\theta \cdot \phi = \sum_{k=1}^m \theta_k \phi_k$, and $\psi(\theta) = \log \int e^{\theta \cdot \phi(x)} \nu(dx)$. We assume that the natural parameter space $\Theta = \{\theta \in \mathbf{R}^m : \psi(\theta) < \infty\}$ is open, i.e. the family is regular (Brown 1986, p.2). If $\nu$ is finite and $\phi$ is bounded, then $\Theta = \mathbf{R}^m$ and the family is clearly regular.

Let $C = \{P : \int \phi(x) P(dx) = \alpha\}$ be the set (hyperplane) of all probability measures for which the expected value of $\phi(X)$ is equal to $\alpha$, where $\alpha \in \mathbf{R}^m$. It turns out that in an information-theoretic sense the set $C$ and the family $\{P_\theta : \theta \in \Theta\}$ are orthogonal. Indeed, all members of the family have the same information projection onto $C$ (in the sense of Csiszár 1975).

The following Lemma recalls for convenience some of the projection properties. We let $\Omega = \{\int \phi \, dP_\theta : \theta \in \Theta\}$ and consider the equation

$$\int \phi \, dP_\theta(dx) = \alpha. \tag{2.12}$$

**Lemma 2.1** (*Information Projection*):

(a) *Uniqueness*: The solution $\theta(\alpha)$ to the equation (2.12) is unique for $\alpha \in \Omega$.

(b) *Likelihood maximization*: If $\alpha \in \Omega$, the function $F(\theta) = \theta \cdot \alpha - \psi(\theta)$ has a unique maximum at $\theta(\alpha)$; if $\alpha$ is not in $\Omega$, then $F(\theta)$ has no local or global maximum in $\Theta$.

For the next two properties, suppose $\alpha \in \Omega$.

(c) *Relative entropy minimization*:

(i) $P^* = P_{\theta(\alpha)}$ uniquely minimizes $D(P \| \nu)$ subject to $P \in C$.

(ii) For any fixed $P \in C$, if $D(P \| P_\theta)$ is finite for some $\theta \in \Theta$, then $\theta^* = \theta(\alpha)$ uniquely minimizes $D(P \| P_\theta)$.

(d) *Pythagorean identity*: For all $P \in C$ and all $\theta \in \Theta$

$$D(P \| P_\theta) = D(P \| P^*) + D(P^* \| P_\theta). \tag{2.13}$$

Thus $P^* = P_{\theta^*}$ is the information projection of every $P_\theta$ onto $C$.

Recall that the relative entropy of a probability measure $P$ with respect to a measure $Q$, in the case that $Q$ is equivalent to $\nu$, is given by

$$D(P \| Q) = \int p(x) \log \frac{p(x)}{q(x)} \nu(dx)$$

when $P << \nu$ and $D(P \| Q) = \infty$ otherwise. Here $p = dP/d\nu$ and $q = dQ/d\nu$. Also recall that if $Q$ and $P$ are probability measures, then $D(P \| Q) \geq 0$ with equality if and only if $P = Q$.

**Remark**: The familiar results (a) and (b) are special cases of Brown (1966, Theorems 3.6 and 5.5). The results (c) and (d) are special cases of results in Csiszár (1975, Section 3). We give a short proof to emphasize the commonality.

**Proof of Lemma 2.1**: Observe that by the positivity of the density (3.1), all the measures $P_\theta$, $\theta \in \Theta$, are equivalent to $\nu$. First we show the Pythagorean identity (3.3). This identity is trivial if $P$ is not absolutely continuous with respect to $\nu$, so we now suppose $P << \nu$. Obviously,

$$\log \frac{p(x)}{p_\theta(x)} = \log \frac{p(x)}{p_{\theta^*}(x)} + \log \frac{p_{\theta^*}(x)}{p_\theta(x)}. \tag{2.14}$$

where $\theta^*$ is any solution to (2.12). Taking the expected value with respect to $P$ establishs (2.13) with $P^* = P_{\theta^*}$ since the second term on the right side of (2.14) has the same expectation with respect to $P$ or $P^*$. A similar identity holds with $\nu$ in place of $P_\theta$, i.e., $D(P \| \nu) = D(P \| P^*) + D(P^* \| \nu)$, from which (c)(i) follows by the positivity of $D(P \| P^*)$.

Now (c)(ii) and (a) follow from the Pythagorean identity (2.13) by the positivity of $D(P^* \| P_\theta)$, with the uniqueness due to the fact that if $D(P_{\theta^*} \| P_\theta) = 0$ then $\log p_{\theta^*} = \log p_\theta$ a.e. and hence $\theta^* = \theta$ by the assumed linear independence of the functions $\phi_k$. For $\alpha \in \Omega$, the maximization of $F(\theta)$ is the same as the minimization of $D(P^* \| P_\theta)$ since $D(P^* \| P_\theta) = F(\theta^*) - F(\theta)$, so the first part of (b) is a special case of (c)(ii).

Finally, the second conclusion of (b) follows from the basic fact (Brown 1986, Thm.2.2) that for $\theta$ in $\Theta$ (which is open), $F(\theta)$ is continuously differentiable with gradient equal to $\alpha - \int \phi \, dP_\theta(dx)$ which cannot be zero if $\alpha$ is not in $\Omega$. □

## 2.3 Bounds

In order to prove the main theorem, we need some upper and lower bounds of Kullback- Leibler number. Also we need some bounds in terms of

distance between the parameters. We divide two parts to discuss it.

### A. $L_2$ 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 2.2:**

$$D(p \| q) \ge \frac{1}{2} e^{-\| \log p/q \|_\infty} \int p (\log \frac{p}{q})^2 \tag{2.15}$$

*and*

$$D(p \| q) \le \frac{1}{2} e^{\| \log p/q - c \|_\infty} \int p (\log \frac{p}{q} - c)^2 \tag{2.16}$$

*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 2.2:** From the Taylor expansion of $e^z$ we have

$$\frac{z^2}{2} e^{-z_-} \le e^z - 1 - z \le \frac{z^2}{2} e^{z_+} \tag{2.17}$$

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

$$\int p \log \frac{p}{q} = \int (p \log \frac{p}{q} + q - p)$$

$$= \int p(e^{-f} - 1 + f)$$

$$\geq \int p \frac{f^2 e^{-(-f)_-}}{2}$$

$$\geq \frac{1}{2} e^{-\|f_+\|_\infty} \int p f^2 \qquad (2.18)$$

which yields inequality (2.15).

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

$$\int p \log \frac{p}{q} \leq \int p(e^{-f} - 1 + f) + 1 + c - e^c$$

$$\leq \int p \frac{f^2 e^{(-f)_+}}{2}$$

$$\leq \frac{1}{2} e^{\|f_-\|_\infty} \int p f^2 \qquad (2.19)$$

which yields the desired inequality. □

We also need the following lemma.

**Lemma 2.3:**

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

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

**Proof:** Use the fact that $|e^z - 1| \le |z| e^{z_+}$ for $-\infty < z < \infty$ to get

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

**B. Bounds for Exponential Families.**

Let $\{p_\theta(x) = e^{\theta \cdot \phi(x) - \psi(\theta)}\}$ be a regular exponential family as in Section 2.2 with bounded functions $\phi_k$, $k=1,\ldots,m$, and a finite dominating measure $\nu(dx)$. 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 measure $Q$. Here $Q$ may be any probability measure having a density function $q = dQ/d\nu$ for which $\log q$ is bounded. If there is interest in certain meaningful parameters (and not just interest in the family of density functions, for which various parameterizations may be chosen) then the assumption of orthonormality is restrictive and the results in this section may be modified to use an eigenvalue assumption (e.g. of the form $\underline{\lambda} \|\theta\|_2 \le \|\theta \cdot \phi\|_{L_2(Q)} \le \overline{\lambda} \|\theta\|_2$ for all $\theta \in \mathbf{R}^m$ and some $0 < \underline{\lambda} \le \overline{\lambda}$).

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

$$\|f_m\|_\infty \le A_m \|f_m\|_{L_2(Q)}. \tag{2.20}$$

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

**Lemma 2.4**: *For* $\theta_0, \theta \in \mathbf{R}^m$

$$\| \log p_{\theta_0}/p_\theta \|_\infty \leq 2A_m \| \theta_0 - \theta \|, \tag{2.21}$$

$$D(P_{\theta_0} \| P_\theta) \leq \frac{b}{2} e^{A_m \| \theta_0 - \theta \|} \| \theta_0 - \theta \|^2, \tag{2.22}$$

*and*

$$D(P_{\theta_0} \| P_\theta) \geq \frac{1}{2b} e^{-2A_m \| \theta_0 - \theta \|} \| \theta_0 - \theta \|^2 \tag{2.23}$$

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

**Proof**: Observe that

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

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

from which it follows that $| \psi(\theta) - \psi(\theta_0) | \leq \| (\theta - \theta_0) \cdot \phi \|_\infty$. Now $\log p_{\theta_0}/p_\theta = (\theta_0 - \theta) \cdot \phi + \psi(\theta) - \psi(\theta_0)$ so it follows that $\| \log p_{\theta_0}/p_\theta \|_\infty \leq 2 \| (\theta - \theta_0) \cdot \phi \|_\infty \leq 2A_m \| \theta_0 - \theta \|$ which gives (2.21). Using the assumed orthonormality of the $\phi_k$, the inequalities (2.22) and (2.23) follow from Lemma 2.2 with $c = \psi(\theta) - \psi(\theta_0)$. This completes the proof. □

Now for the key Lemma in this chapter. Recall that $\theta(\alpha)$ denotes the unique solution to $E_{p_\theta} \phi(X) = \alpha$ (whenever such a solution exists); however, in general there is no explicit formula for $\theta(\alpha)$. The next result establishs sufficient conditions for the existence of a solution at $\alpha$ in terms of the distance from a point $\alpha_0$ in $\mathbf{R}^m$ for which a solution is known to exist. Moreover, the distance between $\theta(\alpha)$ and $\theta(\alpha_0)$ is bounded in terms of the distance between $\alpha$ and $\alpha_0$. Under a different set of assumptions, similar results were obtained by

Portnoy (1988).

**Lemma 2.5:** *Let* $\theta_0 \in \mathbf{R}^m$, $\alpha_0 = \int \phi \, dP_{\theta_0}$, $\alpha \in \mathbf{R}^m$ *be given. Let* $b = e^{\|\log q/p_{\theta_0}\|_\infty}$ *and assume that (2.20) holds. If*

$$\| \alpha - \alpha_0 \|_2 \leq \frac{1}{4ebA_m} \tag{2.24}$$

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

$$\| \theta(\alpha) - \theta(\alpha_0) \|_2 \leq 2be^\tau \| \alpha - \alpha_0 \|_2, \tag{2.25}$$

$$\| \log p_{\theta(\alpha_0)}/p_{\theta(\alpha)} \|_\infty \leq 4be^\tau A_m \| \alpha - \alpha_0 \|_2 \leq \tau, \tag{2.26}$$

*and*

$$D(P_{\theta(\alpha_0)} \| P_{\theta(\alpha)}) \leq 2be^\tau \| \alpha - \alpha_0 \|^2. \tag{2.27}$$

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

For identifying asymptotic rates, adequate bounds may be obtained with $\tau = 1$. The smallest choice $\tau = 4ebA_m \| \alpha - \alpha_0 \|$ yields improved constants.

**Proof of Lemma 2.5:** 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 2.2. Then since $D(P_{\theta_0} \| P_\theta) = (\theta_0 - \theta) \cdot \alpha_0 + \psi(\theta) - \psi(\theta_0)$ we have that for all $\theta \in \mathbf{R}^m$

$$\begin{aligned} F(\theta_0) - F(\theta) &= (\theta_0 - \theta) \cdot \alpha + \psi(\theta) - \psi(\theta_0) \\ &= D(P_{\theta_0} \| P_\theta) - (\theta_0 - \theta) \cdot (\alpha_0 - \alpha). \end{aligned} \tag{2.28}$$

It follows by Lemma 4 and the Cauchy-Schwartz inequality that for all $\theta \in \mathbf{R}^m$,

$$F(\theta_0) - F(\theta) \geq \frac{1}{2b} e^{-2A_m \| \theta_0 - \theta \|} \| \theta_0 - \theta \|^2 - \| \theta_0 - \theta \| \, \| \alpha_0 - \alpha \|. \tag{2.29}$$

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

$$F(\theta_0) - F(\theta) > (e^{\tau - 4A_m e^{\tau} b \| \alpha - \alpha_0 \|} - 1) 2e^{\tau} b \| \alpha - \alpha_0 \|^2. \qquad (2.30)$$

The right side is non-negative when $4ebA_m \| \alpha - \alpha_0 \| \leq \tau \leq 1$. 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, i.e. $\| \theta^* - \theta_0 \| < r$. The gradient of $F$ at $\theta^*$ must be zero which means that $\alpha - \int \phi \, dP_{\theta^*} = 0$, that is $\theta^* = \theta(\alpha)$. Therefore $\| \theta(\alpha) - \theta(\alpha_0) \|_2 < r$ which verifies (2.25). The inequalities (2.26) then follow by applying Lemma 2.4. To verify (2.27), since $F(\theta(\alpha)) \geq F(\theta_0)$, it follows from (2.28) and (2.25) that

$$\begin{aligned} D(P_{\theta_0} \| P_{\theta}) &\leq (\theta(\alpha_0) - \theta(\alpha)) \cdot (\alpha_0 - \alpha) \\ &\leq \| \theta_0 - \theta \| \, \| \alpha_0 - \alpha \| \\ &\leq 2be^{\tau} \| \alpha - \alpha_0 \|^2. \end{aligned}$$

This completes the proof of Lemma 2.5. □

## 2.4 Main Theorem

Here we establish a result which is seen to be equivalent to the proposition in Section 2.1. There we stated the conditions in terms of the $L_2(\nu)$ norm, since with $\nu$ known, the conditions are easier to check. Here we state the conditions in terms of the $L_2(P)$ norm, since this yields simpler bounds and potentially smaller constants. The equivalence of the conditions follows from the assumed boundedness of $\log p$.

Let $A_m = A_m(P)$ be such that for all $f_m \in S_m$

$$\| f_m \|_\infty \leq A_m \| f_m \|_{L_2(P)} \tag{2.31}$$

(Observe that if $a_m = A_m(v)$ satisfies (2.1), then the best constant for (2.31) satisfies $A_m(P) \leq e^{(1/2)\| \log p \|_\infty} a_m$.)

Let $f = \log p$ and assume that there exists $f_m \in S_m$ with $\| f - f_m \|_\infty \leq \gamma$ and

$$\| f - f_m \|_{L_2(P)} \leq \Delta_m. \tag{2.32}$$

(Observe that if $\delta_m$ satisfies (2.4), then the best $\Delta_m$ is not larger than $e^{(1/2)\| \log p \|_\infty} \delta_m$.) Suppose $\Delta_m A_m$ is bounded by the constant $(1/4)e^{-4\gamma-1}$ and let $0 < \tau \leq 1$ be such that, with $C_0 = (1/4)\tau e^{-4\gamma-1}$

$$\Delta_m A_m \leq C_0. \tag{2.33}$$

Let $\hat{p}_{n,m}$ be the density estimate, when it exists, which maximizes the likelihood among all probability density functions with logarithm in $S_m$.

**Theorem 2.1:** *For all* $1 \leq \kappa \leq C_3 n/(mA_m^2)$, $\hat{p}_{n,m}$ *exists and*

$$D(p \| \hat{p}_{n,m}) \leq C_1 \Delta_m^2 + C_2 \frac{m}{n} \kappa \tag{2.34}$$

*except in a set of probability less than* $1/\kappa$. *Here* $C_1 = (1/2)e^{\gamma}$ *and* $C_2 = 2e^{2\gamma+2\tau}$ *and* $C_3 = (1/16)\tau^2 e^{-4\gamma-2\tau}$.

**Remark**: Asymptotically, the constants may approach $C_1 = 1/2$ and $C_2 = 2$. This is the case if $\lim \| f - f_m \|_\infty = 0$, $\lim \Delta_m A_m = 0$, $\lim m_n = \infty$, and $\lim m_n A_{m_n}^2 / n = 0$ as $n \to \infty$, so that $\gamma$ and $\tau$ may be chosen arbitrarily small for $n$ sufficiently large.

**Proof of The Theorem 2.1:** 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 $\int \phi dP_{\theta^*} = \int \phi dP$ and that $\log p/p_{\theta^*}$ is bounded by a constant. This $P_{\theta^*}$ is the information projection referred to in Section 2.2. 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 $\alpha_0 = \int \phi dP_\beta$ where $\beta = (\beta_1,\ldots,\beta_m)$ and set $\alpha = \int \phi dP$. Then the entries in the vector $\alpha - \alpha_0$ are given by $\int ((p-p_\beta)/p)\phi_k dP$ for $k=0,1,\ldots,m$. These entries are seen to be the coefficients in the $L_2(P)$ orthonormal projection of $(p-p_\beta)/p$ onto $S_m$, so by Bessel's inequality and Lemma 3,

$$\begin{aligned} \| \alpha - \alpha_0 \| &\le \| (p-p_\beta)/p \|_{L_2(P)} \\ &\le e^{\| f - f_m \|_\infty - (\beta_0 + \psi(\beta))} \| f - f_m \|_{L_2(P)} \\ &\le e^{2\gamma} \Delta_m \end{aligned} \tag{2.35}$$

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 $2\| f - f_m \|_\infty$ which in turn is bounded by $2\gamma$. Now apply Lemma 2.5 with $\theta_0 = \beta$, $q = p$, $\alpha = \int \phi(x) P(dx)$, and $b = e^{\| \log p/p_\beta \|_\infty} \le e^{2\gamma}$. The condition (2.24) is satisfied if $e^{2\gamma}\Delta_m \le \tau/(4ebA_m)$ which is true for some $0 < \tau \le 1$ if (2.33) holds. In which case we may conclude that $\theta^* = \theta(\alpha)$ exits and that $\| \log p_{\theta^*}/p_\beta \|_\infty \le \tau$. So by the triangle inequality

$$\| \log p/p_{\theta^*} \|_\infty \leq 2\gamma + \tau. \tag{2.36}$$

Now for the second task, we show that $D(p_{\theta^*} \| p_{\hat{\theta}})$ is small with high probability. Lemma 2.5 is applied once more with different choices of the parameters. In particular, take $\theta_0$ to be $\theta^*$: the corresponding $\alpha_0$ is $\int \phi \, dP^*$ (which is the same as $\int \phi \, dP$). For $\alpha$ take $\overline{\phi}_n = (1/n)\sum_{i=1}^n \phi(X_i)$. (Whenever a solution to $\int \phi \, dP_\theta = \overline{\phi}_n$ exists, we recognize this solution $\hat{\theta} = \theta(\overline{\phi}_n)$ as the maximum likelihood estimate.) With these choices $\| \alpha - \alpha_0 \|^2 = \sum_{k=1}^m (\overline{\phi}_{n,k} - E_P \phi_k)^2$. Lemma 2.5 requires that this distance between $\alpha$ and $\alpha_0$ be not too large. By Chebyshev's inequality $\| \alpha - \alpha_0 \|^2 \leq K m/n$ except in a set of probability which satisfies

$$P\{\sum_{k=1}^m (\overline{\phi}_{n,k} - E_P \phi_k)^2 > \frac{m}{n} K\} \leq \frac{n}{mK} E_P[\sum_{k=1}^m (\overline{\phi}_{n,k} - E_P \phi_k)^2]$$

$$= 1/K \tag{2.37}$$

where the last identity is due to the fact that $X_1,\ldots,X_n$ are independent with distribution $P$ and the functions $\phi_k(X)$ are normalized to have zero mean and unit variance with respect to $P$. Now apply Lemma 2.5 with $q = p$ and $b = e^{\| \log p/p_{\theta^*} \|_\infty} \leq e^{2\gamma+\tau}$. If $(K m/n)^{1/2} \leq \tau/(4ebA_m)$ then except in the set above (which has probability less than $1/K$), the conditions of the lemma are satisfied, whence the MLE $\hat{\theta}$ exists and

$$D(p_{\theta^*} \| p_{\hat{\theta}}) \leq 2be^\tau \frac{m}{n} K$$

$$\leq 2e^{2\gamma+2\tau} \frac{m}{n} K. \tag{2.38}$$

Finally, by Lemma 1, the Kullback-Leibler loss decomposes into a sum of approximation error and estimation error terms:

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

The estimation error $D(p^* \| \hat{p})$ has just been shown to be less than $C_2(m/n)\mathrm{K}$ except in a set of probability less than $1/\mathrm{K}$. By Lemma 1 and Lemma 2, the approximation error satisfies

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

This completes the proof of the Theorem. □

## 2.5 Details

In this section, it is shown how the conditions are satisfied for the corollaries given in Section 2.1. First we give a useful Lemma.

**Lemma 2.6**: *If $g(x)$ is a polynomial of degree $\leq d$ on $[a,b]$ then*

$$\sup_{x \in [a,b]} | g(x) | \leq (d+1) \left(\frac{1}{b-a}\right)^{1/2} \left(\int_a^b g^2\right)^{1/2}. \tag{2.39}$$

**Remark 1**: The lemma applies with $d = m$ and $[a,b] = [0,1]$ to show that the condition (2.1) holds with $a_m = m+1$ in the polynomial case.

**Remark 2**: The lemma also applies with $d = s-1$ to each of the polynomial pieces, to obtain that if $g$ is a spline of order $s$ on $[0,1]$ with knots at $\Delta, 2\Delta, \ldots, 1-\Delta$ then

$$\sup_{x \in [0,1]} | g(x) | \leq \max_{j=1,2,\ldots,1/\Delta} s \left(\frac{1}{\Delta}\right)^{1/2} \left(\int_{(j-1)\Delta}^{j\Delta} g^2(x)\,dx\right)^{1/2}$$

$$\leq s \left(\frac{1}{\Delta}\right)^{1/2} \left(\int_0^1 g^2(x)\,dx\right)^{1/2}. \tag{2.40}$$

Setting $\Delta = 1/(m-s+2)$ this shows that condition (2.1) is satisfied with

$a_m = s\sqrt{m-s+2}$ in the spline case.

**Proof of Lemma 2.6**: First note that by scaling the polynomials it suffices to prove the result for $[a,b] = [0,1]$. Let $\phi_k, k=0,1,...,d$ be the orthonormal Legendre polynomials which are bounded in absolute value by $\sqrt{2k+1}$ (see Jackson 1930, p.25). Hence $\sum_{k=0}^{d}(\phi_k(x))^2 \leq (d+1)^2$. (Here equality is obtained at $x=0$ and $x=1$, so this bound can not be improved.) If $g$ is a polynomial of degree $d$, then $g = \sum_{k=0}^{d}\beta_k\phi_k(x)$ for some coefficients $\beta_k$. So by the the Cauchy-Schwartz inequality

$$|g(x)| \leq (\sum_{k=0}^{m}\phi_k^2(x))^{1/2}(\sum_{k=0}^{m}\beta_k^2)^{1/2} \tag{2.40}$$

$$\leq (m+1)(\int_0^1 g^2)^{1/2}$$

uniformly for $x$ in $[0,1]$. This completes the proof of Lemma 2.6. □

Now we separately examine the $L_2$ and $L_\infty$ approximation properties of polynomials, splines and trigonometric series. For the $L_\infty$ results we are able to adapt longstanding results from the approximation theory literature, especially Jackson (1930, Chapter I). Until recently, the $L_2$ approximation theory for polynomials was less well developed.

**Polynomials:** We examine the $L_2$ approximation error for $f \in W_2^r$ using the recent results of Cox (1988). Fix $r \geq 1$. Let $\phi_k(x)$, $k=0,1,..$ denote the normalized Legendre polynomials which are orthonormal with respect to the uniform weight function on $[0,1]$. The system $\{\phi_k^{(r)} : k \geq r\}$ is orthogonal with respect to the weight function $(x(1-x))^r$ on $[0,1]$ and has normalizing constants $c_k^2 = \int_0^1(\phi_k^{(r)}(x))^2(x(1-x))^r dx = (k+1)!/(k-r)!$. If $f$ is in $W_2^r$ then

$\int (f^{(r)}(x))^2(x(1-x))^r dx$ is finite (since $(x(1-x))^r$ is bounded) and the coefficients $\beta_k$ in the expansion of $f(x)$ in terms of the system $\{\phi_k\}$ are the same for $k \geq r$ as the coefficients in the expansion of $f^{(r)}(x)$ in terms of the system $\{\phi_k^{(r)}\}$.Consequently, $\sum_{k \geq r} c_k^2 \beta_k^2 = \int_0^1 (f^{(r)}(x))^2(x(1-x))^r dx$ which is not greater than $(1/4)^r \int (f^{(r)}(x))^2 dx$. Let $f_m(x) = \sum_{k=0}^m \beta_k \phi_k(x)$ be the $m$th degree Legendre polynomial approximation of $f$. Then for $m \geq r$

$$\| f - f_m \|_2^2 = \sum_{k=m+1}^{\infty} \beta_k^2$$

$$\leq \frac{1}{c_{m+1}^2} \sum_{k=m+1}^{\infty} c_k^2 \beta_k^2$$

$$\leq \frac{1}{(m+r+1)\cdot\ldots\cdot(m-r+2)} (\frac{1}{4})^r \| f^{(r)} \|_2^2 \, \varepsilon_m^2 \qquad (2.42)$$

for some $0 \leq \varepsilon_m \leq 1$ with $\lim \varepsilon_m = 0$ since $\lim_m \sum_{k>m} c_k^2 \beta_k^2 = 0$. (The first inequality in (2.42) follows from the monotonicity of the sequence $c_k^2 = (k+r)\cdot\ldots\cdot(k-r+1)$ with increasing $k$.) Consequently, (2.4) and (2.5) hold for all large $m$ with $\delta_m = (1/(2(m-r+2))^r \| f^{(r)} \|_2$. (Note that a slight improvement is obtained by including the factor $\varepsilon_m$ which depends on $f$; however, this improvement is not uniform for $f$ in a Sobolev ball.) The above analysis is a refinement of Cox (1988) who showed that $\| f - f_m \|_2 = O(1/m)^r$ for $f \in W_2^r$ without explicitly identifying the constants. Our analysis shows that in fact $\| f - f_m \|_2 = o(1/m)^r$.

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

$m \geq r$ by

$$|f(x) - f_m(x)| \leq \left(\sum_{k=m+1}^{\infty} \frac{2k+1}{c_k^2}\right)^{1/2} \left(\sum_{k=m}^{\infty} c_k^2 \beta_k^2\right)^{1/2}$$

$$\leq \left(\sum_{k=m+1}^{\infty} \frac{2e^2}{(k+r)^{2r-1}}\right)^{1/2} \left(\frac{1}{2}\right)^r \|f^{(r)}\|_2$$

$$\leq \frac{e^r}{(r-1)^{1/2}(m+r)^{r-1}} \left(\frac{1}{2}\right)^r \|f^{(r)}\|_2. \qquad (2.43)$$

Here we have used the inequality $c_k^2 \geq (k+r)^{2r} e^{-2r}$ (which may be deduced by comparing the sum $\sum_{j=k-r+1}^{k+r} \log j$ to the integral $\int_{k-r}^{k+r} \log x \, dx$) and well as the inequality for the sum $\sum_{k>m} (k+r)^{-2r+1} \leq 2(r-1)(m+r)^{-2(r-1)}$ (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 $\omega(\delta) \leq \delta^{1/2} \|f^{(r)}\|_2$ and then appling bounds from Jackson (1930, p.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 $m \geq s$. For the $L_2$ and $L_\infty$ degrees of approximation we use the results of De Boor and Fix (1973). It is only assumed that $f$ is in $W_2^r$. There is a spline $f_m$ in $S_m$ which approximates $f$ in the $L_2$ sense that $\|f - f_m\|_2 \leq K \Delta^{r^*} \|f^{(r^*)}\|_2$ where $r^* = \min\{r,s\}$ and $K$ is an absolute constant (see De Boor and Fix, Thm. 5.2). Thus we may take $\delta_m$ to be proportional to $(1/m)^{r^*}$. Now $f^{(r^*-1)}$ is continuous with modulus of continuity not greater than $\Delta^{1/2} \|f^{(r^*)}\|_2$. (Indeed if $|x-y| \leq \Delta$, 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-Schwartz inequality.) So by De Boor and Fix (Thm. 2.1), $\| f - f_m \|_\infty \leq K' \Delta^{r^*-1/2} \| f^{(r^*)} \|_2$ where $K'$ is an absolute constant. Thus we may take $\gamma_m$ to be proportional to $(1/m)^{r^*-1/2}$. This completes the details for the spline case.

**Trigonometric Functions:** 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 $\phi_0 = 1$ $\phi_{2k} = \sqrt{2}\cos(2\pi kx)$ and $\phi_{2k+1} = \sqrt{2}\sin(2\pi kx)$ for $0 \leq x \leq 1$. (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 $\| f - f_m \|_2 \leq (\pi(m+2))^{-r} \| f^{(r)} \|_2$. Similarly, applying the Cauchy-Schwartz inequality, it is seen that the Fourier series is absolutely convergent, with error $|f(x) - f_m(x)|$ bounded by $(\sum_{k > m/2} (2\pi k)^{-2r})^{1/2}$ $(\sum_{k > m/2} (2\pi k)^{2r} (\beta_{2k}^2 + \beta_{2k+1}^2))^{1/2}$ which is not greater than $(2r-1)^{-1/2} m^{-(r-1/2)} \pi^{-r} \| f^{(r)} \|_2$. Thus $\| f - f_m \|_\infty \leq O(m^{-(r-1/2)})$ 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, p.22, Cor.IV), yields the slightly worse but also satisfactory rate $\| f - f_m \|_\infty \leq O(m^{-(r-1/2)}) \log m$.]

Finally we determine $a_m$ for the trigonometric case. If $f_m = \sum_{k=0}^{m} \beta_k \phi_k(x)$ then by the Cauchy-Schwartz inequality we have for any $x \in [0,1]$

$$| f_m(x) | \leq \left( \sum_{k=0}^{m} \phi_k^2(x) \right)^{1/2} \left( \sum_{k=0}^{m} \beta_k^2 \right)^{1/2} = (m+1)^{1/2} \| f_m \|_2 \qquad (2.44)$$

where we have used the fact that $\cos^2 + \sin^2 = 1$. It follows that (2.1) holds with $a_m = \sqrt{m+1}$.

**Remark:** We return to the polynomial case and deduce bounds in the case that $f \in W_2^m$. By (2.42) and (2.43) with $r = m$,

$$\| f - f_m \|_2 \leq \left(\frac{1}{(2m+1)!}\right)^{1/2} \left(\frac{1}{2}\right)^m \| f^{(m)} \|_2 \tag{2.45}$$

$$\| f - f_m \|_\infty \leq \frac{e}{2(m-1)^{1/2} (4m/e)^{m-1}} \| f^{(m)} \|_2. \tag{2.46}$$

In particular if $m = 3$ and the third derivative of $f$ has norm equal to 1 then the cubic approximation error has $L_2$ norm not exceeding $1/(8\sqrt{7!})$ (approximately 0.002), which is surprisingly small!

Suppose $f = \log p$ is an infinitely differentiable function on [0,1] and that the sequence of derivatives $f^{(m)}$ have $L_2$ norms which do not grow faster than a factorial: i.e. $\| f^{(m)} \| \leq cm!$ for some constant $c$. From Stirling's formula it is seen that $m!/\sqrt{(2m+1)!} \leq (1/2)^m$ and, for $m > 1$,

$$\| f - f_m \|_2 \leq c\left(\frac{1}{4}\right)^m, \tag{2.47}$$

$$\| f - f_m \|_\infty \leq 4\sqrt{\pi} cm \left(\frac{1}{4}\right)^m. \tag{2.48}$$

In this case, a consequence of the Theorem is that if $m_n$ is chosen to be proportional to $\log n$ (to optimize the sum of the approximation and estimation components) then $D(p \| \hat{p}_n)$ converges to zero in probability at rate

$$\frac{\log n}{n}.$$

Of course this convergence is faster than $n^{-\rho}$ for any $\rho < 1$.

The practical implication of the above remarks 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.

# CHAPTER 3 CONVERGENCE OF $ED(P \| \hat{P}_{n,m})$

## 3.1 Preliminaries

The minimum Kullback-Leibler distance principle for density estimation (Shore and Johnson (1980) ,Van Campenhout and Cover (1981), Blahut (1987)) states that given an initial guess $p_0(x)$ and sample constraints $(1/n)\sum_{i=1}^{n}\phi_k(X_i) = \alpha_k, k = 1,2,\cdots,m$, the density $\hat{p}_n(x)$ should be estimated which minimizes the relative entropy $D(\hat{p} \| p_0) = \int \hat{p}(x)\log \hat{p}(x)/p_0(x)dx$ subject to the constraint that $\int \phi_k(x)\hat{p}(x)dx = \alpha_k$ , $k = 1,2,\ldots,m$. The solution to this minimization problem is the maximum likelihood estimator in the exponential family

$$p_\theta(x) = p_0(x)\exp\{\theta_1\phi_1(x) + \ldots + \theta_m\phi_m(x) - \psi(\theta)\} \tag{3.1}$$

where $\psi(\theta) = \log \int p_0 \exp\{\theta_1\phi_1 + \ldots + \theta_m\phi_m\}$.

Let the random variables $X_1, X_2, \ldots, X_n$ be independent with probability density function $p(x)$ on $[0,1]$ which satisfies the smoothness condition $\int |D^r \log p|^2 < \infty$ for some $r \geq 1$. Take basis functions $\phi_k(x)$ $k = 1,\cdots,m$ to be polynomials, splines or trigonometric series of order $m = o(n^{1/2r+1})$. As we have shown in chapter 2, the Kullback-Leibler distance $D(p \| \hat{p}_n)$ converges to zero in probability at the rate $n^{2r/(2r+1)}$. However, the expected value $ED(p \| \hat{p}_n)$ can fail to converge, indeed in some cases $D(p \| \hat{p}_n)$ can be infinite in a set of small positive probability, for each n.

In this chapter we prove convergence of the expected value of Kullback-Leibler number for maximum posterior likelihood estimators. The prior $\pi_m(\theta)$ is assumed to make $\theta_1, \theta_2, \ldots, \theta_m$ independent Normal random variables. It is

believed that our results are valid for a large class of priors. However, the normal has been the easiest to analyze. The maximum posterior likelihood estimator ( *MPLE* ) is characterized as that member of the exponential family (3.1) for which the expected value of $\phi_k(X)$ satisfies

$$E_{p_\theta}\phi_k(X) = \frac{1}{n}\sum_{i=1}^{n}\phi_k(X_i) - \frac{d}{d\theta}\log \pi(\theta). \tag{3.2}$$

Another characterization of this estimator $\hat{p}_n$ is that it minimizes relative entropy from $p_0$ among all densities ( including those not in the family ) for which $\int \phi_k(x)p(x)dx$ satisfies the constraint (3.2). Subject to a restriction on the rate of growth of $m = m_n$, we show that

$$ED(p \| \hat{p}_n) \leq C_1\delta_m^2 + C_2\frac{m}{n} + o(\frac{m}{n}) \tag{3.3}$$

where the first term $\delta_m = (1/m)^r$ corresponds to the error in the best approximation of $p$ by densities in the family and the second term $m/n$ corresponds to the estimation error associated with inference in the family. When the dimension is chosen to be of order $m = o(n^{1/2r+1})$ the rate of convergence is

$$ED(p \| \hat{p}_n) = O((\frac{1}{n})^{2r/(2r+1)}). \tag{3.4}$$

Section 3.2 contains bounds for the Kullback-Leibler distance in terms of the Euclidean distance between the parameters in an exponential family. Some other useful lemmas are found there for the maximum posterior likelihood estimator. In Section 3.3 we prove the main result. A key trick is to adapt Hoeffding's inequality to handle large deviation events for $\hat{\theta}$. A specialization of inequalities in Hoeffding (1963) is as follows:

**Hoeffding's Inequality:** *If* $X_1, X_2, \ldots, X_n$ *are independent and* $a \leq X_i \leq b$, *then for* $t > 0$

$$P\{\bar{X} - u \geq t\} \leq \exp\{-2nt^2/(b-a)^2\}$$

where $u = E\bar{X}$, $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$.

### 3.2 Lemmas

We recall that $D(p \| q) = \int p(x)\log(p(x)/q(x))v(dx)$ is the relative entropy or Kullback-Leibler distance between $p$ and $q$. The reason the *MPLE* approach works is because this prior distribution forces the maximum of the posterior likelihood function to exist and to be unique. Thus it can avoid the chance that $D(p \| \hat{p}_n)$ is infinity. Some crucial bounds for the Kullback-Leibler distance are found in this section and its implications will be needed in proving the main theorem. We assume the prior is chosen such that the $\theta_k$ are independent Gaussian random variables with mean zero and variance $\sigma^2$ and such that the prior distribution of $\theta_1, \theta_2, \ldots, \theta_m$ has support equal to all of $\mathbf{R}^m$ for every $m \geq 1$. Given data $X^n = (X_1, X_2, \ldots, X_n)$, let $F_n(\theta) = \theta\cdot\alpha - \psi(\theta) - \frac{\|\theta\|^2}{2n\sigma^2}$ with $\alpha = \bar{\phi}_n = (1/n)\sum_{i=1}^{n}\phi(X_i)$. We observe that $nF_n(\theta)$ is the posterior log-likelihood function.

The following lemma demonstrates that the maximum of $F_n(\theta)$ exists and is unique for each fixed n.

**Lemma 3.1:** *Let*

$$F_n(\theta) = \theta\cdot\alpha - \psi(\theta) - \frac{\|\theta\|^2}{2n\sigma^2}. \tag{3.5}$$

*There exists a unique* $\theta_n(\alpha)$ *achieving the maximum value of* $F_n(\theta)$, *for each* $\alpha \in \mathbf{R}^m$ *and each* $n$.

**Proof of Lemma 3.1**: We observe that $F(\theta) = \theta\cdot\alpha - \psi(\theta)$ is equal to $1/n$ times the log-likelihood function , so $F(\theta)$ is a concave function. By the property of concavity, for every $\theta_0 \in \mathbf{R}^m$, we can find $a \in \mathbf{R}^m$ and $b \in R$ such that $F(\theta)$ is bounded by the hyperplane $\theta\cdot a + b$. Fix $n$ and let $g(\theta) = a\cdot\theta + b - \frac{\|\theta\|^2}{2n\sigma^2}$ where $a\cdot\theta + b$ is the tangent to $\theta\cdot\alpha - \psi(\theta)$ at $\theta_0$. We observe that $g(\theta)\rightarrow -\infty$ as $\|\theta\|\rightarrow\infty$.

Thus there exists $r$ such that $g(\theta) < g(\theta_0)$ for all $\theta$ with $\|\theta - \theta_0\| > r$. For such $\theta$,

$$F_n(\theta) \leq g(\theta) < g(\theta_0) = F_n(\theta_0). \tag{3.6}$$

Thus we have that the value of $F_n$ at $\theta_0$ is greater than the value for all $\theta$ outside the ball. Consequently the maximum exists in the ball. Since $F_n(\theta)$ is strictly concave, the maximum is unique. □

Let $\theta_n(\alpha)$ denote the unique maximizer of the function $F_n(\theta) = F(\theta) - \frac{\|\theta\|^2}{2n\sigma^2}$, where $F(\theta) = \theta\cdot\alpha - \psi(\theta)$ ,then $\theta_n(\alpha)$ has the following property.

**Lemma 3.2:** *If $F_n^*$ is the maximum value of $F_n(\theta)$, then for each $n$,*

$$\|\theta_n(\alpha)\|^2 \leq 2n\sigma^2 F^*. \tag{3.7}$$

**Proof of Lemma 3.2**: The value of $F(\theta) - \frac{\|\theta\|^2}{2n\sigma^2}$ at $\theta_n(\alpha)$ is greater or equal to the value at zero. Thus

$$F(0) - 0 \leq F(\theta_n(\alpha)) - \frac{\|\theta_n(\alpha)\|^2}{2n\sigma^2} \leq F_n^* - \frac{\|\theta_n(\alpha)\|^2}{2n\sigma^2}. \tag{3.8}$$

But $F(0) = \alpha\cdot 0 - \psi(0) = -\log\int e^{0\cdot\phi(x)}p(x)dv = 0$. Whence (3.8) yields $0 \le F_n^* - \frac{\|\theta_n\|^2}{2n\sigma^2}$ as desired. In the other words, $\theta_n$ is inside of the ball of squared radius $2n\sigma^2 F_n^*$. □

Let $\{p_\theta(x) = e^{\theta\cdot\phi(x) - \psi(\theta)}\}$ be a regular exponential family as in (3.1) with bounded functions $\phi_k$, $k=1,\ldots,m$, and a finite dominating measure $v(dx)$. We now assume that the functions $1,\phi_1,\ldots,\phi_m$ are chosen to be an *orthonormal* basis for $S_m$ with respect to a probability measure $Q$. Here $Q$ may be any probability measure having a density function $q = dQ/dv$ for which $\log q$ is bounded.

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

$$\|f_m\|_\infty \le A_m \|f_m\|_{L_2(Q)}. \tag{3.9}$$

We need to relate distances between the densities in the parametric family to distances between the parameters. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbf{R}^m$. This following lemma is the improvement of Lemma 4 of Chapter 2.

**Lemma 3.3**: *For* $\theta_0,\theta \in \mathbf{R}^m$

$$D(P_{\theta_0}\|P_\theta) \le Min\{\frac{b}{2}e^{A_m\|\theta_0-\theta\|}\|\theta_0-\theta\|^2, 2A_m\|\theta_0-\theta\|\} \tag{3.10}$$

and

$$D(P_{\theta_0}\|P_\theta) \ge \frac{1}{2be}Min\{\|\theta_0-\theta\|, \frac{1}{2A_m}\}\|\theta_0-\theta\| \tag{3.11}$$

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

**Proof of Lemma 3.3**: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 $D(P_{\theta_0}\|P_\theta) \le 2A_m\|\theta_0-\theta\|$. Then the inequality (3.10) follows from Lemma

4 of Chapter 2.

Let $\theta_t = \theta_0 + t \cdot b_m$ where $t \geq 0$ and $b_m = \frac{\theta - \theta_0}{\|\theta - \theta_0\|}$, $b_m \in \mathbf{R}^m$. Then since $D(P_{\theta_0} \| P_\theta) = (\theta_0 - \theta) \cdot \alpha_0 + \psi(\theta) - \psi(\theta_0)$, we have that $J(t) = D(P_{\theta_0} \| P_{\theta_t}) = -tb_m \cdot \alpha_0 + \psi(\theta_0 + tb_m) - \psi(\theta_0)$. It follows that $\frac{\partial}{\partial t} J(t) = -b_m \cdot \alpha_0 + b_m \cdot E_{\theta_t} \phi(x)$ and $\frac{\partial^2}{\partial t^2} J(t) = b_m \cdot \Sigma_{\theta_t} \cdot b_m^t \geq 0$ where $\Sigma_\theta = E(\phi(X) - \alpha_\theta)(\phi(X) - \alpha_\theta)^T$. Consequently, $J(t)$ is a convex function of t.

By the definition of $J(t)$, we observe $J(0) = 0$, and $J(\|\theta - \theta_0\|) = D(P_{\theta_0} \| P_\theta)$. So by convexity of $J(t)$, If $\|\theta - \theta_0\| \geq 1/(2A_m)$, we have

$$D(P_{\theta_0} \| P_\theta) \geq 2A_m \|\theta_0 - \theta\| J\left(\frac{1}{2A_m}\right)$$

$$\geq \frac{1}{2be} \frac{1}{2A_m} \|\theta_0 - \theta\|.$$

If $\|\theta - \theta_0\| \leq 1/(2A_m)$, $D(P_{\theta_0} \| P_\theta) \geq 1/(2be) \|\theta - \theta_0\|^2$ follows by applying (2.23) of Chapter 2 with $t = \|\theta - \theta_0\|$. This completes the proof of Lemma 3.3. □

**Remark 1:** If we choose $0 < \tau \leq 1$, then (3.11) can be generalized to

$$D(p_{\theta_0} \| p_\theta) \geq \frac{1}{2be^\tau} Min\{\|\theta_0 - \theta\|, \frac{\tau}{2A_m}\} \|\theta - \theta_0\|. \qquad (3.12)$$

**Remark 2:** To gain an intuitive feeling of this lemma from a geometric point of view, we find the largest $a$ such that the straight line $(y = ar)$ intersects the curve $(y = \frac{r^2}{2b} e^{-2A_m r})$. The solution occurs at $a = \frac{1}{2A_m} \cdot \frac{1}{2b} \cdot e^{-1}$.

In the proof of the theorem, we need a finer upper bound for the distance between $\theta_n(\alpha)$ and $\theta(\alpha_0)$. In other words, we need a bound which is approximated by $\|\alpha - \alpha_0\|$.

**Lemma 3.4:** *Let* $\theta_0 \in \mathbf{R}^m$, $\alpha_0 = \int \phi \, dP_{\theta_0}$, $\alpha \in \mathbf{R}^m$, *and* $\tau \in (0,1]$ *be given. Let* $b = e^{\|\log q/p_{\theta_0}\|_\infty}$ *and* $c = 2be^{\tau}$. *If* $\theta_n(\alpha)$ *maximizes* $F_n(\theta) = \theta\cdot\alpha - \psi(\theta) - \frac{\|\theta\|^2}{2n\sigma^2}$ *then*

$$\|\theta_n(\alpha) - \theta_0\|$$

$$\leq Max\{\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\| c, (\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\| - \frac{\tau}{2A_m c})n\}. \qquad (3.13)$$

**Proof of Lemma 3.4:** Let $F_n(\theta) = \theta\cdot\alpha - \psi(\theta) - \frac{\|\theta\|^2}{2n\sigma^2}$. Then since $D(P_{\theta_0}\| P_\theta) = (\theta_0 - \theta)\cdot\alpha_0 + \psi(\theta) - \psi(\theta_0)$, we have that for all $\theta \in \mathbf{R}^m$,

$$F_n(\theta_0) - F_n(\theta) = (\theta_0 - \theta)\cdot\alpha + \psi(\theta) - \psi(\theta_0) + \frac{\|\theta\|^2}{2n\sigma^2} - \frac{\|\theta_0\|^2}{2n\sigma^2}$$

$$= D(P_{\theta_0}\| P_\theta) - (\theta - \theta_0)\cdot(\alpha - \alpha_0) + \frac{\|\theta\|^2}{2n\sigma^2} - \frac{\|\theta_0\|^2}{2n\sigma^2}.$$

We observe that $\frac{\|\theta\|^2}{2n\sigma^2} - \frac{\|\theta_0\|^2}{2n\sigma^2} = \frac{1}{n}[\theta_0\cdot(\theta - \theta_0)] + \frac{1}{2n\sigma^2}\|\theta - \theta_0\|^2$. It follows from (3.12) and Cauchy-Schwartz inequality that for all $\theta \in \mathbf{R}^m$,

$$F_n(\theta_0) - F_n(\theta)$$

$$\geq \{\frac{1}{2be} Min\{\|\theta - \theta_0\|, \frac{\tau}{2Am}\} - \|\alpha - \alpha_0 + \frac{\theta_0}{n}\sigma^2\| + \frac{\|\theta - \theta_0\|}{2n\sigma^2}\}\|\theta - \theta_0\|.$$

This inequality is seen to be strict for $\theta \neq \theta_0$. Consider $\theta$ on the sphere

$\{\theta : \| \theta - \theta_0 \| = r\}$ where

$$r = Max\{\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\| c, (\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\| - \frac{\tau}{2A_m c}) 2n\sigma^2\}.$$

For all $\theta$ on this sphere, if $\|\theta - \theta_0\| \leq \frac{\tau}{2A_m}$ then

$$F_n(\theta_0) - F_n(\theta)$$

$$> (\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\| - \|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\|) r + \frac{r^2}{2n\sigma^2}$$

$$> 0.$$

If $\|\theta - \theta_0\| \geq \frac{\tau}{2A_m}$ then

$$F_n(\theta_0) - F_n(\theta) > (\frac{\tau}{2A_m} - \frac{\tau}{2A_m}) c + 0 = 0.$$

Thus the value of $F_n$ at $\theta_0$ ( inside the sphere ) is larger than all the values $F_n(\theta)$ on the sphere. Applying Lemma 1, it is then seen that $F_n$ has an extreme point $\theta^*$ which is inside the sphere i.e. $\|\theta^* - \theta_0\| < r$. Consequently, $\theta^* = \theta_n(\alpha)$. Therefore $\|\theta_n(\alpha) - \theta_0\| < r$. □

**Remark** :In Chapter 2, we needed to establish sufficient conditions for the distance between $\theta(\alpha)$ and $\theta(\alpha_0)$ to be bounded in terms of the distance between $\alpha$ and $\alpha_0$, but the probability that the conditions fail to be achieved is positive. In this chapter, for fixed n, we have found upper bound for the distance between $\theta_n(\alpha)$ and $\theta_0$, which are valid for all $\alpha$. These bounds are helpful in showing that $ED(p \| \hat{p}_{n,m})$ converges to zero at rate $n^{-2r/(2r+1)}$.

### 3.3 The MPLE Approach

Let $(\mathbf{X},\mathbf{B})$ be a measurable space and let $\nu$ be a fixed probability measure on this space. For $m \geq 1$, let $S_m$ be a linear space of dimension $m$ spanned by bounded measurable functions $\phi_{0,m}(x) = 1, \phi_{1,m}(x),\ldots,\phi_{m,m}(x)$. It is assumed that there exist positive numbers $a_m$ such that for all $f_m \in S_m$

$$\| f_m \|_\infty \leq a_m \| f_m \|_2. \tag{3.14}$$

Here $\| f \|_\infty$ is the essential supremum of $| f |$ and $\| f \|_2 = (\int f^2 d\nu)^{1/2}$. ( In particular $\|\phi_k\|_\infty \leq a_m$).

Let $\gamma_0$ and $\gamma$ be arbitrary positive constants. Let $f = \log p$ be bounded by $\gamma_0$ and assume that there exists $f_m \in S_m$ with $\| f - f_m \|_\infty \leq \gamma$ and $\| f - f_m \|_2 \leq \delta_m$ . Suppose $\delta_m$ is sufficiently small that satisfy $\delta_m a_m \leq c_0$ where $c_0 = 1/(4e^{\gamma_0+4\gamma+1})$.

Consider the family of probability density functions with respect to $\nu$ for which the logarithm of the density is in $S_m$. A parameterization of this family $\{ p_\theta(x) : \theta \in \mathbf{R}^m \}$ is

$$p_\theta(x) = \exp\{\sum_{k=1}^{m} \theta_k \phi_{k,m}(x) - \psi_m(\theta)\} \tag{3.15}$$

where $\psi_m(\theta) = \log \int \exp\{\sum \theta_k \phi_{k,m}(x)\}\, \nu(dx)$, $\theta \in \mathbf{R}^m$.

The prior density $\pi(\theta)$ is chosen such that the $\theta_k$ are independent Gaussian random variables with mean zero and finite variance $\sigma^2$ for $k \geq 1$. Given data $X^n = (X_1, X_2, \ldots, X_n)$ , then we call $\pi(\theta)p(X^n|\theta)$ the posterior likelihood function.

Let $\hat{p}_{n,m}$ be the density estimator, which maximizes the posterior likeli-

hood among all probability density functions with logarithm in $S_m$.

**Theorem 3.1:** *If* $\frac{ma_m^4}{n} = o(\frac{1}{\log n})$

*then*

$$ED(p \| \hat{p}_{n,m}) \leq C_1\delta_m^2 + C_2\frac{m}{n} + o(\frac{m}{n}) \tag{3.16}$$

*Here* $C_1 = (1/2)e^{\gamma+\gamma_0}$ *and* $C_2 = 2e^{2\gamma_0+2\gamma+1}$

**Remark**: Asymptotically, the constants may approach $C_1 = (1/2)e^{\gamma_0}$ and $C_2 = 2e^{2\gamma_0}$. This is the case if $\lim \| f - f_m \|_\infty = 0$, $\lim \delta_m a_m = 0$, $\lim m_n = \infty$, and $\lim m_n a_{m_n}^4 / n = 0$ as $n \to \infty$, so that $\gamma$ and $\tau$ may be chosen arbitrarily small for $n$ sufficiently large.

**Proof of the Theorem 3.1**: 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 $\nu$. 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 $\alpha_0 = \int \phi p_\beta d\nu$ where $\beta = (\beta_1,\ldots,\beta_m)$ and set $\alpha = \int \phi p d\nu$. Then the entries in the vector $\alpha - \alpha_0$ are given by $\int (p - p_\beta)\phi_k d\nu$ for $k = 0,1,\cdots,m$. By Bessel's inequality, we obtain

$$\begin{aligned} \| \alpha - \alpha_0 \| &\leq \| p - p_\beta \|_{L_2(\nu)} \\ &\leq e^{\gamma_0 + \| f - f_m \|_\infty - (\beta_0 + \psi(\beta))} \| f - f_m \|_{L_2(\nu)} \\ &\leq e^{\gamma_0 + 2\gamma}\delta_m . \end{aligned}$$

Now apply the same technique as in Chapter 2, but with $b = e^{\| \log p_\beta \|_\infty} \leq e^{\gamma_0 + 2\gamma}$, to conclude that $\theta^* = \theta(\alpha)$ exists and that

$\| \log p_{\theta^*} \|_\infty \leq \gamma_0 + 2\gamma + \tau$. For convenience, we can choose $\tau = 1$.

In Chapter 2, we chose $p_{\hat{\theta}}$ to be the maximum likelihood estimator and showed that $D(p_{\theta^*} \| p_{\hat{\theta}})$ is small with high probability under assumption (2.24). In this chapter we use $\hat{\theta}$ to be the maximum posterior likelihood estimator; we show that $ED(p_{\theta^*} \| p_{\hat{\theta}})$ is small. In particular, take $\theta_0$ to be $\theta^*$: the corresponding $\alpha_0$ is $\int \phi p^* dv$ (which is same as $\int \phi p dv$). For $\alpha$ take $\bar{\phi}_n = (1/n)\sum_{i=1}^n \phi(X_i)$. With these choices $\| \alpha - \alpha_0 \|^2 = \sum_{k=1}^m (\bar{\phi}_{n,k} - E_p \phi_k)^2$. By Lemma 3, the expected value of the estimation error satisfies

$$ED(p_{\theta^*} \| p_{\hat{\theta}}) \leq \frac{b}{2} e^\tau E \| \theta^* - \hat{\theta} \|^2 1_{\{\|\hat{\theta} - \theta^*\| \leq \frac{\tau}{a_m}\}}$$
$$+ 2a_m E \| \theta^* - \hat{\theta} \| 1_{\{\|\hat{\theta} - \theta^*\| > \frac{\tau}{a_m}\}}. \qquad (3.17)$$

Now apply Lemma 3.4. The first term of (3.17) is bounded by $\frac{b}{2} e^\tau E \| \alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2} \|^2 c^2$. So by the triangle inequality, the upper bound is $2b^3 e^{3\tau}(E_p \| \alpha - \alpha_0 \|^2 + E_p \| \frac{\theta_0}{n\sigma^2} \|^2)$, where $c = 2be^\tau$, $b = e^{\| \log p_{\theta^*} \|_\infty}$. From the fact that $X_1, \ldots, X_n$ are independent with common density $p$ and that the functions $\phi_k(X)$ are normalized to have zero mean and unit variance with respect to $\nu$, it follows that

$$E_p[\sum_{k=1}^m (\bar{\phi}_{n,k} - E_p \phi_k)^2] = \frac{1}{n} \sum_{k=1}^m Var_p(\phi_k(X))$$

$$\leq \frac{1}{n} \sum_{k=1}^m e^{\gamma_0} \int \phi_k^2 dv$$

$$= \frac{m}{n} e^{\gamma_0}.$$

Consequently, the first part of (3.17) is bounded by $C_2(m/n)$ plus a term

which goes to zero faster than $1/n$, as $n \to \infty$.

Now apply Lemma 3.4 and the Cauchy-Schwartz inequality. We obtain

$$E \|\theta^* - \hat{\theta}\| 1_{\{\|\hat{\theta} - \theta^*\| > \frac{\tau}{a_m}\}} \leq \tag{3.19}$$

$$n\sigma^2 (E(\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\| - \frac{\tau}{a_m c})^2)^{1/2} (P\{\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\|^2 > \frac{\tau^2}{a_m^2 c^2}\})^{1/2}.$$

We need to prove that the probability of $\{\|\alpha - \alpha_0 + \theta_0/n\sigma^2\|^2 > \tau^2/a_m^2 c^2\}$ is exponentially small. We know that

$$P\{\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\|^2 > \frac{\tau^2}{a_m^2 c^2}\}$$

$$\leq P\{\max_k (\bar{\phi}_k - E\phi_k + \frac{\theta_{0,k}}{n\sigma^2})^2 > \frac{\tau^2}{m a_m^2 c^2}\}$$

$$\leq \sum_{k=1}^{m} P\{(\bar{\phi}_k - E\phi_k + \frac{\theta_{\max}}{n\sigma^2} > \frac{\tau^2}{m a_m^2 c^2}\}$$

$$\leq \sum_{k=1}^{m} P\{(\bar{\phi}_k - E\phi_k) > \frac{\tau}{m^{1/2} a_m c} - \frac{\theta_{\max}}{n\sigma^2}\}$$

$$+ \sum_{k=1}^{m} P\{(-\bar{\phi}_k + E\phi_k) > \frac{\tau}{m^{1/2} a_m c} + \frac{\theta_{\max}}{n\sigma^2}\}. \tag{3.20}$$

Here $\theta_{\max} = \max_k |\theta_{0,k}| : k = 1, \ldots, m$. We know that $\theta_{\max} \leq (\sum_{k=0}^{m} \theta_{0,k}^2)^{1/2}$. By Parserval's identity, $(\sum_{k=0}^{m} \theta_{0,k}^2)^{1/2} = \|\log p_{\theta_0}\|$, and $\|\log p_{\theta_0}\| \leq \|\log p_{\theta_0}\|_\infty \leq \gamma_0 + 2\gamma + \tau$. Thus we find $\theta_{\max}$ is bounded by a constant which is not increasing with $m$.

Now we examine the terms in the bound (3.20). The assumption (3.14) implies that $-a_m \leq |\phi_k(X_i)| \leq a_m$. Set $t = \frac{\tau}{m^{1/2} a_m c} - \frac{\theta_{max}}{n\sigma^2}$. For all sufficiently large n ($n > m^{1/2} a_m c \theta_{max}/\tau\sigma^2$), under the assumption of the theorem, we get $t > 0$. Since $\phi_k(X_1), \phi_k(X_2), \ldots, \phi_k(X_n)$ are independent, then by Hoeffding's inequality

$$P\{(\bar{\phi}_k - E\phi_k) > \frac{\tau}{m^{1/2} a_m c} - \frac{\theta_{max}}{n\sigma^2}\}$$

$$\leq \exp\{-\frac{n\tau^2}{2ma_m^4 c^2} - \frac{\theta_{max}^2}{2a_m^2 n\sigma^4} + \frac{\tau\theta_{max}}{a_m^2\sigma^2(ma_m^2c^2)^{1/2}}\}.$$

By the same method it is seen that $P\{(-\bar{\phi}_k + E\phi_k) > \frac{\tau}{m^{1/2} a_m c} + \frac{\theta_{max}}{n\sigma^2}\}$ satisfies the same bound. Plugging these bounds into (3.20) and pulling out a common factor yields

$$P\{\|\alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\|^2 > \tau^2/a_m^2 c^2\}$$

$$\leq 2m\exp\{-\frac{n\tau^2}{2ma_m^4 c^2} - \frac{\theta_{max}^2}{2a_m^2 n\sigma^4} + \frac{\tau\theta_{max}}{a_m^2\sigma^2(ma_m^2c^2)^{1/2}}\}$$

$$\leq 2m\exp\{-n\,(\frac{\tau^2}{2ma_m^4 c^2} - \frac{\theta_{max}^2}{2a_m^2 n^2\sigma^4} + \frac{\tau\theta_{max}}{na_m^2\sigma^2(ma_m^2c^2)^{1/2}}\}$$

$$= 2m\exp\{-n\,(\frac{\tau^2}{2ma_m^4 c^2} - \varepsilon_{m,n})\}.$$

where $\varepsilon_{m,n} = \frac{\tau\theta_{max}}{na_m^2\sigma^2(ma_m^2c^2)^{1/2}} - \frac{\theta_{max}^2}{2a_m^2 n^2\sigma^4}$.

We check that $\varepsilon_{m,n} \le \frac{1}{2}(\tau^2/2ma_m^4c^2)$ for all large $n$, satisfying the assumption of the theorem. In particular, if $n \ge 4m^{1/2}a_m\theta_{max}c/\sigma^2\tau$ then $\varepsilon_{m,n} < \tau^2/(4ma_m^4c)$. Therefore

$$P\{\| \alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\|^2 > \tau^2/a_m^2c^2\} < 2m\exp\{-n(\frac{\tau^2}{4ma_m^4c^2})\}$$

for all large $n$. This bound is of smaller order than $m/n^c$ for any polynomial rate $c$, under the assumption of the Theorem.

Now we are prepared to handle the factors in (3.19). By expanding the square $E(\|\alpha - \alpha_0 + \frac{\theta_0}{n}\sigma^2 - \frac{\tau}{a_m c})^2$ and using (3.18) we have $E(\|\alpha - \alpha_0 + \frac{\theta_0}{n}\sigma^2\|^2) = O(m/n)$ and we get $E(\|\alpha - \alpha_0 + \frac{\theta_0}{n}\sigma^2\|) \le O(m^{1/2}/n^{1/2})$ by Jensen's inequality. We also see the constant part is order of $O(1/a_m^2)$. Thus $[E(\|\alpha - \alpha_0 + \frac{\theta_0}{n}\sigma^2 - \frac{\tau}{a_m c})^2]^{1/2}$ is bounded by $1/a_m$. We assume that $\frac{ma_m^4}{n} = o(\frac{1}{\log n})$ holds. More accurately, if we let $b_n = \frac{ma_m^4}{n}(4c^2\tau^2)$ then $1/(b_n \log n) \to \infty$ as $n \to \infty$ which implies that $1/(b_n \log n) - 4 \to \infty$ as $n \to \infty$ and $n^4\exp\{-1/b_n\} \to 0$. That means $n^2\exp\{-1/(2b_n)\} \to 0$. We obtain that

$$n^2[P\{\| \alpha - \alpha_0 + \frac{\theta_0}{n\sigma^2}\|^2 > \tau^2/a_m^2c^2\}]^{1/2} \le n^2(2m)^{1/2}\exp\{-1/(2b_n)\}.$$

Putting them back into (3.19), we shows that $2a_m E\|\theta^* - \hat{\theta}\|1_{\{\|\hat{\theta} - \theta^*\| > \frac{\tau}{a_m}\}}$ is order of $o(m^{1/2}/n)$. It is dominated by order of $o(m/n)$.

Finally, the expected value of Kullback-Leibler loss decomposes into a sum of approximation error and estimation error terms:

$$ED(p \| \hat{p}) = ED(p \| p^*) + ED(p^* \| \hat{p}).$$

The estimation error $ED(p^* \| \hat{p})$ has just been shown to be $C_2(m/n) + o(m/n)$. The approximation error satisfies

$$\begin{aligned} ED(p \| p^*) &= D(p \| p^*) \\ &\leq D(p \| p_\beta) \\ &\leq \frac{1}{2} e^{\| f - f_m \|_\infty + \| f \|_\infty} \| f - f_m \|_2^2 \\ &\leq \frac{1}{2} e^{\gamma + \gamma_0} \delta_m^2. \end{aligned}$$

This completes the proof of the Theorem. □

The examinations of the $L_2$ and $L_\infty$ approximation properties of polynomials , splines and trigonometric series are same with section 5 of chapter 2. The assumptions of these three cases are similar to chapter 2. For polynomial case : We choose $r \geq 3$ and assume $\frac{m^5 \log n}{n} \to 0$. For splines and trigonometric series cases : We choose $r \geq 2$ and $\frac{m^3 \log n}{n} \to 0$. More specific we choose $s = r$ in spline case and boundary conditions $f^{(j)}(0) = f^{(j)}(1)$ for $0 \leq j < r$ need to be satisfied in trigonometric case. Thus we have the following result.

**Corollary:** *If* $\log p \in W_2^r$ *and* $m \to \infty$ *then the expected value of Kullback-Leibler distance* $ED(p \| \hat{p}_{n,m})$ *converges to zero at rate*

$$(\frac{1}{m})^{2r} + \frac{m}{n}.$$

*In particular if m is chosen to be proportional to* $n^{1/(2r+1)}$ *then the expected value of Kullback-Leibler number converge to zero at rate* $n^{-2r/(2r+1)}$.

# CHAPTER 4 MULTIVARIATE DENSITY ESTIMATION

## 4.1 $L_2$ Bounds on Approximation

We have discussed several methods of density estimation in the univariate case in the previous chapters. Most of these methods can be generalized to the multivariate case, but the study of the large sample properties of these estimators becomes complicated by the dimensionality of the problem. Some basic structures are given in this section. We also discuss the $L_2$ approximation error in three different settings which are Fourier system, polynomials, and multivariate splines. Theorem 2.1 and Theorem 3.1 are applicable to multivariate cases, but we need to check the $L_2$ and $L_\infty$ assumptions given there. The motivation for examining the Kullback-Leibler number in multivariate case is not only for statistical reasons but also for its applications.

We first assume $I = [0,1]$, whose variable is denoted by $x$; $\Omega$ the product $I^d$ whose variable is denoted by $\mathbf{x} = (x^{(j)})_{j=1,\cdots,d}$; for a multi-integer $\mathbf{k} \in Z^d$, we set $|\mathbf{k}|^2 = \sum_{j=1}^{d} |k_j|^2$ and $|\mathbf{k}|_\infty = \max_{1 \le j \le d} |k_j|$. Set

$$L_2(\Omega) = \{\phi : \Omega \to C \mid \phi \text{ is measurable and } (\phi,\phi) < +\infty\}$$

equipped with the inner product

$$(\phi,\psi) = \int_\Omega \phi(\mathbf{x})\overline{\psi(\mathbf{x})}\, d\mathbf{x}.$$

Let $\|\cdot\|$ denote the $L_2$ norm on $\Omega$. For any positive integer $r^*$, set

$$\mathbf{H}_{r^*}(\Omega) = \{\phi \in L_2(\Omega) \mid \|\phi\|_{r^*} < +\infty\},$$

where

$$\|\phi\|_{r^*}^2 = \sum_{\substack{\mathbf{k}\in N^d \\ k_1 + k_2 + \cdots + k_d \le r^*}} \int_\Omega |(\prod_{j=1}^{d} D^{k_j})\phi|^2 d\mathbf{x}.$$

**I. Fourier System:** We consider the set $\{\phi_{\mathbf{k}} \mid \mathbf{k} \in Z^d\}$ with $\phi_{\mathbf{k}} = \exp\{2\pi i \mathbf{k}\cdot\mathbf{x}\}$ for $\mathbf{x}$ in $R^d$ which forms a completely orthonormal system in $L_2(\Omega)$. Let $r = (r_1, \cdots, r_d)$ be a multi-integer with nonnegative entries. Denote $|r| = r_1 + \cdots + r_d$, and define $D^r = \dfrac{\partial^{r_1}}{\partial x_1^{r_1}} \cdots \dfrac{\partial^{r_d}}{\partial x_d^{r_d}}$.

For any positive integer $m$, set $\mathbf{S}_m = span\{\phi_{\mathbf{k}} \mid \mathbf{k} \in Z^d \mid \mathbf{k}|_\infty \le m\} \subset L_2(\Omega)$. For any $f = \sum_{\mathbf{k}\in Z^d} \theta_{\mathbf{k}}\phi_{\mathbf{k}}$, we have $D^j\phi_{\mathbf{k}} = i2\pi k_j$. For functions $f$ that satisfy a boundary conditon, then a modification of C. Canuto and A. Quarteroni (1982 theorem 1.1) yields the following result.

**Theorem 4.1:** *For $r^* \ge 0$, there exists a constant $c = 1/(4\pi^2)$ such that $\|f - f_m\|^2 \le cm^{-2r^*} \|f\|_{r^*}$ for any $f \in \mathbf{H}^{r^*}(\Omega)$, satisfying the boundary condition $D^k f(0) = D^k f(1)$, for $0 \le \sum_{i=1}^{d} k_i \le r^*$.*

**Proof:** For any $f_m \in \mathbf{S}_m$, one has

$$\|f - f_m\|^2 = \sum_{|\mathbf{k}|_\infty > m} |\theta_{\mathbf{k}}|^2$$

$$\le \sum_{j=1}^{d} \sum_{\mathbf{k}:k_j > m} \frac{k_j^{2r^*}}{m^{2r^*}} |\theta_{\mathbf{k}}|^2$$

$$\le \frac{1}{4\pi^2 m^{2r^*}} \sum_{j=1}^{d} \sum_{\mathbf{k}} 4\pi^2 k_j^{2r^*} |\theta_{\mathbf{k}}|^2$$

$$= \frac{1}{4\pi^2 m^{2r^*}} \sum_{j=1}^{d} \int \left\| \frac{\partial r^*}{\partial x_j^{r^*}} f \right\|^2. \quad \square \tag{4.1}$$

**II. The Polynomial System:** Let $\{\phi_k\}_{k=0}^{\infty}$ be the orthogonal polynomials in $L_2([0,1])$ with $deg\,\phi_k = k$, then the system

$$\{\phi_{\mathbf{k}}\}_{\mathbf{k}\in N^d}, \quad \text{where } \phi_{\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{d}\phi_{k_j}(x^{(j)}),$$

is orthonormal and complete in $L_2(\Omega)$, where $N = \{0,1,2,...\}$. Thus any $v\in L_2(\Omega)$ can be written as

$$v = \sum_{\mathbf{k}\in N^d}\theta_{\mathbf{k}}\phi_{\mathbf{k}},\ \theta_{\mathbf{k}} = (v,\phi_{\mathbf{k}}) \tag{4.2}$$

with

$$\| v \|_0^2 = \sum_{\mathbf{k}\in N^d}|\theta_{\mathbf{k}}|^2.$$

Setting $\mathbf{S}_m = \mathbf{S}_m(\Omega) = span\,\{\phi_{\mathbf{k}} \mid \mathbf{k}\in N^d, |\,\mathbf{k}\,|_\infty \le m\}$ and $f_m = \sum_{\mathbf{k}\in \mathbf{S}_m}\theta_{\mathbf{k}}x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}$ where $\mathbf{x}\in[0,1]^d$, the dimension of $\mathbf{S}_m$ is equal to $(m+1)^d$. We denote

$$D^{\mathbf{r}}\phi_{\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{d}D^{r_j}\phi_{k_j}(x_j).$$

If we let

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in N^d}\theta_{\mathbf{k}}\phi_{\mathbf{k}}(\mathbf{x})$$

then

$$D^{\mathbf{r}}f(\mathbf{x}) = \sum_{\mathbf{k}}\theta_{\mathbf{k}}D^{\mathbf{r}}\phi_{\mathbf{k}}(\mathbf{x}).$$

The following identity is a multivariate extension of two identities in Cox (1988, equations 3.8 & 3.9 ).

$$\int_{[0,1]^d}(D^{\mathbf{r}}\phi_{\mathbf{k}}(\mathbf{x}))D^{\mathbf{r}}\phi_{\mathbf{l}}(\mathbf{x})(\mathbf{x}(1-\mathbf{x}))^{\mathbf{r}}d\mathbf{x} = \begin{cases}0 & , \text{if } \mathbf{l}\neq\mathbf{k}\\ c(\mathbf{k},\mathbf{r}) & , \text{if } \mathbf{l}=\mathbf{k}\end{cases} \tag{4.3}$$

where

$$c(\mathbf{k},\mathbf{r}) = \begin{cases} \prod_{j:r_j \neq 0} \{(k_j + r_j) \cdots (k_j - r_j + 1)\}; \mathbf{k} \geq \mathbf{r} \text{ and at least one } r_j > 0 \\ 1 \qquad\qquad\qquad\qquad\qquad\qquad ; \mathbf{r} = 0 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad ; \text{otherwise.} \end{cases}$$

Here $\mathbf{k} \geq \mathbf{r}$ means $k_j \geq r_j$ for every $j$. Consequently,

$$\int (D^{\mathbf{r}} f(\mathbf{x}))^2 (\mathbf{x}((1-\mathbf{x}))^{\mathbf{r}} = \sum_{\mathbf{k} \geq \mathbf{r}} c(\mathbf{k},\mathbf{r}) \theta_{\mathbf{k}}^2. \tag{4.4}$$

Right now, we want to find the minimax value of $c(\mathbf{k},\mathbf{r})$ under the constraint $\sum r_j \leq r^*$ in the complement set of $\mathbf{S}_m$. We assume that $r^*$ is a positive integer.

**Lemma 4.2 :** *If* $0 < r^* \leq m$, *then*

$$c_{m,r^*} \equiv \underset{\mathbf{k} \in \mathbf{S}_m^c}{Min} \; \underset{\mathbf{r}:\sum r_j \leq r^*}{Max} c(\mathbf{k},\mathbf{r}) = (m + 1 + r^*) \cdots (m - r^* + 2). \tag{4.5}$$

**Proof of Lemma 4.2:** We divide the proof into two parts. Letting $\mathbf{r}^* = (r^*,0,0, \ldots,0)$, the first part is to show that $c(\mathbf{k}^*,\mathbf{r}^*)$ maximizes $c(\mathbf{k}^*,\mathbf{r})$ for $\mathbf{r} \in N^d$ given $\mathbf{k}^* = (m+1,1,\ldots,1)$. While the second part is to prove that for each $\mathbf{k} \neq (m+1,1,1,\ldots,1)$ and $\mathbf{k} \in \mathbf{S}_m^c$, there exists at least one $\mathbf{r} = (r_1,\ldots,r_d)$ satisfying $\sum_{j=1}^{d} r_j \leq r^*$ such that $c(\mathbf{k},\mathbf{r}) \geq c(\mathbf{k}^*,\mathbf{r}^*)$.

For the first part, without loss of generality, we assume that $r_1 \geq r_2 \geq \cdots \geq r_d$ for any $\mathbf{r}$ satisfying $\sum_{j=1}^{d} r_j \leq r^*$. Whenever $r_j > 1$ for some $j = 2,\ldots,d$, such that $r_j > k_j^*$, then by the definition of $c(\mathbf{k},\mathbf{r})$, it follows that $c(\mathbf{k}^*,\mathbf{r}) = 0$. Thus the value of $\mathbf{r}$ to achieve the maximization of $c(\mathbf{k}^*,\mathbf{r})$ will occur at $r_j \leq k_j = 1$, $j = 2,\ldots,d$. It follows

$$c(\mathbf{k}^*,\mathbf{r}) = (m+1+r_1) \cdots ((m+1)-r_1+1) 2^{d-1-\#\{r_j: r_j=0,\, j=2,\ldots,d\}}. \tag{4.6}$$

From the assumption $\sum r_j \leq r^*$, it follows that

$$c(\mathbf{k}^*,\mathbf{r}) \le (m+1+r_1)\cdots((m+1)-r_1+1)\,2^{(r^*-r_1)} \tag{4.7}$$

This bound has a total of $r^* + r_1$ factors (writing $2^{(r^*-r_1)}$ as a product of 2's), which are term by term less than or equal to the terms in

$$c(\mathbf{k}^*,\mathbf{r}^*) = (m+1+r^*)\cdots((m+1)-r^*+1). \tag{4.8}$$

Now for the second part, let $k_1 \ge k_2 \ge \ldots \ge k_d$. If $k_1 > m+1$, then it is obvious that $c(\mathbf{k},r^*) \ge c(\mathbf{k}^*,r^*)$. If $k_1 = m+1$ , then $c(\mathbf{k},\mathbf{r}^*) \ge c(\mathbf{k}^*,\mathbf{r}^*)$ by choosing $r_1 = r^*$ and $r_j = 0$ for $j = 2,\cdots,d$. This completes the proof of the second part.

From those two parts, we obtain

$$c_{m,r^*} = c(\mathbf{k}^*,\mathbf{r}^*) = (m+1+r^*)\cdots(m-r^*+2). \tag{4.9}$$

This completes the proof of Lemma 4.2. □

**Remark:** Also $\underset{\mathbf{k}\in S_m^c}{Min}\ Max\ c(\mathbf{k},\mathbf{r}) = c_{m,r^*}$, where the $Max$ is taken over the set of $\mathbf{r}$ of the form $\mathbf{r} = (0,\ldots,0,r^*,0,\ldots,0)$, where the $r^*$ may be placed at any coordinate.

We can now derive the rate of convergence ( with respect to $N$ ) for the approximation error $f - f_m$ in the Sobolev norms.

**Theorem 4.2:** *For any* $0 < r^* \le m$, *if* $f \in \mathbf{H}_{r^*}(\Omega)$, *then*

$$\| f - f_m \|^2 \le \frac{1}{c_{m,r^*}} \sum_{\mathbf{r}:\sum r_j \le r^*} \| D^{\mathbf{r}} f \|^2. \tag{4.10}$$

*Moreover,*

$$\| f - f_m \|^2 \le \frac{1}{c_{m,r^*}} \left(\frac{1}{4}\right)^{r^*} \sum_{j=1}^{d} \left\| \frac{\partial^{r^*}}{\partial x_j^{r^*}} f \right\|^2. \tag{4.11}$$

**Remark:** The result (4.10) has previously been obtained ( see, e.g. C. Canuto

and A. Quarteroni (1982), J-L Lions (1987)). The result (4.1.12) is an improved bound. Note that it involves no cross-partial derivatives.

**Proof:** If $f_m \in \mathbf{S}_m$, we have

$$\| f - f_m \|^2 = \sum_{\mathbf{k} \in \mathbf{S}_m^c} \theta_{\mathbf{k}}^2$$

$$\leq \frac{1}{c_{m,r^*}} \sum_{\mathbf{k} \in \mathbf{S}_m^c} \underset{\mathbf{r}}{Max}\, c(\mathbf{k},\mathbf{r})\, \theta_{\mathbf{k}}^2$$

$$\leq \frac{1}{c_{m,r^*}} \sum_{j=1}^{d} \sum_{\mathbf{k} \in \mathbf{S}_m^c} c(\mathbf{k}, \mathbf{r}_j^*)\, \theta_{\mathbf{k}}^2$$

$$\leq \frac{1}{c_{m,r^*}} \sum_{j=1}^{d} \int \left( \frac{\partial r^*}{\partial x_j^{r^*}} f(\mathbf{x}) \right)^2 (x_j(1-x_j))^2 d\mathbf{x}$$

$$\leq \frac{1}{c_{m,r^*}} \left(\frac{1}{4}\right)^{r^*} \sum_{j=1}^{d} \left\| \frac{\partial r^*}{\partial x_j^{r^*}} f \right\|^2. \qquad (4.12)$$

Here the $Max_{\mathbf{r}} c(\mathbf{k},\mathbf{r})$ is taken over the set of $\mathbf{r}$ of the form $\mathbf{r}_j^* = (0, \ldots, 0, r^*, 0, \ldots, 0)$ where the nonzero value is in the $j^{th}$ position, $j = 1,2, \ldots, n$. This completes the proof of Theorem 4.2. □

**Remark:** Thus we have proved that prove the $L_2$ approximation error converges to zero as faster as $m^{-2r^*}$.

**III. Spline Systems:**

Piecewise polynomials play an important role in approximation theory and statistics. An important type of piecewise polynomial is what is known as a spline. We gave some properties of splines of univariate case in Chapter 2. We extend the same idea into the multivariate case. Let $\mathbf{S}_m$ be the space of multivariate splines of order $s$ on $[0,1]^d$. Here we have equal size boxes with

edge width $\Delta = 1/(m-s+2)$. This choice for $\Delta$ makes the dimension of $\mathbf{S}_m^s$ be $(m+1)^d$ as required. Fix $s$ and consider $m \geq s$. It is assumed that $f$ lies in the Sobolev space $W_2^k(\Omega)$; i.e.,

$$W_2^k(\Omega) = \{f \in L_2(\Omega) \mid \sum_{|\mathbf{r}| \leq k} \int_{\Omega} |D^{\mathbf{r}} f(\mathbf{x})|^2 \, d\mathbf{x} < \infty \}.$$

De Boor and Fix (1973) proved that there is a spline $f_m$ in $\mathbf{S}_m$ such that

$$\| f - f_m \|^2 \leq K'(\Delta)^{2r^*} \| f \|_{r^*}^2 \tag{4.13}$$

where $r^* = \min \{k, s\}$ and $K'$ is an absolute constant depending at most on the mesh ratios.

## 4.2 Main Results

Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ ($\mathbf{X}_j = (X_j^{(1)}, \cdots, X_j^{(d)})$ be multivariate independent random variables in $I^d$ with a common unknown density function $p(\mathbf{x})$ with respect to a known dominating measure $\lambda(d\mathbf{x})$. The asymptotics of density estimators is considered in terms of the Kullback - Leibler number.

$$D(p \| \hat{p}) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\hat{p}(\mathbf{x})} \lambda(d\mathbf{x}). \tag{4.14}$$

Let $\mathbf{S}_m$ be spanned by bounded measurable functions. It is assumed that there exists positive $\mathbf{a}_m$ such that for all $f_m \in \mathbf{S}_m$, $\| f_m \|_\infty \leq \mathbf{a}_m \| f_m \|_2$, here we state the conditions in terms of the $L_2(P)$ norm. Let $f(\mathbf{x}) = \log p(\mathbf{x})$ and $\gamma_0$, $\gamma$ be arbitrary positive constants. We consider functions $f$ for which there exists $f_m \in \mathbf{S}_m$ such that $\| f \|_\infty \leq \gamma_0$, $\| f - f_m \|_\infty \leq \gamma$ and $\| f - f_m \| \leq \delta_m$. It is assumed that the numbers $\delta_m$ and $\mathbf{a}_m$ satisfy $\delta_m \mathbf{a}_m \leq c_0$. We denote $C_m$ as the dimension of $\mathbf{S}_m$.

Let $\hat{p}_n$ be the multivariate density estimate, when it exists, which maximizes the likelihood among all probability density functions with logarithm in $\mathbf{S}_m$.

**Theorem 4.3 :** *For every* $1 \le K \le C_3 \frac{n}{C_m a_m^2}$,

$$P\{D(p \| \hat{p}_n) \ge C_1 \delta_m^2 + C_2 \frac{C_m}{n} K\} \le \frac{1}{K} \tag{4.15}$$

*Here* $C_1 = (1/2)e^{\gamma+\gamma_0}$, $C_2 = 2e^{2\gamma+2}$ *and* $C_3 = (1/16)e^{-4\gamma-2}$.

This is the same result as Theorem 2.1. Notationally the only difference is that here the dimension of the approximating linear space $S_m$ is $C_m$ instead of m. For the product basis funtions used below $C_m = (m+1)^d$. We have repeated the statement of the theorem for convenience in checking the condition in the multivariate case.

The assumptions of Theorem 4.3 need to be proved in the three different cases. We did the examination of $L_2$ approximations in Section 4.1. Now we examine $L_\infty$ approximation properties of polynomials, splines and trigonometric series. Given a linear space $S_m$, define $A_m(Q) = A_m(Q, S_m)$ by

$$A_m(Q) = \sup_{f \in S_m} \frac{\|f\|_\infty}{\|f\|_{L_2(Q)}}.$$

**Lemma 4.4 :**

$$A_m(Q) = \sup_x \Big( \sum_{|k| \le m} \phi_k^2(x) \Big)^{1/2}. \tag{4.16}$$

*where the* $\phi_k$ *are orthonormal basis with respect to* $Q$ *and span* $S_m$.

**Proof:** For any $f \in S_m$,

$$\|f\|_\infty = \sup_x \sum \theta_k \phi_k(x)$$

$$\le \sup_x \Big( \sum_{|k| \le m} \phi_k^2(x) \Big)^{1/2} \|\theta\|_2$$

$$= \sup_x \| \phi(x) \| \, \|f\|_2.$$

Here we have used the Cauchy-Schwartz inequality. Let $x^*$ achieve $\sup_x ( \sum_{|k| \leq m} \phi_k^2(x))^{1/2}$. Then equality in the Cauchy-Schwartz is achieved at $x = x^*$, when $\theta_k$ is proportional to $\phi_k(x^*)$. This completes the proof. □

**Remark:** We showed in Chapter 2 that with $Q$ = Lebesgue measure on [0,1], $a_m = A_m(Q)$ is equal to $m+1$ in the polynomial case, $\sqrt{m+1}$ in the trigonometric case, and in the spline case it is bounded by $a_m \leq s\sqrt{m-s+2}$. To handle the multivariate case, we simply take the $d$ *th* power as the following Lemma shows.

Assume that $\mathbf{S}_m$ is the linear space of d-dimensional functions spanned by the functions $\phi_{\mathbf{k}}(\mathbf{x}) = \prod_{j=1}^{d} \phi_{k_j}(x^{(j)})$ which are products of one-dimensional basis functions $\phi_k(x)$ which span $S_m$. Let $Q^d$ denote the product measure.

**Lemma 4.5:** $A_m(Q^d, \mathbf{S}_m) = (A_m(Q, S_m))^d$ .

**Proof:** By Lemma 4.4

$$A_m(Q^d, \mathbf{S}_m) = \sup_{\mathbf{x}} \Big( \sum_{|\mathbf{k}| \leq m} \prod_{j=1}^{d} \phi_{k_j}^2(x_j) \Big)^{1/2}$$

$$= \sup_{\mathbf{x}} \prod_{j=1}^{d} \Big( \sum_{k=1}^{m} \phi_k^2(x_j) \Big)^{1/2}$$

$$= \prod_{j=1}^{d} \Big( \sup_{\mathbf{x}} \sum_{k=1}^{m} \phi_k^2(x_j) \Big)^{1/2}$$

$$= (A_m(Q, S_m))^d. \; \square$$

Therefore, in the multivariate case $\mathbf{a}_m$ is proportional to $m^d$ for the polynomial

system and and it is proportional to $m^{d/2}$ for the Fourier and spline system.

**Polynomials:** Now we examine the $L_\infty$ error for the Legendre polynomial approximation, assuming $\sum_{j=1}^{d} \| \frac{\partial r^*}{\partial x_j^{r^*}} f \|^2 \leq \infty$. We extend the technique from one dimension to multiple dimensions. For $m \geq r^*$ and $r^* > d$ we apply the Cauchy-Schwartz inequality to the series $\sum \theta_{\mathbf{k}} \phi_{\mathbf{k}}(\mathbf{x})$. We found

$$\|f - f_m\|_\infty \leq \Big( \sum_{\mathbf{k} \in S_m^c} \theta_{\mathbf{k}}^2 \operatorname*{Max}_{\mathbf{r}} c(\mathbf{k},\mathbf{r}) \Big)^{1/2} \Big( \sum_{\mathbf{k} \in S_m^c} \frac{\phi_{\mathbf{k}}^2(\mathbf{x})}{\operatorname*{Max}_{\mathbf{r}} c(\mathbf{k},\mathbf{r})} \Big)^{1/2}. \qquad (4.17)$$

The first factor on the right side of (4.17) is bounded by $(1/4)^{r^*} \sum_{j=1}^{d} \| \frac{\partial r^*}{\partial x_j^{r^*}} f \|^2$. We observe that $S_m^c = \{\phi_{\mathbf{k}} : |\mathbf{k}|_\infty > m\}$ $= \bigcup_{i=1}^{\infty} \{\phi_{\mathbf{k}} : |\mathbf{k}|_\infty = m+i\}$. Therefore,

$$\sum_{\mathbf{k} \in S_m^c} \frac{\phi_{\mathbf{k}}^2(\mathbf{x})}{\operatorname*{Max}_{\mathbf{r}} c(\mathbf{k},\mathbf{r})} = \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{\substack{\mathbf{k}: k_j = m+i, \\ k_l \leq m+i, l \neq j}} \frac{\prod_{j=1}^{d} \phi_{k_j}^2(x_j)}{\operatorname*{Max}_{\mathbf{r}} c(\mathbf{k},\mathbf{r})}.$$

By the bound on Legendre polynomials, $|\phi_k(x)|^2 \leq 2k+1$ it follows that

$$\sum_{\mathbf{k} \in S_m^c} \frac{\phi_{\mathbf{k}}^2(\mathbf{x})}{\operatorname*{Max}_{\mathbf{r}} c(\mathbf{k},\mathbf{r})} \leq \sum_{j=1}^{d} \sum_{i=1}^{\infty} \sum_{\substack{\mathbf{k}: k_j = m+i, \\ k_l \leq m+i, l \neq j}} \frac{(2(m+i)+1)^d}{(m+i+r^*)(m+i+r^*-1)\ldots(m+i-r^*+1)}.$$

The sets $\{\mathbf{k}: k_j = m+i, k_l \leq m+i, l \neq j\}$ have $(m+i)^{(d-1)}$ terms. Consequently,

$$\sum_{\mathbf{k} \in S_m^c} \frac{\phi_{\mathbf{k}}^2(\mathbf{x})}{\operatorname*{Max}_{\mathbf{r}} c(\mathbf{k},\mathbf{r})} \leq 2^d \sum_{j=1}^{d} \sum_{i=1}^{\infty} \frac{(m+i+1)^d (m+i)^{d-1}}{(m+i+r^*)\ (m+i+r^*-1)\ldots(m+i-r^*+1)}.$$

For $r^* > d$, we found $(m+i+r^*)(m+i-(2d-r^*-2)) \geq (m+i+1)(m+i)$. Then we obtain

$$\sum_{\mathbf{k}\in S_m^c} \frac{\phi_{\mathbf{k}}^2(\mathbf{x})}{\underset{\mathbf{r}}{Max}\, c(\mathbf{k},\mathbf{r})} \le 2^d \sum_{j=1}^{d}\sum_{i=1}^{\infty} \frac{1}{(m+i-r^*+1)^{2r^*-2d+1}}$$

$$= 2^d \sum_{j=1}^{d} \sum_{k=m-r^*+2}^{\infty} \frac{1}{k^{2r^*-2d+1}}$$

$$\le 2^d \sum_{j=1}^{d} \int_{m-r^*+1}^{\infty} \frac{1}{x^{2r^*-2d+1}} dx$$

$$= \frac{d\ 2^d}{2(r^*-d)} \frac{1}{(m-r^*+1)^{2r^*-2d}}.$$

Therefore $\gamma_m$ is proportional to $(1/m)^{(r^*-d)}$.

**Remark:** An alternative proof of this rate can be obtained by using same method deriving in univariate case. We obtain

$$\|f - f_m\|_\infty \le 2e^{2r^*}(r^*-d)\left(\frac{1}{m+r^*}\right)^{2(r^*-d)} \left(\frac{1}{4}\right)^{r^*} \sum_{j=1}^{d} \left\| \frac{\partial r^*}{\partial x_j^{r^*}} f \right\|^2.$$

**Splines:** DeBoor and Fix ( 1973 ) extend the quasinterpolant construction to include functions of d variables. Let $S_m^s$ be the space of multivariate splines on $[0,1]^d$ of order $s$ with knots equally spaced in each coordinate. The width $\Delta$ was defined as previous. They proved that the global error estimate is approximately bounded by modulus of continuity multiply $(\Delta)^{r^*}$, where $r^* = \min\{k, s\}$. Then applying Cauchy-Schwartz inequality, there is a constant $K$ such that

$$\|f - f_m\|_\infty \le K(\Delta)^{r^* - 1/2} \|f\|_{r^*}. \qquad (4.18)$$

**Fourier System:** The boundary conditions $D^k \log p(0) = D^k \log p(1)$, for $\mathbf{k} < \mathbf{r}$ are required. For simplicity we focus on the case that $m$ is even. The approximation rates are easier to obtain in the Fourier case than in the polyno-

mial case. By a similar technique, we obtain the bound

$$\| f - f_m \|_\infty \leq \left( \sum_{j=1}^{d} \int \| \frac{\partial r^*}{\partial x_j^{r^*}} f \|^2 \right)^{1/2} \left( \frac{d(2r^*-d)}{m^{(2r^*-d)}} \right)^{1/2}. \tag{4.19}$$

Thus $\| f - f_m \|_\infty \leq O(m^{-(r^*-d/2)})$, for $f$ in $\mathbf{H}_{r^*}(\Omega)$.

We need to choose $r^* > d$ and assume that $m^{2d+1}/n \to 0$ for the polynomial system. For the spline and Fourier systems, we choose $r^* > d/2$ and $m^{d+1}/n \to 0$. This is needed so that the following result can be achieved.

**Corollary 4.3:** *If* $\log p \in \mathbf{H}^{r^*}(\Omega)$ *and* $m \to \infty$ *then the Kullback-Leibler number* $D(p \| \hat{p}_n)$ *converges to zero in probability at rate*

$$\left( \frac{1}{m+1} \right)^{\frac{2r^*}{d}} + \frac{(m+1)^d}{n}. \tag{4.20}$$

*In particular, if* $m$ *is chosen to be proportional to* $n^{1/(2r^*+d)}$ *then the Kullback-Leibler number* $D(p \| \hat{p}_n)$ *converges to zero in probability at rate* $n^{-2r^*/(2r^*+d)}$.

Is it possible to get the rate of convergence of the expected value of the Kullback-Leibler loss for multivariate case? The answer is yes. We are going to discuss the details in this section.

Let random variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent with unknown probability density function $p(\mathbf{x})$ with respect to a known dominating measure $\lambda(d\mathbf{x})$. This density function $p(\mathbf{x})$ is assumed to satisfy the smoothness condition $\sum_{|\mathbf{r}| \leq r^*} \int |D^{\mathbf{r}} f(\mathbf{x})|^2 d\mathbf{x} < \infty$. For a given set of functions $\phi_1(\mathbf{x}), \cdots, \phi_{C_m}(\mathbf{x})$ and a density function $p_0(\mathbf{x})$, consider the exponential family of densities

$$p_\beta(\mathbf{x}) = p_0(\mathbf{x}) \exp\{\beta_1 \phi_1(\mathbf{x}) + \cdots + \beta_{C_m} \phi_{C_m}(\mathbf{x}) - \psi(\beta)\} \tag{4.21}$$

where $\psi(\beta) = \log \int p_0 \exp\{\beta_1\phi_1 + \cdots + \beta_{C_m}\phi_{C_m}\}$. This family is characterized by the property that given a constraint on the expected values $E_p\phi_k(\mathbf{x}) = \alpha_k$, $k = 1, \cdots, C_m$, the density which minimizes the relative entropy $D(p \| p_0)$ subject to this constrained is in the family (4.21).

We assume the prior information is chosen such that $\beta_k$ are independent normal random variables with mean zero and variance $\sigma^2$ ( or $\sigma_k^2$ ). Once we draw the random sample $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ from a probability distribution $P$ which has a density $p$ with respect to $\nu$, we can find the corresponding maximum posterior likelihood estimator which depends on the sample sizes.

Let $\hat{p}_{n,m}$ be the multivariate density estimate, which maximizes the posterior likelihood among all probability density functions with logarithm in $\mathbf{S}_m$. We now restate Theorem 3.1 in the present context. The same $L_2$ and $L_\infty$ approximation conditions are required. $C_m$ is the dimension of $\mathbf{S}_m$, which is $(m+1)^d$ for the linear space product bases that we have examined.

**Theorem 4.4:** *If* $\dfrac{C_m \mathbf{a}_m{}^4}{n} = o(\dfrac{1}{\log n})$ *then*

$$ED(p \| \hat{p}_{n,m}) \leq C_1\Delta_m^2 + C_2\frac{C_m}{n} + o(\frac{C_m}{n}). \tag{4.22}$$

*Here* $C_1 = (1/2)e^{\gamma + \gamma_0}$ *and* $C_2 = 2e^{2\gamma_0 + 2\gamma + 1}$

We choose $r^* > 2d$ and assume $m^{5d}(\log n)/n \to 0$ in polynomial system, choose $r^* > d$ and assume $m^{3d}(\log n)/n \to 0$ in the other two systems. Thus we have the following result.

**Corollary 4.4:** *If* $\log p \in \mathbf{H}_{r^*}(\Omega)$, *then the expected value of Kullback-Leibler distance* $ED(p \| \hat{p}_{n,m})$ *converges to zero at rate*

$$(\frac{1}{C_m})^{2r} + \frac{C_m}{n}$$

*If we choose* $m$ *to be proportional to* $n^{1/(2r^* + d)}$, *then* $ED(p \| \hat{p}_{n,m})$ *converges to zero at rate* $n^{-2r^*/2r^* + d}$.

The verification of the conditions proceeds in the same way as for Corrollary 4.3.

Stegbuchner

Sterbuchner (1977) set $\mathbf{S}_m = \{\mathbf{k}: \prod_{j=1}^{d} \max\{1, |k_j|\} \leq m, \mathbf{k} \in Z^d\}$ for estimating density in the unit cube in $R^d$. The cardinal number of $\mathbf{S}_m$ is of order $\mathbf{S}_m \sim cm \log^{d-1} m$, where the constant $c$ is independent of $d$. He considered a nonparametric multivariate density estimator, based on multidimensional Fourier series and kernel function. He show that the rate of convergence is of order $n^{-2r/(2r+1)}(\log n)^{d-1}$ which is much better than order $n^{-2r/(2r+d)}$; however his smoothness requirement on the density is different. We tried several different choices for $\mathbf{S}_m$, compared the relationships among them, and decided to choose the one we defined in the beginning of this chapter. The rate of convergence of Kullback-Leibler number is of order $n^{-2r/(2r+d)}$.

In most practical problems, even when using a high-speed computer it is important to take some care in calculating multivariate density estimates. For example, if we assume that $\gamma_0 = 1/2$ and $\gamma = 1$ and $r^* = 3$ then we almost need $10^8$ samples when $d = 2$ for polynomial case, $10^4$ samples for spline or fourier system to meet the requirements of assumption. If we extend the dimension to $d > 3$, then the minimum sample size increases extremely fast to fit the assumptions. Connections between high-dimensional fitting in statistics and approximation theory are being explored.

Even with large data sets, fitting unrestricted nonparametric models when the number of independent variables is large leads to unreliable predictions. This has come to be called the "curse of dimensionality" and can be viewed as a variation on the theme that it is not so much the number of observations that matters but rather the number of observations per parameter. To deal with the "curse of dimensionality", the additive regression model of Stone (1988) has been proposed and are beginning to be studied. Projection pursuit method has been presented by Friedman (1974,1981,1984) and Huber (1985). Barron and Barron (1988) show that the concept of the statistical learning networks provides a unifying framework which encompasses traditional neural network models, adaptively synthesized networks and nonparametric statistical techniques. Future work is needed to make comparisons among different modeling methods (such as spline smoothing, projection pursuit and learning network) and different model selection criterions (such as cross validation, Akaike's information criterion and the minimum complexity criterion).

# CHAPTER 5 APPLICATIONS

## 5.1 Coding

As we pointed out in the beginning of this thesis, the motivation for examining the Kullback-Leibler number and its expectation is not only the mathematical results but also its applications. We discuss two applications in this chapter. We determine the rate of convergence of the redundancy of a code for data compression based on $\hat{P}$ compared to the optimal code based on the unknown $P$. In Section 5.2 we apply the results we obtained in Chapter 4 to compare the relationship between the actual wealth and the optimal wealth, in a stock market setup.

Suppose $X$ is a discrete random variable with a countable range space $A$ and a probability mass function $p(x)$, $x \in A$. By the theorem of Kraft (1955), there exists a uniquely decodable code $\Phi: A \to \{0,1\}^*$ with code length $l(\Phi(x))$, if and only if

$$\sum_{a \in A} 2^{-l(\Phi(x))} \leq 1. \tag{5.1}$$

In particular, if $q$ is any probability mass function on $A$ then there exists a code (called the Shannon code with respect to $q$) with length $\lceil \log 1/q(x) \rceil$ The redundancy of a code is the difference between its expected length and the entropy.

Let $X_1, X_2, \ldots, X_n$ be independent random variables with unknown distribution $P$. Thus the true distribution for $X^n = (X_1, X_2, \ldots, X_n)$ is $P^n$, the product measure on $A^n$. We are free to choose a Shannon code with respect to

any probability mass function $q^n$ on $A^n$. The redundancy (per sample) is

$$R_n(p) = \frac{1}{n}(E_P \lceil \log \frac{1}{q^n(X^n)} \rceil - E \log \frac{1}{p^n(X^n)}). \tag{5.2}$$

We note that

$$\frac{1}{n} D(p^n \| q^n) \leq R_n(p) \leq \frac{1}{n} D(p^n \| q^n) + \frac{1}{n}. \tag{5.3}$$

We let $\hat{p}_k$ be an estimator of $p(x)$ based on the data $x^{k-1} = (x_1, \ldots, x_{k-1})$, $k = 1,2, \ldots, n$. ( $\hat{p}_1 = p_0$ is an initial guess based on no data ). It is required that for each $x^{k-1}$, $\hat{p}_k$ is non-negative and sums to one. Then the sequence of estimates for $k = 1,2, \ldots, n$ yields the following joint probability mass functions.

$$q^n(x_1, \ldots, x_n) = \prod_{k=1}^{n} q(x_k \mid x_1, \ldots, x_{k-1}) = \prod_{k=1}^{n} \hat{p}_k(x_k). \tag{5.4}$$

Note that with respect to $Q^n$, the random variables $X_1, X_2, \ldots, X_n$ are no longer independent. Now by chain rule

$$D(p^n \| q^n) = \sum_{k=1}^{n} E \log \frac{p(x_k)}{q(x_k \mid x_1, \cdots, x_{k-1})} = \sum_{k=1}^{n} E \log \frac{p(x_k)}{\hat{p}_k(x_k)}. \tag{5.5}$$

The terms in the sum are just $ED(p \| \hat{p}_k)$, which we recognize as the relative entropy risk of the density estimator with a sample of size k. Dividing by the sample size n, the relative entropy between $p^n$ and $q^n$ is the Cesaro average of the risk

$$\frac{1}{n} D(p^n \| q^n) = \frac{1}{n} \sum_{k=1}^{n} ED(p \| \hat{p}_k). \tag{5.6}$$

Thus the code with lengths $l(\Phi(X^n) = \lceil \log \frac{1}{\prod_{k=1}^{n} \hat{p}_k(X_k)} \rceil$ has redundancy $R_n(p)$ within $1/n$ of $\frac{1}{n} \sum_{k=1}^{n} ED(p \| \hat{p}_k)$.

In the case of continuous random variables with density $p(x)$ and density estimates $\hat{p}_k(x) = \hat{p}_k(x : x_1, \ldots, x_{k-1})$ for $k = 1,2, \ldots, n$, the same construction leads to a probability measure $Q^n$ on $X^n$ with joint density $q^n(x_1, \ldots, x_n) = \prod_{k=1}^{n} \hat{p}_k(x_k ; x_1, \ldots, x_{k-1})$. and $D(p^n \| q^n) = \sum_{k=1}^{n} ED(p \| \hat{p}_k)$.

In the continuous case, the real values random variables $(X_1, X_2, \ldots, X_n)$ can not be described exactly. Nevertheless, if we consider quantizations of the real line, $D(p^n \| q^n)$ is seen to provide a bound on the redundancy. Indeed for any partition $\Pi$, let $D^{\Pi}(p^n \| q^n) = \sum p^n(B) \log p^n(B)/q^n(B)$ where the summation is for all rectangles $B \in \Pi^n$. It is known that $D^{\Pi}(p^n \| q^n) \leq D(p^n \| q^n)$ uniformly over all partitions. Moreover $D = \sup_{\Pi} D^{\Pi}$ for sufficiently fine quantizations ( Pinsker (1964) ).

If the Shannon code based on $q^n$ is used to describe the quantized data, then the redundancy $R_n^{\Pi}(p)$ is within $1/n$ of the relative entropy $(1/n)D^{\Pi}(p^n \| q^n)$ which is made arbitrarily close to $(1/n)D(p^n \| q^n)$ by the choice of a sufficiently fine partition. Defining $R_n(p) = \sup_{\Pi} R_n^{\Pi}(p)$ to be the redundancy in the continuous case, we see as (5.4) and (5.5) that the redundancy of the code based on $\prod_{k=1}^{n} \hat{p}_k(x_k)$ is $(1/n)D(p^n \| q^n)$ to within $1/n$ bits.

We gave conditions in Section 4 such that $ED(p \| \hat{p}_n) = O((1/n))^{2r/(2r+1)}$. Consequently, we have the following.

**Proposition 5.1:** *The redundancy $R_n(p)$ for densities with $\int (D^r \log p)^2 < \infty$ is*

$$R_n(p) = O(\frac{1}{n})^{\frac{2r}{2r+1}}. \tag{5.7}$$

*Moreover, this bound holds uniformly over log-densities in a Sobolev ball. Conse-*

*quently, $n^{-2r/(2r+1)}$ bounds the minimax redundancy.*

**Proof:** We assume the initial guess $\hat{p}_1 = p_0$ satisfies the condition $ED(p \| \hat{p}_1) \leq c$ and that $ED(p \| \hat{p}_k) \leq c(1/k)^{2r/2r+1}$. By (5.6), it follows that

$$R_n(p) \leq \frac{1}{n}\sum_{k=1}^{n} c(\frac{1}{k})^{2r/(2r+1)} + \frac{1}{n}$$
$$= c'n^{-2r/(2r+1)} + n^{-1}.$$

This complete the proof of this proposition. □

**Remarks:** (i) This redundancy contrasts with the rate $(\log n)/n$ which could be achieved if $P$ were known to be a member of a smooth finte dimensional-parametric family (Clarke and Barron (1988)). Indeed a redundancy of $(\log n)/n$ obtains using any estimator for which $D(P \| \hat{P}_n) = O(1/n)$. (Note that the series $1/k$ $k = 1, \ldots, n$ yields a sum of order $(\log n)/n$ instead of $1/n$). (ii) Barron and Cover (1989) give another universal code for the same class of densities. The overall length of the code is (within one bit of) $L_n(\hat{p}_n) + \log 1/\hat{p}_n(X^n)$, (where $\hat{p}_n$ is chosen to achieve $\min_q(L_n(q) + \log 1/q(X^n))$ over a class of densities $\Gamma$, and $L_n(q), q \in \Gamma$ denotes the codelength for q). With a specific choice for $\Gamma$, they proved the redundancy of the minimum two-stage code is bounded by $O(n^{-2r/(2r+1)} \log n)$ for the same class of densities as in propositions 5.1.

## 5.2 Portfolio Selection

Almost everyone owns a portfolio (group) of assets. This portfolio is likely to contain real assets such as a car, a house, or a book, as well as financial assets such as stocks and bonds. An invester is faced with a choice from

among an enormous number of assets. When one considers the number of possible assets and the various possible proportions in which each can be held, the decision process seems overwhelming. Portfolio analysis is concerned with finding the most desirable group of assets to hold, given the properties of each of the assets, and with showing how the composition of the preferred portfolios can be determined.

Now we give a simple example and learn from it. A decision maker is faced with a collection of stocks $\mathbf{X} = (X_1, X_2, \ldots, X_d)$, where $X_i$ is the number of units returned from an investment of one unit in the ith stock. We assume that $X_i$ are nonnegative. A portfolio $\mathbf{b} = (b_1, \ldots, b_d)$, $b_i \geq 0$, $\sum_i b_i = 1$, is the proportion of the current capital invested in each of the d stocks. The decision maker is trying to find the best portfolio in the set of all portfolios **b**.

The return on a portfolio of assets is simply a weighted average of the return on the individual assets. The weight applied to each return is the fraction of the portfolio invested in that asset. If **S** is the return on the portfolio and $b_i$ is the fraction of the investor's funds invested in the ith asset, then

$$\mathbf{S} = \mathbf{b}^t \mathbf{X} = \sum_{i=1}^{d} b_i X_i. \tag{5.8}$$

A criterion for selecting **b**, that of maximizing $E \log \mathbf{S}$, has been put forth by Kelly (1956) and Breiman (1961). Cover and Gluss (1986) consider sequential investment in a stock market with the goal of performing as well as if we know the empirical distribution of future market performance. In particular, they wish to outperform the best stock. Barron and Cover (1988) show that the increase in exponential growth of wealth achieved by the knowledge of the stock market distribution $F$ over that achieved under incorrect knowledge

$G$ is bounded above by Kullback-Leibler number $D(F \| G)$.

Haugen (1986) discuss the correlation (relationship) among the individual assets, and derive a single and multiple model of assets in parametric family of dependence among the stocks. But if we can estimate the unknown joint distribution $P$ for the stocks, then we can analyze the relationship among them. It motivate us to estimate the unknown distribution function $P$.

If $\mathbf{b}$ is used for $n$ investment periods, then the stock sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ results in wealth $\mathbf{S}_n$ at time $n$ given by

$$\mathbf{S}_n = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{X}_i$$

$$= 2^{n(\frac{1}{n}\sum_{i=1}^{n} \log \mathbf{b}^t \mathbf{X}_i)}$$

The logarithm in here is to the base 2. Barron and Cover (1988) define the doubling rate $W(\mathbf{X})$ for the market by

$$W(\mathbf{X}) = \max_{\mathbf{b}} \int \log \mathbf{b}^t \mathbf{x} \, dF(\mathbf{x}). \tag{5.9}$$

The Kuhn-Tucker conditions ( $E(X_i / \mathbf{b}^t \mathbf{X}) = 1$, for $b_i > 0, E(X_i / \mathbf{b}^t \mathbf{X}) \leq 1$, for $b_i = 0$) characterizing $\mathbf{b}^*(F)$ which maximizing $E \log \mathbf{b}^t \mathbf{X}$. By the strong law of large number, we observe that

$$(\mathbf{S}_n^*)^{1/n} = 2^{(1/n)\sum_{i=1}^{n} \log \mathbf{b}^{*t} \mathbf{X}_i} \to 2^W, \tag{5.10}$$

with probability one.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of random stock vectors with joint probability distribution $P^n$. If the true distribution $P^n$ were known, we may use the portfolio $\mathbf{b}_i^* = \mathbf{b}^*(P_{\mathbf{X}_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}})$, which maximizes the conditional

expected value of $\log \mathbf{b}^t \mathbf{X}_i$ given that $\mathbf{X}_1 = \mathbf{x}_1, \cdots, \mathbf{x}_{i-1}$. Suppose that instead of $\mathbf{b}^*$, we use portfolio $\hat{b}_i = b^*(\hat{P}_{\mathbf{X}_i \mid \mathbf{x}_1, \cdots, \mathbf{x}_{i-1}})$ which are optimal for a sequence of estimators of the true distribution $P^n$. Let $P_{\mathbf{X}_i \mid \mathbf{x}_1, \cdots, \mathbf{x}_{i-1}}$ and $\hat{P}_{\mathbf{X}_i \mid \mathbf{x}_1, \cdots, \mathbf{x}_{i-1}}$ be conditional distributions associated with $P^n$ and $\hat{P}^n$ respectively. Let $\hat{\mathbf{S}}_n = \prod_{i=1}^{n} \hat{\mathbf{b}}_i^t \mathbf{X}_i$ and $\mathbf{S}_n^* = \prod_{i=1}^{n} \mathbf{b}^{*t} \mathbf{X}_i$. Barron and Cover (1988) proved that

$$0 \leq E \log \frac{\mathbf{S}_n^*}{\hat{\mathbf{S}}_n} \leq D(P^n \| \hat{P}^n). \tag{5.11}$$

We gave conditions in Chapter 4 such that $ED(p \| \hat{p}_n) \leq cn^{-\gamma}$, for some $0 < \gamma < 1$. In which case we have following result.

**Proposition 5.2:** *If* $ED(p \| \hat{p}_k) \leq ck^{-\gamma}$, *then*

$$E\frac{1}{n} \log \frac{\mathbf{S}_n^*}{\hat{\mathbf{S}}_n} \leq \frac{c}{1-\gamma} n^{-\gamma}.$$

**Proof:** We assume $ED(p \| \hat{p}_k) \leq ck^{-\gamma}, k = 1, \ldots, n$. It yields that $E\frac{1}{n} \log \frac{\mathbf{S}_n^*}{\hat{\mathbf{S}}_n} \leq \frac{1}{n} \sum_{k=1}^{n} ED(p \| \hat{p}_k) \leq \frac{1}{n} \sum_{k=1}^{n} ck^{-\gamma}$. By calculations, we obtain

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^{n} ED(p \| \hat{p}_k) &\leq \frac{c}{n}(1 + \int_1^n x^{-\gamma} dx) \\ &= \frac{c}{n}\left(\frac{-\gamma}{1-\gamma} + \frac{n^{(1-\gamma)}}{1-\gamma}\right) \\ &\leq \frac{c}{1-\gamma} n^{-\gamma}. \quad \square \end{aligned}$$

**Corollary:** *If* $\int (D^r(\log p))^2 < \infty$ *and* $\gamma = 2r/(2r+d)$, *then using the maximum posterior likelihood estimator in exponential familes as in Corollary 4.4, we have*

$$E\frac{1}{n} \log \frac{\mathbf{S}_n^*}{\hat{\mathbf{S}}_n} \leq O(n^{-2r/(2r+d)}).$$

It follows that the actual wealth $\hat{\mathbf{S}}_n$ is close to the log-optimal wealth $\mathbf{S}_n^*$.

**Proposition 5.3:** *Let $\hat{p}_k$ be a sequence of estimators of the true density $p$ such that $ED(p \| \hat{p}_k) \leq ck^{-\gamma}, 0 < \gamma < 1$. Set $D_n = cn^{-\gamma}$. Let $\mathbf{S}_n^* = \prod_{i=1}^{n} \mathbf{b}^{*t}\mathbf{X}_i$ be the optimal wealth sequence. If $\hat{\mathbf{S}}_n = \prod_{i=1}^{n} \hat{\mathbf{b}}_i^t \mathbf{X}_i$ where $\hat{\mathbf{b}}_i = \mathbf{b}^*(\hat{p}_i)$, then*

$$\hat{\mathbf{S}}_n = \mathbf{S}_n^* 2^{-n\, O_{pr}(D_n)}$$

**Proof:** By Markov's inequality

$$P\{\frac{\hat{\mathbf{S}}_n}{\mathbf{S}_n^*} > 2^{nD_n}\} \leq 2^{-nD_n} E(\frac{\hat{\mathbf{S}}_n}{\mathbf{S}_n^*}) \leq 2^{-nD_n},$$

which tends to zero at rate $2^{-cn^{(1-\gamma)}}$. The inequality $E(\frac{\hat{\mathbf{S}}_n}{\mathbf{S}_n^*}) \leq 1$ from the Kuhn-Tuker conditions for the optimality of $\mathbf{b}^*$.

On the other side, for every $\mathrm{K} > 0$

$$P\{\frac{\mathbf{S}_n^*}{\hat{\mathbf{S}}_n} > 2^{nD_n \mathrm{K}}\} \leq \frac{1}{D_n \mathrm{K}}[\frac{1}{n} E \log \frac{\mathbf{S}_n^*}{\hat{\mathbf{S}}_n} + \frac{1}{n}]$$

$$\leq \frac{1}{\mathrm{K}}(\frac{1}{1-\gamma} + 1).$$

where the first inequality is as in Barron and Cover (1988, equation 38), and the second inequality is by the assumptions and Proposition 5.2. This bound is made arbitrarily small with sufficiently large K. Therefore, $\frac{1}{n} |\log \hat{\mathbf{S}}_n / \mathbf{S}_n^*| = O_{pr}(D_n)$ in probability as desired. □

If we only have two choices $X_1$ and $X_2$, then $W^* = \underset{b_1, b_2}{Max}\, E \log \mathbf{b}^t \mathbf{X}$. We

observe that

$$\underset{b_1,b_2}{Max}\ E\ \log \mathbf{b}^t\mathbf{X} = \underset{b_1,b_2}{Max}\ E\ \log \frac{b_1 X_1 + b_2 X_2}{X_1 + X_2} + E\ \log (X_1 + X_2)$$

$$= \underset{b}{Max}\ E\ \log \left(b\left(\frac{X_1}{X_1 + X_2}\right) + (1-b)\left(1-\frac{X_1}{X_1+X_2}\right)\right) + E\ \log (X_1 + X_2).$$

Thus we can simplify the two dimensional problem to one dimension. That means we just need to estimate the distribution function of $X_1/(X_1 + X_2)$.

## 5.3 Example

Nonparametric function estimation is a very visual area. In this section, we apply the method we presented in Chapter 2 to a real data and a simulated data. The computations were obtained using a program by Gayle Nygaard which performs Newton's algorithm to maximize the likelihood. We assume

$$p_\theta(x) = \exp\{\sum_{k=1}^{m} \theta_k \phi_{k,m}(x) - \psi_m(\theta)\}$$

where $\psi_m(\theta) = \log \int \exp\{\sum \theta_k \phi_{k,m}(x)\}\, p_0(x)\, dx$, $\theta \in \mathbf{R}^m$. The data $x_1, \ldots, x_n$ may be from a real data. We state the algorithm for finding the *MLE* $\hat{\theta}$ as follows:

step 0 The $p_0(x)$, dimension of the parameter vector, and the endpoints of the interval must be specified.

step 1 Values are provide for $\theta_i$, $i = 1, \ldots, m$, to initialize the search.

step 2 Read data.

step 3 Compute sufficient statistics

$$\overline{\phi}_{k,m} = \frac{1}{n}\sum_{i=1}^{n} \phi_{k,m}(X_i),\ k = 1,\ldots,m.$$

step 4 Compute $\psi_m(\theta)$

We call a subroutine *qng* from *CMLIB* to evaluate definite integrals of functions of one variable.

step 5 Do 5 k= 1,m

sum = sum + $\overline{\phi}_{k,m}$ * $\theta_k$

5 Continue

log-likelihood function = sum - log $\psi_m(\theta)$.

step 6 Use subroutine *qng* to compute the derivative of the likelihood function and the second partial derivative of the likelihood function with respect to the $i$ and $j$ coordinates.

Do 6 k= 1,m

dlike(k) = $\overline{\phi}_{k,m}$ - $\dfrac{d\psi_m}{\psi_m}$

sum2 = sum2 + dlike(k) * dlike(k)

6 Continue

step 7 Compute the new $\theta_i$ using Newton's method. We can use Gaussian elimination or Cholesky decomposition to solve the system of linear equations.

step 8 If ( sum2 < 0.00001 ) and ( $|\theta_{new,i} - \theta_i| < 0.00001$ )

then $\hat{\theta} = \theta_{new}$ and stop

else

$\theta_i = \theta_{new,i}$ and goto step 4.

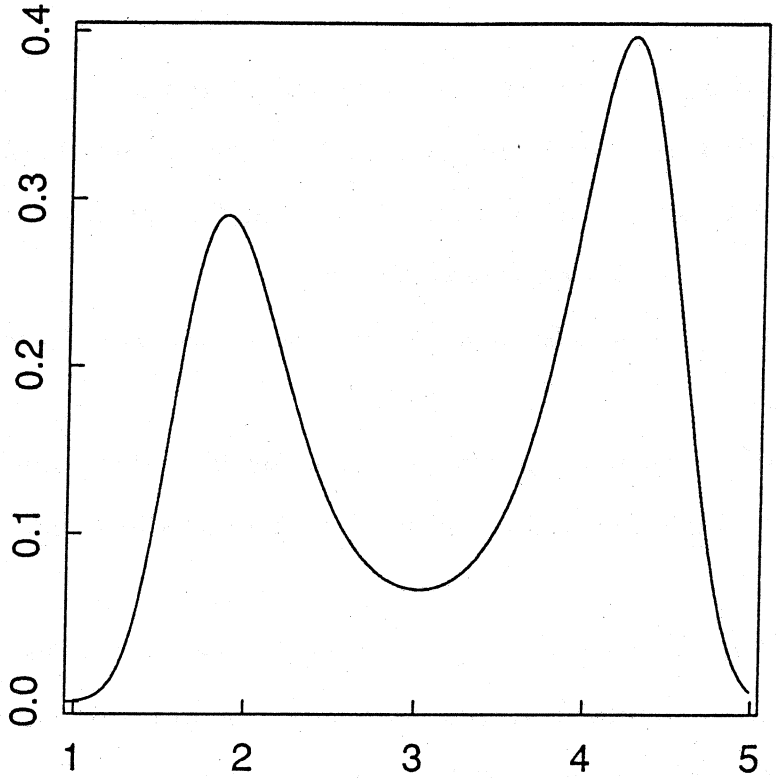
After we find the *MLE* $\hat{\theta}$, we know the density estimator $p_{\hat{\theta}}$. Then we can draw the picture of $p_{\hat{\theta}}$ and in simulated experiments we can compute the Kullback-Leibler number $D(p \| p_{\hat{\theta}})$ if we know the true density function.

**Example 1**:There are plenty of real data sets around from which an example could be constructed. 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, p.8). Using an exponential family with a polynomial of degree 4 on [1,5], we obtain the density estimate plotted in the figure of next page. The reference density $p_0$ is taken to be uniform on [1,5]. To avoid numerical overflow problems in the parameter search we found it advisable to scale the data to [-1,1] and to use the Legendre polynomial basis. The answer is then scaled back to the original interval.

We state the details as follows. Let $Y = (X-3)/2$ be the scaled data from original $X$. Let $p_\theta(y) = 1/(\psi_m(\theta))\sum_{k=1}^{m}\theta_k \phi_k(y)$, here $\phi_k, k = 1, \ldots, m$ are Legendre polynomials (i.e. $\phi_1(y) = y, \phi_2 = (3y^2 - 1)/2$, $\phi_3(y) = (5y^3 - 3y)/2$, $\phi_4(y) = (35y^4 - 30y^2 + 3)/8$). After using the algorithm we presented above, we find $\hat{\theta}_1 = 1.0288867$, $\hat{\theta}_2 = -1.2040582$, $\hat{\theta}_3 = 0.5387992, \hat{\theta}_4 = -2.8748367$ and $\psi_m(\hat{\theta}) = 3.228884$. We know $p_{\hat{\theta}}(x) = (p_{\hat{\theta}_y}((x-3)/2))/2$ by transformation.

The degree four 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, p.17). [For other estimates based on the same data see p.9,13,20 of Silverman]. A difference is that the kernel estimate has noticeable 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).




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

## References

Barron, A. R. and Cover, T. M. (1988). A bound on the financial value of information. *IEEE Trans. Inform. Theory,* vol.34, No.5, September 1988.

Barron, A. R. and Cover, T. M. (1989). Minimum Complexity Density Estimation, submitted to *IEEE Trans. Inform. Theory.*

Blahut, R. E. (1987). *Principles and Practice of Information Theory.* Reading: Addison-Wesley

Breiman, L. (1961). Optimal gambling systems for favorable games. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* **1,** 65-78.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* Monograph **9,** Institute of Mathematical Statistics, Hayward, California.

Canuto, C. and Quarteroni, A. (1982). Approximation Results for Orthogonal Polynomials in Sobolev Spaces. *Mathematics of Computation.* **38** 67-86.

Clarke, B. and Barron, A. R. (1988). Information theoretic asymptotics of Bayes method, submitted to *IEEE Trans. Inform. Theory.*

Cover, T. and Gluss, D. H. (1986). Empirical Bayes Stock Market Portfolios. *Adv. Appl. Math.* **7** , 170-181.

Cox, D. D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **18** 713-732.

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.

De Boor, C. and Fix, G. J. (1973). Spline approximation by quasiinterpolants. *J. Approximation Theory* **8** 19-45.

Devroye, L. and Gyorfi, L. (1985). *Nonparametric density estimation. The L1 View.* Wiley, New York.

Efroimovich, S. Y. and Pinsker, M. S. (1983). Estimation of square-integrable probability density of a random variable. *Probl. Inform. Transmission* **18** 175-189. (Translated from *Probl. Peredachi Inform.* **18** 19-38, 1982).

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers,* **23,** 881-889.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal American Statistical Association,* **76,** 817-823.

Friedman, J. H., Stuetzle, W. and Schroeder, A. (1984). Projection Pursuit density estimation. *Journal American Statistical Association,* **79,** 599-608.

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

Haugen, R. A. (1986). *Modern investment theoryi.* Prentice Hall.

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

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statistics,* **13,** 435-525.

Jackson, D. (1930). *The Theory of Approximation.* American Mathematical Society, New York.

Kelly, J. (1956). A new interpretation of information rate. *Bell System Tech. J.* 917-926.

Kraft, C (1955). Some conditions for consistency and uniform consistency of statistical procedures, *University of California Publications in Statistics,* Vol.2, pp.125-141.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.,* **22,** 79-86.

Lions J-L (1987). *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer-Verlag.

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 non-parametric estimates for the density function. *Theory Probab. Appl.* **19** 133-141.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*,**33**, 1065-1076

Pinsker, M. S. (1964). Information and Information Stability of Random Variables. *Translated by A. Feinstein, Holden-Day, San Francisco.*

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

Rosenblatt, M. (1956). Remarks on some nonaprametric estimates of a density function. *Ann. Math. Statist.*,**27**, 832-837.

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

Sterbuchner, H. (1980). On nonparametric multivariate density estimation. *Rev. Roum. Math. Pures Et Appl,* Tome XXV, No 1, 111-118

Stone, C. J. and Koo, C.-Y. (1986). Logspline density estimation. In *Contemporary Mathematics* **59** 1-15. American Mathematics Society, Providence, RI.

Van Campenhout, J. M. and Cover, T. M. (1981). Maximum entropy and conditional probability. *IEEE Trans. Inform. Theory.* **27** 483-489.

Wahba, G. (1975). Optimal convergence properties of variable knot, kernel and

orthogonal series methods for density estimation. *Ann. Statist.* **3** 15-29.

Zabell, S. L. (1980). Rates of convergence for conditional expectations. *Ann. Probab.* **8** 928-941.

# Vita

Chyong-Hwa Sheu was born on March 20, 1956 in Ilan, Taiwan. He moved a lot during the past thirty years. He studied in many schools before he attended National Kaohsiung Teacher's College in 1974, where he received his Bachelor of Science in 1979. Then he taught mathematics in a high school for two years. He was chosen to be an exchange student to Eastern Illinois University in 1982. He received his Master's degree in mathematics there. He entered the University of Illinois at Urbana-Champaign in August 1984. During all periods at the University of Illinois, he was either a teaching assistant or a research assistant in the Department of Statistics. He met so many great teachers and friends at the University of Illinois. For all these, he is filled with thanksgivings. He will never forget the colorful life he had in Urbana-Champaign.