Density Estimation Through Convex Combinations of Densities; Approximation and Estimation Bounds

Assaf J. Zeevi* and Ronny Meir

Faculty of Electrical Engineering,
Technion, Haifa 32000, Israel

Abstract

We consider the problem of estimating a density function from a sequence of independent and identically distributed observations $\mathbf{x}_i$ taking value in R^d. The estimation procedure constructs a convex mixture of 'basis' densities and estimates the parameters using the maximum likelihood method. Viewing the error as a combination of two terms, the approximation error measuring the adequacy of the model, and the estimation error resulting from the finiteness of the sample size, we derive upper bounds to the expected total error. These results then allow us to derive explicit expressions relating the sample complexity and model complexity

1 Introduction

The problem of density estimation is one of great importance in many domains of engineering and statistics, playing an especially significant role in pattern recognition and regression. There have traditionally been two principal approaches to dealing with density estimation, namely the parametric view which makes stringent assumptions about the density, and the nonparametric approach which is essentially distribution free. In recent years, a new approach to density estimation, often referred to as the method of sieves [10], has emerged. In this latter approach, one considers a family of parametric models, where each member of the family is assigned a 'complexity' index in addition to the parameters. In the process of estimating the density one usually sets out with a simple model (low complexity index) slowly increasing the complexity of the model as the need may be. This general strategy seems to exploit the benefits of both the

*Also affiliated with the Faculty of Industrial Engineering, Technion

parametric as well as the nonparametric approaches, namely fast convergence rates and universal approximation ability, while not suffering from the drawbacks of the other methods. As has been demonstrated by White [27], the problem of learning in feed-forward neural networks can be viewed as a specific implementation of the method of sieves. Barron [3], has recently studied a density estimator based on sequennces of exponential families, and established convergence rates, in the Kulback - Leibler measure. In a related context, very encouraging results have been obtained recently by Barron concerning the convergence rates for function approximation [5] and estimation [6] using neural networks.

The purpose of this paper is to apply some of Barron's results [5] to the problem of density estimation. We also utilize the general results of White [26], concerning estimation in a misspecified framework, deriving upper bounds on the approximation and estimation error terms. However, rather than representing the density as an arbitrary combination of non-linearly parameterized functions, as in the function approximation framework, we demand that the representation be given by a convex combination of density functions. While this requirement seems rather stringent, it will turn out that a very broad class of densities can be closely approximated by this model. The main result is an upper bound on the total error between a target density and a finite mixture model estimator. This construction actually permits an interpretation of a broad class of densities as mixture models. Furthermore, as long as the 'basis' densities belong to a broad class of densities (the so-called exponential family) a very efficient learning algorithm, known as the EM algorithm, exists [21].

¿From the point of view of density estimation, there are two basic questions of interest. First, the approximation problem refers to the question of whether the representation is sufficiently powerful to parsimoniously represent a broad class of density functions. Assuming the answer to this question is affirmative (as we demonstrate below), the question arises as to whether one can find an efficient estimation scheme, which allows one to compute the optimal values of the parameters from a finite set of examples. As we show, the answer to this question is also affirmative. From the approximation point of view, our results can be viewed as an extension of a well known result which we have traced to Fergusson [9], stating that any density function may be approximated to arbitrary accuracy by a convex combination of normal densities. Normal, or Gaussian, densities appear also, in the approximation literature in the more general form of Radial Basis Functions (RBF). This class has been studied extensively in the approximation literature (see [19] for instance), and has found applications also in neural network models in the form of RBF networks [17]. In the framework we present the approximating class of densities is not necessarily constituted of the Gaussian type, rather we present the general functional form of which RBF is a specific admissable choice..

Another model ,introduced recently in by Jacobs *et al.* . [11], termed the mixture of experts model (MEM), is motivated by the concept of mixture models. It is demonstrated (see for instance [12]) that an efficient learning algorithm (EM) is applicable in this case and results in superior convergence rates and robustness [14]. The results we obtain herein, may be applied in the case of the MEM to relate model complexity and sample complexity, and extend the estimation results to misspecified scanrios (i.e., when the data generating probability law is not a subset of the models used to estimate it).

It should be noted that utilizing the recent results concerning function approximation [5], it is possible to achieve a representation for density functions, by transforming the outputs of a neural network into exponential form and normalizing the density appropriately. However, we believe that representing a general density as a convex combination of densities affords much insight as well as giving rise to efficient learning algorithms which are not available in the case of neural network models.

The remainder of the paper is organized as follows. We present an exact definition of the problem in section 2, relating it to the general issue of function approximation. In section 3 we then present some preliminary results which are needed in deriving the main theorems. Section 4 of the paper then proceeds to present the theorems concerning the approximation and estimation error for the convex combination of densities. A specific estimation scheme ('learning algorithm') is presented in section 5, and compared with standard approaches used in the neural network literature. A summary of our results, together with current research directions, is presented in section 6. Some of the technical details are relegated to the appendix, for the sake of coherence of presentation.

2 Definitions, Notation and Statement of the Problem

The problem of density estimation can be decomposed into two basic issues. The first question is related to the quality of approximation, namely how well can a class of functions approximate an unknown probability density. Assuming the approximation issue has been addressed, one still has to deal with the question of whether an algorithm exists to find the best approximation, and to characterize the dependence of the algorithm on the size of the data set. The latter problem is usually referred to as the problem of estimation.

The problem of density approximation by convex combinations can be phrased as follows: we wish to approximate a class of density functions, by a convex combination of 'basis' densities. Let us start clarifying this objective by introducing the following function classes:

$$\mathcal{F}_c = \left\{ f \mid f \in C_c(\mathbb{R}^d),\ f \geq 0,\ \int f = 1 \right\} \tag{1}$$

which is the class of all continuous densities with compact support in $\mathbb{R}^d$, denoted: In general we can consider a target density to be any unknown, continuous, density, restricted to some compact domain, where the approximation results are valid. We define the class of admissable *target densities* as

$$\mathcal{F}_{c,\eta} = \{ f \in \mathcal{F}_c \mid \forall f \ \ \exists \eta,\ s.t.\ f \geq \eta > 0 \} \,. \tag{2}$$

This class is composed of all compactly supported continuous densities, bounded below by some positive constant which we generically denote as η. While this requirement may seem somewhat unnatural at this point, it is needed in the precise statement of the theorems stated in section 4. Since we will be utilizing the KL divergence (to be defined) as a discrepency measure, it is quite natural to consider densities that are bounded from below. Unless this condition is satisfied, densities may be arbitrarily close in the L_1 metric, while the KL divergence is arbitrarily large (see for example Wyner and Ziv [29] for a discussion in the context of discrete probability measures). Having defined the above classes, we note in passing that the following relation holds $\mathcal{F}_{c,\eta} \subset \mathcal{F}_c$.

With the class of target densities at hand, we proceed by defining the class of 'basis' densities, which will serve as the approximation building blocks. These 'basis' densities are then used to build a nested family of convex models. We begin by denoting the class of continuous densities by

$$\Phi = \left\{ \phi \mid \phi \in C(\mathbb{R}^d),\ \phi > 0,\ \int \phi = 1 \right\}. \tag{3}$$

Recalling our restricted target class $\mathcal{F}_{c,\eta}$ and considering the characteristics of convex combinations, we define

$$\Phi_\eta = \{\phi \in \Phi \mid \phi \geq \eta > 0\}. \tag{4}$$

Obviously, from the design standpoint, given some apriori knowledge concerning $\mathcal{F}_{c,\eta}$ characterizing the target density's lower bound, the densities $\phi \in \Phi$ may be chosen accordingly. This generic class of densities will now be endowed with a parametric form,

$$\Phi_{\eta,\tau} = \left\{ \phi_\sigma \in \Phi_\eta \mid \phi_\sigma \stackrel{\triangle}{=} \sigma^{-d}\phi\left(\frac{\cdot - \boldsymbol{\mu}}{\sigma}\right),\ \boldsymbol{\mu} \in \mathbb{R}^d,\ \sigma \in R,\ s.t.\ \sigma \geq \tau > 0 \right\}. \tag{5}$$

The motivation for this parameterization will be made below, when we introduce the approximating class of densities, and discussed further in section 3. Notice that ϕ_σ is merely $\phi(\cdot/\sigma)$ normalized in the d-dimensional space. This form of parameterization formally classifies the 'basis' densities as members of the scale-location family of densities. We make the parameterization of ϕ implicit by defining the 'basis' densities as $\{\phi_\sigma(\cdot;\boldsymbol{\theta})\}$ where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma)$. Although we do not specify the exact functional form of these densities, we consider some possible choices of multidimensional 'basis' densities. The following two candidates are adapted from the common kernel functions, used in multidimensional nonparametric regression and density estimation (see for example [23]).

- **Product kernel -** Each ϕ_σ can be written as a product of d univariate kernels. In this case, the structure of each kernel usually depends on a separate *smoothing factor* in each dimension, i.e $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_d)$. The univariate 'basis' density may be chosen from a list of common kernel functions such as: the triweight, epanechnikov, normal etc.

- **Radial Basis Functions -** The 'basis' densities are of the form $\phi_\sigma(\cdot/\sigma) \equiv \phi_\sigma(\|\cdot\|/\sigma)$, that is a Euclidean norm is used as the metric. In this formulation only one *smoothing parameter* is used in each 'basis' density. This requires a pre-scaling or pre-whitening of the data, since the 'basis' density function scales equally in all directions. The formulation can, of course, be extended to handle a vector of smoothing parameters (like the product kernel case). In any such case the vector of parameters remains of dimension $O(nd)$ where n is the complexity index of the model, and d is the dimension of the data.

The form of the 'basis' density can be chosen from the list of common kernel functions, all of which are radially symmetric and unimodal. Such kernels may be the multivariate Gaussian kernel or the multivariate epanechnikov kernel, endowed with the Euclidean distance norm.

As noted before, the latter functional class is of particular interest in function approximation problems in general, and an enormous literature exists, ranging from approximation theory results (see [19] and [16] for some results in the context of neural netwroks), to applications. The original proof establishing the universal approximation capability of convex combinations of Gaussian densities (traced to [9]) also falls into this category.

We note that $\Phi_{\eta,\tau} \subset \Phi_\eta \subset \Phi$ and $\Phi_{\eta,\tau} \subset \mathcal{F}_{c,\eta}$ (considering a restriction to a compact domain). As stated previously, our objective is to approximate the target density by convex combinations of the predefined, 'basis' densities. We now define the *approximation* class

$$\mathcal{G}_n = \left\{ f_n^\theta \mid f_n^\theta(\cdot) = \sum_{i=1}^{n} \alpha_i \phi_\sigma(\cdot; \boldsymbol{\theta}_i),\ \phi_\sigma \in \Phi_{\eta,\tau},\ \alpha_i > 0,\ \sum_{i=1}^{n} \alpha_i = 1 \right\} \tag{6}$$

so that $\mathcal{G}_n$ is the class of convex combinations of parameterized densities consisting of n components. Note that $\mathcal{G}_n$ constitutes a nested family so that

$$\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots \subset \mathcal{G}_n \subset \ldots \subset \mathcal{G} \tag{7}$$

where $\mathcal{G} = \cup \mathcal{G}_n$. We denote the full set of parameters by $\boldsymbol{\theta}$, namely $\boldsymbol{\theta} = \{\{\alpha_i\}, \{\boldsymbol{\theta}_i\}\}$. Note that the number of parameters in $\boldsymbol{\theta}$ is proportional to n, which will henceforth be referred to as the *complexity index* or model complexity term. This formulation is quite similar in content to that of finite mixture models (see for example Titternigton [24]), though we take a different approach in defining the classes of basis densities. Moreover, we seek a relationship between the sample size and the complexity of the model, through the upper bounds on the expected total error.

According to the approximation objective, we wish to find values $\boldsymbol{\theta}^*$ such that for any $\varepsilon > 0$

$$d(f, f^*) \leq \varepsilon \tag{8}$$

where f^* is the value of f_n^θ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Here $d(f,g)$ represents some generic distance function between densities f and g, whose exact form will be specified in the next section. As discussed above, establishing the existence of a good approximating density f^* is only the first step. One still needs to consider an effective procedure, whereby the optimal function can be obtained.

The estimation objective may be stated as follows: Given a sample (data) set $\mathcal{D}_N\{\mathbf{x}_i\}_{i=1}^N$ drawn from the underlying *target* density $f \in \mathcal{F}_{c,\eta}$, we estimate a density $\hat{f}_{n,N} \in \mathcal{G}_n$ by means of maximum likelihood (i.e. maximizing the empirical likelihood). The following step will be to assess the performance of this estimator. We shall carry this out by defining an appropriate metric that will subsequently be used in establishing upper bounds on the total error. In this work we utilize the Hellinger distance as a measure of divergence between the target density and the estimator.

In summary then, the basic issue we address in this work is related to the relationship between the approximation and estimation errors and (i) the dimension of the data, d, (ii) the sample size, N, and (iii) the complexity of the model class parameterized by n.

3 Preliminaries

We devote this section to some technical definitions and lemmas which will be utilized in the following section, where the main results are stated and derived. In order to measure and discuss the accuracy of the estimation (and approximation), we must define an appropriate distance measure, $d(f,g)$, between densities f and g. A commonly used measure of discrepancy between densities is the so-called Kullback-Leibler (KL) divergence (sometimes referred to as relative entropy), given by

$$D(f||g) \triangleq \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})}\, d\mathbf{x}. \tag{9}$$

As is obvious from the definition, the KL divergence is not a true distance function since it is not symmetric nor does it obey the triangle inequality. To circumvent this problem one often resorts to an alternative definition of distance, namely the squared Hellinger distance

$$d_{\mathrm{H}}^2(f,g) \triangleq \int \left(\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})}\right)^2 d\mathbf{x}, \tag{10}$$

which can be shown to be a true metric (obeying the triangle inequality) and is particularly useful for problems of density estimation (see Le Cam [15]). Finally, for the sake of completeness we define the L_p distance

$$d_p(f,g) \triangleq \left(\int |f(\mathbf{x}) - g(\mathbf{x})|^p d\mathbf{x}\right)^{1/p} \tag{11}$$

We quote below without proof three lemmas relating the various distances. These inequalities will be used in section 4 in the derivation of the estimation error.

Lemma 3.1 (Devroye & Györfy, 1985) *The Hellinger distance is related to the L_1 distance as follows:*

$$\left(\frac{1}{2} d_1(f,g)\right)^2 \le d_{\mathrm{H}}^2(f,g) \le d_1(f,g). \tag{12}$$

Lemma 3.2 *For all densities f and g, the squared Hellinger distance is bounded by the KL divergence as follows*

$$d_{\mathrm{H}}^2(f,g) \le D(f||g), \tag{13}$$

Lemma 3.3 *For any two strictly positive densities f and g, such that $g, f \ge 1/\gamma^2$, the KL divergence is bounded as follows*

$$D(f||g) \le \gamma^2 d_2^2(f,g) \tag{14}$$

Proof: By Jensen's inequality

$$D(f||g) = E_f \log \frac{f}{g} \le \log E_f \frac{f}{g} = \log \int \frac{f^2}{g}$$

and upper bound on the logarithm

$$\log \int \frac{f^2}{g} \le \int \frac{f^2}{g} - 1 = \int \frac{(f-g)^2}{g} \le \gamma^2 d_2^2(f,g) \qquad \square$$

A crucial step in establishing our results is given by the following theorem, which allows one to represent an $L_p(\mathbb{R}^d)$ function to arbitrary accuracy by a convolution with a function $\phi \in L_1(\mathbb{R}^d)$. Formally we have (see Petersen [18]):[1]

Lemma 3.4 *(*Petersen, 1983*) Let $1 \le p < \infty$ and let $\phi \in L_1(\mathbb{R}^d)$, $\int \phi = 1$. Letting $\phi_\sigma(\mathbf{x}) = \sigma^{-d}\phi(\mathbf{x}/\sigma)$,then for any $f \in L_p(\mathbb{R}^d)$ we have $\phi_\sigma * f \to f$ in $L_p(\mathbb{R}^d)$ as $\sigma \to 0$, where*

$$(\phi_\sigma * f)(\mathbf{x}) \triangleq \int \phi_\sigma(\mathbf{x}-\mathbf{y}) f(\mathbf{y}) d\mathbf{y} \tag{15}$$

This statement establishes that Φ is dense in $L_p(\mathbb{R}^d)$. It is immediately obvious from Lemma 3.4, that the only requirement of the function ϕ is that it integrate to unity. This condition immediately raises the possibility of considering ϕ to be a density function, which imposes the further condition (allowed by the lemma) that $\phi \ge 0$. Although Lemma 3.4 refers to the general space $L_p(\mathbb{R}^d)$, the result obviously holds also for $C_c(\mathbb{R}^d)$ (for discussion see Adams [1], 1975, pp. 28-31). At this point the motivation for the classes of 'basis' densities is quite obvious: by a correct choice of ϕ_σ we can approximate any function in $L_p(\mathbb{R}^d)$ to any degree of accuracy, in the integral representation. This continuous representation will later be seen to be merely the limit of a convex combination of infinitely many 'basis' densities. The lemma states that for all $\varepsilon > 0$ there exists a positive constant $\tau > 0$ and some $\sigma \ge \tau$ such that

$$\|f - \bar{f}\|_p < \varepsilon \tag{16}$$

for $1 \le p < \infty$, where $\bar{f} \equiv f * \phi_\sigma$. Since f is a density function, and both f and ϕ_σ are continuous functions, it follows that the integrand of the convolution (15) is continuous a.e. and thus from the Riemann theory of integration we have:

[1]One of the referees has pointed out that this lemma predates the refernece we quote here.

Corollary 3.1 *The function $\bar{f}$ belongs to the closure of the convex hull of $\Phi_{\eta,\tau}$, namely $\bar{f} \in \bar{co}\Phi_{\eta,\tau}$.*

At this point we have shown that any density function can be approximated to arbitrary accuracy by an infinitely countable convex combination of densities $\phi_\sigma(\mathbf{x};\boldsymbol{\theta}) = \sigma^{-d}\phi\left((\mathbf{x}-\boldsymbol{\mu})/\sigma\right)$, comprising $\bar{f}$. The question arises, however, as to how many terms are needed in the convex combination in order to approximate $\bar{f}$ to some arbitrary $\varepsilon > 0$. From Corollary 3.1 we infer that $\bar{f}$ belongs to the closure of the convex hull of the set of functions $\Phi_{\eta,\tau}$, thus we can immediately make use of the following remarkable result attributed to Maurey and proved for example in Barron [5]. Denoting by $\|f\|_2$ the L_2 norm of the function f, we have:

Lemma 3.5 (Maurey, Barron 1993) *If $\bar{f}$ is in the closure of the convex hull of a set G in Hilbert space, with $\|g\|_2 \leq b$ for each $g \in G$, then for every $n \geq 1$, and every $c > (b^2 - \|\bar{f}\|_2^2)^{1/2}$, there is a function f_n^0 in the convex hull of n points in G such that*

$$d_2^2(\bar{f}, f_n^0) \leq \frac{c}{n} \ . \tag{17}$$

Proof Sketch: The main idea of the proof follows from a random coding argument. Think of the functions as elements in a probability space, and the function in the closure of the convex hull as the mean (w.r.t. a discrete probability measure). By application of Chebychev's inequality, it is seen that there is a positive probability that any function in the convex hull and the average of n functions (independently drawn) are $1/\sqrt{n}$ far apart.

Let us now consider the results of Lemma 3.5 in the context of the approximation classes defined in the previous section. Recall the class $\mathcal{G}_n$ which was defined as the set of convex combinations of n points in the convex hull of $\Phi_{\eta,\tau}$. By Corollary 3.1 we have $\bar{f} \in \bar{co}\Phi_{\eta,\tau}$, thus restating the result of Lemma 3.5 we have that for every $\bar{f}$ there exists an approximation $f_n^0 \in \mathcal{G}_n$ such that

$$d_2^2(\bar{f}, f_n^0) \leq \frac{c}{n}. \tag{18}$$

By Lemma 3.4 we have, for some fixed accuracy measure $\varepsilon > 0$ and target density $f \in \mathcal{F}_{c,\eta}$ there exists an $\bar{f}$ so that

$$d_2^2(f, \bar{f}) \leq \varepsilon \tag{19}$$

where $\bar{f}$ is the convolution of f with the kernel function ϕ_σ. Combining (18) and (19) we have , by the triangle inequality

Corollary 3.2 *For any $f \in \mathcal{F}_{c,\eta}$ and some fixed accuracy measure $\varepsilon > 0$, there exists a convex combination f_n^0, in the class $\mathcal{G}_n$, such that*

$$d_2^2(f, f_n^0) \leq \varepsilon + \frac{c}{n}$$

.

This result establishes the relation between the approximation error and the number of terms in the convex combination model. In the following section we shall make use of this result in the context of the maximum-likelihood estimator, $\hat{f}_{n,N}$. The existence of an $f_n^0 \in \mathcal{G}_n$ for every $f \in \mathcal{F}_{c,\eta}$ establishes, in essence, the approximation bound for the maximum-likelihood estimator.

4 Main Results

As we have shown in the previous section, given any $\varepsilon > 0$ one can construct a convex combination of densities, $f^\theta \in \mathcal{G}_n$, in such a way that the squared L_2 distance between an arbitrary density $f \in \mathcal{F}_{c,\eta}$ and the model is smaller than $\varepsilon + c/n$. We consider now the problem of estimating a density function from a sequence of d-dimensional samples, $\{\mathbf{x}_i\}$, $i = 1, 2, \ldots, N$, which will be assumed throughout to be independent and identically distributed according to $f(\mathbf{x})$. Following the definition of the approximation class in eq. (6), we let n denote the number of components in the convex combination. The total number of parameters will be denoted by m, which in the problem studied here is equal to $n(d+2)$.

In the remainder of this section we consider the problem of estimating the parameters of the density through a specific estimation scheme, namely maximum likelihood. Defining the log-likelihood function

$$l(\mathbf{x}^N; \boldsymbol{\theta}) = \frac{1}{N} \sum_{k=1}^{N} \log f_n^\theta(\mathbf{x}_k) \tag{20}$$

where $\mathbf{x}^N = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and $f_n^\theta(\mathbf{x}) = \sum_{i=1}^n \alpha_i \phi_\sigma(\mathbf{x}; \boldsymbol{\theta}_i)$, the method of maximum likelihood attempts to maximize l in order to find the optimal $\boldsymbol{\theta}$. Denoting the value of the maximum likelihood estimate by $\hat{\boldsymbol{\theta}}_{n,N}$ we have (by definition)

$$\hat{\boldsymbol{\theta}}_{n,N} = \arg\max_\theta l(\mathbf{x}^N; \boldsymbol{\theta}). \tag{21}$$

We denote the value of f_n^θ evaluated at the maximum likelihood estimate by $\hat{f}_{n,N}$. Now, for a fixed value of n, the finite mixture model, f_n^θ, may not be sufficient to approximate the density f, to the required accuracy. Thus, the model for finite n falls into the so called class of *misspecified* models [25] and the procedure of maximizing l should more properly be referred to as *quasi maximum likelihood* estimation. Thus, $\hat{\boldsymbol{\theta}}_{n,N}$ is the quasi maximum likelihood estimator. Since the data are assumed to be i.i.d, it is clear from the strong law of large numbers (given that the $D(f \| f_n^\theta) < \infty$) that

$$\frac{1}{N} l(\mathbf{x}^N; \boldsymbol{\theta}) \to \mathrm{E} \log f_n^\theta(\mathbf{x}) \qquad \text{(almost surely as } N \to \infty\text{)}, \tag{22}$$

where the expectation is taken with respect to the true (but unknown) density, $f(\mathbf{x})$, generating the examples. From the trivial equality

$$\mathrm{E} \log f_n^\theta(\mathbf{x}) = -D(f \| f_n^\theta) + \mathrm{E} \log f(\mathbf{x})$$

we see that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{n,N}$ is asymptotically given by $\boldsymbol{\theta}_n^*$, where

$$\boldsymbol{\theta}_n^* = \arg\min_{\theta} D(f \| f_n^{\theta}). \tag{23}$$

We assume for simplicity that $\boldsymbol{\theta}_n^*$ is unique, and denote the value of f_n^{θ} evaluated at $\boldsymbol{\theta}_n^*$ by f_n^* (for a detailed discussion see White [25] and [26]). In order not to encumber the text, we have collected the various technical assumptions needed in Appendix **??**..

Now, the quantity of interest in density estimation is the distance between the true density, f, and the density obtained from a finite sample of size N. Using the previous notation and the triangle inequality for metric $d(\cdot,\cdot)$ we have

$$d(f, \hat{f}_{n,N}) \leq d(f, f_n^*) + d(f_n^*, \hat{f}_{n,N}) \tag{24}$$

This inequality stands at the heart of the derivation which follows. We will show that the first term, namely the *approximation error*, is small. This follows from Lemma 3.5 as well as the inequalities presented in section 3. In order to evaluate the second term, the *estimation error*, we make use of the results of White [25] concerning the asymptotic distribution of the quasi maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{n,N}$. The splitting of the error into two terms in (24), is closely related to the expression of the mean squared error in regression as the sum of the bias (related to the approximation error) and the variance (akin to the estimation error).

A stated in the previous section, Corollary 3.2 provides us with an existence proof, in the sense that there exists a parameter value $\boldsymbol{\theta}^0$ such that the approximation error of the n-term convex combination model (6) - belonging to $\mathcal{G}_n$ - is smaller than $\varepsilon + c'/n$. Since we are dealing here with a specific estimation scheme, namely maximum likelihood, which asymptotically approaches a particular parameter value $\boldsymbol{\theta}_n^*$, the question we ask is whether the parameter $\boldsymbol{\theta}_n^*$, obtained through the maximum likelihood procedure, also gives rise to an approximation error of the same order as that of $\boldsymbol{\theta}^0$. The answer to this question is affirmative, as we demonstrate in the next lemma.

Lemma 4.1 (Approximation error) *Given Assumption B.7, for any target density $f \in \mathcal{F}_{c,\eta}$, the Hellinger distance between f and the density f_n^*, minimizing the Kullback-Leibler divergence, is bounded as follows:*

$$d_{\mathrm{H}}^2(f, f_n^*) \leq \varepsilon' + \frac{C_{\mathcal{F},\Phi}}{n} \tag{25}$$

where $C_{\mathcal{F},\Phi}$ is a constant depending on the class of target densities $\mathcal{F}_{c,\eta}$ and the family of basis densities $\Phi_{\eta,\tau}$, and ε' is some predetermined precision constant.

Proof: From Lemma 3.2 we have that

$$d_{\mathrm{H}}^2(f, f_n^*) \leq D(f \| f_n^*). \tag{26}$$

Denoting by f_n^0 the value of f_n^θ evaluated at the point $\boldsymbol{\theta}^0$, and obeying $d_2^2(f, f_n^{\theta^0}) \leq \varepsilon + c/n$ (for some $c > 0$, known to exist from Corollary 3.2), we have

$$D(f||f_n^*) \overset{(a)}{\leq} D(f||f_n^0) \overset{(b)}{\leq} \gamma^2 d_2^2(f, f_n^0) \overset{(b)}{\leq} \gamma^2 \varepsilon + \gamma^2 \frac{c}{n}, \tag{27}$$

where $\gamma^2 = 1/\eta$ (η is the lower bound on the target density, over the compact domain X, and the bound is valid by Lemma 3.3 and Assumption B.7). The inequality (a) follows from the fact that f_n^* minimizes the KL divergence between f and f_n^θ. The second inequality (b) follows from (26) and (c) follows from Corollary 3.2. Combining (26) and (27) we obtain the desired result

$$d_{\mathrm{H}}^2(f, f_n^*) \leq \varepsilon/\eta + \frac{c/\eta}{n}, \tag{28}$$

with $\varepsilon' \equiv \varepsilon/\eta$ and $C_{\mathcal{F},\Phi} \equiv c/\eta$ $\square$

We stress that the main point of theorem 4.1 is the following. While Corollary 3.2 assures the existence of a parameter value $\boldsymbol{\theta}^0$ and a corresponding function f_n^0 which lies within a distance of $\varepsilon + O(1/n)$ from f, it is not clear apriori that f_n^*, evaluated at the quasi maximum likelihood estimate, $\boldsymbol{\theta}_n^*$, is also within the same distance from f. Theorem 4.1 establishes this fact.

Up to now we have been concerned with the first part of the inequality (24). In order to bound the *estimation error* resulting from the maximum likelihood method, we need to consider now the second term in the same equation. To do so we make use of the following lemma, due to White [25], which characterizes the asymptotic distribution of the estimator $\hat{\boldsymbol{\theta}}_{n,N}$ obtained through the quasi maximum likelihood procedure. The specific technical assumptions needed for the lemma are detailed in Appendix **??**. A quantity of interest, which will be used in the lemma is

$$C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) A(\boldsymbol{\theta})^{-1}, \tag{29}$$

where

$$\begin{aligned} A(\boldsymbol{\theta}) &= \mathrm{E}\left[\boldsymbol{\nabla}\boldsymbol{\nabla}^T \log f_n^\theta(\mathbf{x})\right], \\ B(\boldsymbol{\theta}) &= \mathrm{E}\left[\left(\boldsymbol{\nabla} \log f_n^\theta(\mathbf{x})\right)\left(\boldsymbol{\nabla} \log f_n^\theta(\mathbf{x})\right)^T\right], \end{aligned} \tag{30}$$

and the expectations are with respect to the true density f. The gradient operator $\boldsymbol{\nabla}$ represents differentiation with respect to $\boldsymbol{\theta}$.

Lemma 4.2 (White 1982) *Given assumptions B.1 - B.6,*

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{n,N} - \boldsymbol{\theta}_n^*\right) \sim AN\left(0, C^*\right), \tag{31}$$

where $AN(0, C^*)$ should be interpreted as 'asymptotically normal with mean zero and covariance matrix $C^* \equiv C(\boldsymbol{\theta}_n^*)$'.

Finally, we will make use of the Fisher information matrix defined with respect to the density f_n^*, which we shall refer to as the *pseudo-information matrix*, given by

$$I^* = \mathrm{E}_*[\boldsymbol{\nabla} \log f_n^*(\mathbf{x}) \boldsymbol{\nabla} \log f_n^*(\mathbf{x})^T] \ , \tag{32}$$

The expectation in (32) is taken with respect to f_n^*, the density f_n^θ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

With Lemma 4.2 in hand we are ready now to derive the main result of this paper, concerning the expected estimation error for the maximum likelihood based estimator, in the context of the convex combination model. Denoting expectations over the data (according to the true density f) by $\mathrm{E}_{\mathcal{D}}[\cdot]$, we have:

Theorem 4.1 (Expected error bound) *For sample size N sufficiently large, and given assumptions B.1 - B.7, the expected estimation error, $E_{\mathcal{D}}\left[d_{\mathrm{H}}^2(f, \hat{f}_{n,N})\right]$ to some predetermined accuracy ε, obtained from the quasi maximum likelihood estimator $\hat{f}_{n,N}$, is bounded as follows:*

$$E_{\mathcal{D}}\left[d_{\mathrm{H}}^2(f, \hat{f}_{n,N})\right] \leq \varepsilon + O\left(\frac{C_{\mathcal{F},\Phi}}{n}\right) + O\left(\frac{m^*}{N}\right) \tag{33}$$

where $m^* = \mathrm{Tr}(C^* I^*)$ with C^* and I^* given in eq. (31) and (32) respectively.

Proof: See Appendix A.

At this point we make several comments, regarding the result of the theorem, and draw attention to some points which have been temporarily overlooked in the process of derivation.

Remark 4.1 The three terms on the right hand side of eq. (33) may be interpreted as follows. The accuracy measure ε results from the lower bound τ on the parameter σ in ϕ_σ, which restricts the approximation power of the family $\Phi_{\eta,\tau}$. The second term is a direct result of Lemma 3.5 concerning the degree of approximation obtained by the class Φ. These two terms together constitute the approximation error. Finally, the third term results from the estimation error of the maximum-likelihood estimator.

Remark 4.2 For n sufficiently large, the matrix C^* converges to the inverse of the 'true density' (i.e., the approximation term becomes negligble) Fisher information matrix, which we shall denote by $I^{-1}(\boldsymbol{\theta})$, and the pseudo-information matrix, I^*, converges to the Fisher information $I(\boldsymbol{\theta})$. This argument follows immediately from Lemma 4.1, which ensures the convergence of the misspecified model to the 'true', underlying density (to the ε specified accuracy). Therefore their product will be of order m, where m denotes the dimension of the parameter vector $m \equiv n(d+2)$. The bound on the estimation error will therefore be given by

$$\mathrm{E}_{\mathcal{D}}\left[d_{\mathrm{H}}^2(f, \hat{f}_{n,N})\right] \leq \varepsilon + O\left(\frac{C_{\mathcal{F},\Phi}}{n}\right) + O\left(\frac{nd}{N}\right) . \tag{34}$$

Otherwise, the trivial bound on $\mathrm{Tr}\{C^* I^*\}$ is only $O(n^4 d^4)$.

Remark 4.3 The optimal complexity index n may be obtained from eq. (34)

$$n_{\rm opt} = \left(\frac{C_{\mathcal{F},\Phi} N}{d}\right)^{1/2} \tag{35}$$

where d is the dimension of the data in the sample space.

Remark 4.4 The parameter m^* may be interpreted as the *effective number of parameters* of the model, under the misspecification of finite n. This parameter correlates the misspecified model's generalized information matrix C^*, with the pseudo-information matrix related to the density f_n^*, so that the effect of misspecification results in a modification in the number of *effective* parameters. We have argued that if n is sufficiently large, the number of parameters is given by $m \equiv n(d+2)$, which is exactly the number of parameters in the model. This result is related to those obtained by Amari and Murata [2], in a similar context. However, the latter authors considered the Kullback-Leibler divergence, and moreover did not study the approximation error.

Remark 4.5 How is the estimation affected by the dimensionality of the data? Obviously, the parameter m^*, which was observed to be the *effective* number of parameters, is proportional to d. The bound obtained in (34) makes this relation more transparent. However, the so called 'curse of dimensionality' is still an intrinsic part of the bound, though not quite evident by first inspection. The constant $C_{\mathcal{F},\Phi}$ embodies the dimensionality, giving rise to fixed term which may be exponential in the dimension d. This is made clear by observing the different sources comprising this constant, namely d-dimensional integrals due to the norms over the 'basis' densities and the target density (see Lemma 3.5). As a result we would expect that, although the approximation error converges at a rate of $O(1/n)$, the number of terms in the convex combination which is actually needed to reach a sufficiently small approximation error, may be exponentially large in the dimension.

Recall that the ε precision term appears in the error bound due to the insufficient representational power of the 'basis' functions (due to the bound on σ). Yet, under some specific conditions, this term can be removed yielding a bound which is only dependent on $C_{\mathcal{F},\Phi}$, and the parameters N, n, m^*. Following Barron & Cover [4] we have

Definition 1 *The information closure of the approximation class $\{\mathcal{G}_n\}$ is defined as*

$$\bar{\mathcal{G}} = \left\{ f \in \mathcal{F}_{c,\eta} \;\middle|\; \inf_{g\in\mathcal{G}} D(f||g) = 0 \right\} \tag{36}$$

where $\mathcal{G} = \cup \mathcal{G}_n$.

In other words, the densities in this class can be expressed in terms of an integral representation, in accordance with the definition of $\bar{f}$ (see eq. (16)).

Since a target density which is in the information closure of the approximation class may be approximated to any arbitrary degree of accuracy, we obtain in a similar manner to Corollary 3.2, that $d_H^2(f, f_n^0) \leq c/n$. Applying this result to Lemma 4.1 we have for all $f \in \bar{\mathcal{G}}$

$$d_H^2(f, f_n^*) \leq \frac{C_{\mathcal{F},\Phi}}{n} \tag{37}$$

where f_n^* is the density in $\mathcal{G}_n$ minimizing the KL divergence. Given a target density in the information closure of $\mathcal{G}$, and in view of the approximation bound (37), Theorem 4.1 may be restated accordingly. The expected error, comprised of the approximating error and the estimation error, will, under this assumption, be upper bounded by:

$$\mathrm{E}_{\mathcal{D}}\left[d_{\mathrm{H}}^2(f, \hat{f}_{n,N})\right] \leq O\left(\frac{C_{\mathcal{F},\Phi}}{n}\right) + O\left(\frac{m^*}{N}\right) \tag{38}$$

An alternative statement of the main result can be made by application of the Chebychev inequality to yield a bound in probability as follows.

Theorem 4.2 *Suppose assumptions B.1 - B.7 hold. Then for any $\delta > 0$, $\varepsilon_\tau > 0$ and N sufficiently large, the total error, $d_{\mathrm{H}}^2(f, \hat{f}_{n,N})$ (where $\hat{f}_{n,N}$ is the quasi maximum likelihood estimator) is bounded as follows:*

$$\begin{aligned} d_{\mathrm{H}}^2(f, \hat{f}_{n,N}) \quad &\leq \quad \varepsilon_\tau + \frac{C_{\mathcal{F},\Phi}}{n} + \frac{\mathrm{Tr}\{C^* I^*\}}{4N} \\ &\quad + \frac{1}{2N}\sqrt{\frac{\mathrm{Tr}\{C^* I^* C^* I^*\}}{2\delta}} + o\left(\frac{1}{N}\right) \end{aligned} \tag{39}$$

with probability $1 - \delta$.
The matrix C^* is the asymptotic covariance matrix defined in Lemma 4.2, I^* is the pseudo information matrix defined in eq. (32), and ε_τ is the resolution parameter which may be set to zero if the target density belong to the information closure of $\mathcal{G}$.

Proof See Appendix A.

5 Learning Algorithm

Having established the global error bound, eq. (33), we devote this short section to the final subject of interest, namely a learning algorithm which allows the parameters of the model to be estimated in an efficient manner. As we have shown in Theorem 4.1, the maximum likelihood estimation procedure can lead to efficient asymptotic estimation bounds. However, in order to compute the parameter values resulting from maximum likelihood estimation, one needs an efficient procedure for calculating the maximum of the likelihood function for any sample size. We shall focus on an iterative estimation procedure first formalized by Dempster *et al.* [8] and termed by them the expectation and maximization algorithm (EM). The EM algorithm in the context of mixture density estimation problems has been studied extensively since its formal introduction and has been at the heart of several recent research directions. Since the learning algorithm is not the main focus of this work, we content ourselves with a few brief remarks concerning the EM algorithm, referring the reader to the literature for a detailed discussion of the algorithm (see for example [21] for a comprehensive review, and [12] for an interesting recent contribution).

In order to apply the EM algorithm to our problem, we first need to fix n, the number of components in the mixture. This can be done using the asymptotic approximation given in eq. (35) and any a-priori knowledge about the constant c. We now wish to estimate the parameters of the model according to the method of maximum likelihood, and thus seek a point $\boldsymbol{\theta} \in \Theta$ (the parameter space) which is an extremum (local maximum) of the likelihood function. The likelihood equation, in the case of mixture models, is typically a complicated nonlinear function of the parameters, and thus requires an iterative optimization technique, in search of the maximum likelihood estimate point. However, it turns out that as long as the basis densities ϕ_σ belong to the class of exponential densities, the EM algorithm gives rise to a very efficient estimation scheme [21]. One of the attractive features of the algorithm is its global convergence (i.e. convergence from any initial condition). While the rate of convergence of the algorithm is still a matter of debate, there seem to be indications that in certain cases the convergence is in fact superlinear.

It is useful in this context to draw attention to an obvious implementation of density estimation using a neural network, transforming the output by an exponential function and normalizing appropriately, thus transforming the output into a density. Such a model would obviously be capable of approximating a given density to any accuracy (given the universal approximation power of neural nets) and following the recent results of Barron [5] regarding the degree of approximation characteristic of sigmoidal neural nets, an approximation bound could be derived. Since an EM algorithm is not available in the case of general function approximation, one would need to resort to some gradient-based procedure, such as conjugate gradients or quasi-Newton methods. While these procedures have some desirable theoretical attributes, they seem to scale more poorly with the complexity of the problem (expressed through the input dimension d and number of components n), and are often very sensitive to numerical errors.

As a final comment concerning learning we note that an entirely satisfactory approach to estimation in the context of convex combinations of densities would adaptively estimate the required number of components, n, without any need to assign it some prior value. In fact, such an adaptive scheme has been recently proposed and studied by Priebe [20] in the context of density estimation. While Priebe was able to prove that the algorithm is asymptotically consistent, it seems much harder to establish convergence rates.

6 Discussion

We have considered in this paper the problem of estimating a density function over a compact domain X. While the problem of density estimation can be viewed as a special case of function estimation, we believe that by constraining the study to densities (implying non-negativity and normalization of the functions), much insight can be gained. Specifically, the problem is phrased in the language of mixture models, for which a great deal of theoretical and practical results are available. Moreover, one can immediately utilize the powerful EM algorithm for estimating the parameters.

While we have restricted the mathematical analysis to continuous densities, so that the theory of Riemann integration can be used, we believe that our results can be extended to more general scenarios. We have been able, using Theorem 4.1, to present an upper bound to the error of the maximum likelihood (functional) estimator.

Barron [5] has recently presented upper bounds on the same quantity, in the context of function approximation, using an entirely different approach based on complexity regularization by the index of resolvability [4]. In this latter approach, one considers a finite covering of the parameter space, which allows one to define a new complexity limited estimator based on minimizing the sum of the log likelihood function and a complexity term, related to the size of the covering. An astute choice of complexity term then allows Barron to obtain an upper bound on the estimation error.

As opposed to Barron we have not added any complexity term, but rather used the results of White (1982) concerning misspecified models, together with the preliminary approximation (Lemma 3.4) and degree of approximation (Lemma 3.5) results, to obtain the required upper bounds. No need to discretise the parameter space, as has been done by Barron, is required in our approach. Furthermore, the approach of Barron [6] gives rise to an extra factor of $\log N$ in the second term on the rhs of eq. (33), making our bound in fact tighter. We believe the reason for this extra tightness in our case is related to the fact that White's results yield the exact asymptotic behavior of the quasi maximum likelihood estimator. In Barron's approach, however, a rather general form for the complexity function is used, which does not take into account the specific details of the estimation procedure.

Note, however, that the results we obtain concerning the approximation error, contain an extra factor of ε, which although arbitrarily small, cannot be set to zero due to Lemma 3.5. Moreover, unlike Barron's results [6] we do not prove the consistency of the estimator, and merely give upper bounds on the total error. The main contribution of this work is the upper bounds on the total error between a finite mixture model estimator, and an admissable target density. The issue of consistency can be approached using the method of sieves as in [10] and [27].

We believe that our results concerning the estimation error are not restricted to density estimation, and can be directly applied to function estimation using, for example, leastsquares estimation and the results of White [28] w.r.t. non-linear regression . In this context, we recently established upper bounds in the context of functional estimation using the mixture of experts model [30]. These bounds are derived in the framework of non-linear regression and utilize the results of White [28].

Acknowledgment The authors thank Manfred Opper for sending them copies of his work prior to publication, Robert Adler, Allan Pinkus and Joel Ratsaby for helpful discussions, and Paul Feigin for his constructive comments on the manuscript.

A Proof of Main Theorems

We present the proof of the main theorem.

Proof of Theorem 4.1: The proof proceeds by using first order Taylor expansion with remainder, applied to the Hellinger distance. Expanding around the point $\boldsymbol{\theta}^*$ we have:

$$\begin{aligned} d_H^2(f_n^*, \hat{f}_{n,N}) &= \int (\sqrt{f_n^*} - \sqrt{\hat{f}_{n,N}})^2 d\mathbf{x} = \int f_n^* \left(1 - \sqrt{\frac{f_{n,N}}{f_n^*}}\right)^2 d\mathbf{x} \\ &= \int f_n^* \left(1 - \left(\frac{f_n^* + (\boldsymbol{\theta}_N - \boldsymbol{\theta}^*)^T \nabla f_n^*}{f_n^*}\right)^{1/2}\right)^2 d\mathbf{x} + o_p\left(\frac{1}{N}\right) \quad . \end{aligned}$$

Denoting $\Delta\boldsymbol{\theta} = (\boldsymbol{\theta}_N - \boldsymbol{\theta}^*)$ and performing a first order binomial approximation one easily finds that

$$d_H^2(f_n^*, \hat{f}_{n,N}) = \frac{1}{4}\Delta\boldsymbol{\theta}^T \left[\int f_n^* (\nabla \log f_n^*)(\nabla \log f_n^*)^T d\mathbf{x}\right] \Delta\boldsymbol{\theta} + o_p\left(\frac{1}{N}\right) \tag{40}$$

where the order of the remainder follows from the results of Lemma 4.2. Denoting $I^* \stackrel{\triangle}{=} E_{\theta^*}[(\nabla \log f_n^*)(\nabla \log f_n^*)^T]$ we have

$$d_H^2(f_n^*, \hat{f}_{n,N}) = \frac{1}{4}\Delta\boldsymbol{\theta}^T I^* \Delta\boldsymbol{\theta} + o_p\left(\frac{1}{N}\right) \tag{41}$$

and by taking expectation with respect to the data $\mathrm{E}_{\mathcal{D}}[\cdot]$ we have the following expression as an approximation to the *estimation* error

$$\mathrm{E}_{\mathcal{D}}\left[d_H^2(f_n^*, \hat{f}_{n,N})\right] = \frac{1}{4}\mathrm{E}_{\mathcal{D}}[\Delta\boldsymbol{\theta}^T I^* \Delta\boldsymbol{\theta}] + o\left(\frac{1}{N}\right) = \frac{1}{4}\mathrm{E}_{\mathcal{D}}\left[\mathrm{Tr}\left(\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T I^*\right)\right] + o\left(\frac{1}{N}\right)$$

where the matrix I^* may be interpreted as the pseudo-information matrix, taken with respect to the density f_n^*. In order to evaluate the expectation term, we use Lemma 4.2,

$$\Delta\boldsymbol{\theta} \sim AN(0, \frac{1}{N}C^*)$$

from which we infer that

$$\mathrm{E}_{\mathcal{D}}\left[d_H^2(f_n^*, \hat{f}_{n,N})\right] \approx \frac{1}{4N}\mathrm{Tr}(C^* I^*) = O\left(\frac{m^*}{N}\right) \ ,$$

where $m^* = \mathrm{Tr}(C^* I^*)$. Finally, using Theorem 4.1 and the triangle inequality, eq. (24), we have

$$\begin{aligned} \mathrm{E}_{\mathcal{D}}[d_H^2(f, \hat{f}_{n,N})] &\leq \mathrm{E}_{\mathcal{D}}[d_H^2(f, f_n^*)] + \mathrm{E}_{\mathcal{D}}[d_H^2(f_n^*, \hat{f}_{n,N})] \\ &\leq O\left(\frac{C_{\mathcal{F},\Phi}}{n}\right) + O\left(\frac{m^*}{N}\right) \qquad \square \end{aligned} \tag{42}$$

Proof of Theorem 4.2 The proof follows from Chebychev's inequality:

$$\mathbb{P}\left\{ \mid d_H^2(f, \hat{f}_{n,N}) - \mathrm{E}[d_H^2(f, \hat{f}_{n,N})] \mid < \sqrt{\frac{\mathrm{Var}[d_H^2(f, \hat{f}_{n,N})]}{\delta}} \right\} > 1 - \delta, \qquad \forall \delta > 0. \tag{43}$$

The first moment of the squared Hellinger distance (between f_n^* and $\hat{f}_{n,N}$) was esablished in eq. (41), thus by applying the triangle inequality and utilizing the bound on the approximation error the result follows. The variance follows from the statistical properties of the asymptotic expansion which yields a quadratic form of Gaussian r.v.'s, as given by the expression in eq. (41). We omit the derivation of the variance expression and refer the reader to [13], where the fundamental properties of quadratic form of normal variables are studied. Plugging the moment expressions in eq. (43) we have the result. $\square$.

B Technical Assumptions

?? This appendix contains a list of the various assumptions needed in the proofs of the theorems in Section 4. Assumptions B.1-B.6 are simple restatements of those in White [25], whose results are utilized throughout the paper. Since we are concerned in this paper only with Riemann-Stieljes integration over compact domains, we have simplified somewhat the technical requirements appearing in White's paper. Assumption B.7 is essential for the proof of Theorem 4.1. In essence, this assumption ensures that the target function, as well as the approximant f_n^θ are positive, and greater than some threshold η, so that the bound given in Lemma 3.3 is applicable. The precise details of which assumptions are needed for proving each theorem, appear in the statement of the theorems in Section 4.

Assumption B.1 The random variables $\{\mathbf{x}_i\}_{i=1}^N$ whose density is estimated, are independent and identically distributed according to a probability density $f(\mathbf{x})$, where $\mathbf{x} \in X \subset R^d$.

Assumption B.2 Each member of the family of densities $f_n^\theta(\mathbf{x})$, is piece-wise continuous for each value of the parameter $\boldsymbol{\theta}$ taking values in a compact subset, Θ, of $p-$dimensional Euclidean space.

Assumption B.3 (a) $E[\log f(\mathbf{x})]$ exists and $|\log f_n^\theta(\mathbf{x})| \leq m(\mathbf{x})$ for all $\boldsymbol{\theta} \in \Theta$, where $m(\mathbf{x})$ is integrable with respect to f. (b) $E[\log(f/f_n^\theta)]$ has a unique minimum at $\boldsymbol{\theta}^*$ in Θ.

Assumption B.4 $\partial \log f_n^\theta(\mathbf{x})/\partial\theta_i$, $i = 1, 2, \ldots, p$, are integrable functions of $\mathbf{x}$ for each $\theta \in \Theta$ and continuously differentiable functions of $\boldsymbol{\theta}$ for each $\mathbf{x} \in X$.

Assumption B.5 $|\partial^2 \log f_n^\theta(\mathbf{x})/\partial\theta_i\partial\theta_j|$ and $|\partial f_n^\theta(\mathbf{x})/\partial\theta_i \cdot \partial f_n^\theta(\mathbf{x})/\partial\theta_j|$, $i, j = 1, 2, \ldots, p$ are dominated by functions integrable with respect to f for all $\mathbf{x} \in X$ and $\boldsymbol{\theta} \in \Theta$.

Assumption B.6 (a) $\boldsymbol{\theta}^*$ is interior to Θ; (b) $B(\boldsymbol{\theta}^*)$ is nonsingular; (c) $\boldsymbol{\theta}^*$ is a regular point of $A(\boldsymbol{\theta})$, namely $A(\boldsymbol{\theta})$ has constant rank in some open neighborhood of $\boldsymbol{\theta}^*$.

Assumption B.7 The convex model $f_n^\theta \in \mathcal{G}_n$ obeys the η positivity requirement for a sufficiently large complexity index n. Equivalently, $\exists n_0$ s.t. $\forall n > n_0$ we have $\inf_{x \in X} f_n^\theta(x) \geq \eta$.

References

[1] Adams, R.A. *Sobolev Spaces*, Academic Press, New York, 1975.

[2] Amari, S.I. and Murata, N. "Statistical Theory of Learning Curves under Entropic Loss Criterion", *Neural Computation*, vol. 5: 140-153, 1993.

[3] Barron, A.R. and Sheu, C.H. "Approximation of Density Functions By Sequences of Exponential Families," Annals of Statis., vol. 1 no.3, pp. 1347-1369, 1991.

[4] Barron, A.R. and Cover, T.M. "Minimum Complexity Density Estimation," *IEEE Trans. Inf. Theory*, vol. IT-37 no. 4, 1034-1054, 1991.

[5] Barron, A.R. "Universal Approximation Bound for Superpositions of A Sigmoidal Function," *IEEE Trans. Inf. Theory*, vol. IT-39, pp. 930-945, 1993.

[6] Barron, A.R. "Approximation and Estimation Bounds for Artificial Neural Networks", *Machine Learning*, vol. 4, pp. 115-133, 1994.

[7] Devroye, L. and Györfy, L.*Nonparametric Density Estimation: The L_1 View*, John Wiley & Sons, Inc., New York, 1985.

[8] Dempster, A.P. Laird, N.M. and Rubin, D.B. "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Roy. Statis. Soc.*, vol. B39, pp 1-38, 1977.

[9] Fergusson, T. *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, 1967.

[10] Geman S. and Hwang, C.R. "Nonparameteric Maximum Likelihood Estimation by the Method of Sieves", *Annals of Stats.*, vol. 10:2, 401-414, 1982.

[11] Jacobs, R.A., Jordan, M.I., Nowlan, S.J. and Hinton, G.E. "Adaptive Mixtures of Local Experts",*Neural Computation*, vol. 3:79-87, 1991.

[12] Jordan, M.I. and Jacobs, R.A. "Hierarchical mixtures of experts and the EM algorithm", *Neural Computation*, vol. 6:181-214, 1994.

[13] Johnson, N.L. and Kotz, S. *Distributions in Statistics. Continuous Univariate Distributions - 2.* Wiley, New-York, 1972.

[14] Jordan, M.I. and Xu, L. "Convergence Results for the EM Approach to Mixtures of Experts Architectures", *Neural Networks*, to appear.

[15] Le Cam, L. *Asymptotics in Statistics: Some Basic Concepts*, Springer Verlag, Berlin, 1990.

[16] Mhaskar, H. "Versatile Gaussian Networks", unpublished manuscript, 1995.

[17] Park, J., and Sandberg, I.W. "Universal Approximation Using Radial-Basis Function Networks", *Neural Computation*, vol. 3, pp. 246-257, 1991.

[18] Petersen, B.E. *Introduction to the Fourier Transform and Pseudo-Differential Operators*, Pitman Publishing, Boston, 1983.

[19] Powel, M.J.D. "The Theory of Radial Basis Function Approximation", pp. 105-210, in *Advances in Numerical Analysis*, ed. Light W., vol. 2, Oxford University Press, 1992.

[20] Priebe, C.E. "Adaptive Mixtures", *J. Amer. Statis. Assoc.* vol. 89:427, pp. 796-806, 1994.

[21] Redner, R.A. and Walker, H.F. "Mixture Densities, Maximum Likelihood and the EM Algorithm", *SIAM Review*, vol. 26, 195-239, 1984.

[22] Rudin, W. *Real and Complex Analysis*, Second Edition, McGraw-Hill, New York, 1987.

[23] Silverman, B.W. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, NY, 1986.

[24] Titterington, D.M., Smith, A.F.M., and Makov, U.E. *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York, 1985.

[25] White, H. "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, vol. 50 no. 1, 1-25, 1982.

[26] White, H. *Estimation, Inference and Specification Analysis*, Cambridge university press, 1994.

[27] White, H. "Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings", *Neural Networks*, vol. 3, 535-549, 1991.

[28] White, H. "Consequences and Detection of Misspecified Nonlinear Regression Models", *J. Amer. Statis. Assoc.*, vol. 76, 419-433, 1981.

[29] Wyner, A.D. and Ziv, Y. "Universal Classification with Finite Memory", to appear in *IEEE Trans. on Info. Theory*, 1996.

[30] Zeevi, A.J., Meir, R. and Maiorov, V. "Error Bounds for Functional Approximation and Estimation Using Mixtures of Experts", submitted to *IEEE Trans. on Info. Theory*, 1995.