Hierarchical Mixtures-of-Experts for Generalized Linear Models: Some Results on Denseness and Consistency

Wenxin Jiang and Martin A. Tanner
Department of Statistics, Northwestern University
Evanston, IL 60208, USA

Abstract

We investigate a class of hierarchical mixtures-of-experts (HME) models where exponential family regression models with generalized linear mean functions of the form $\psi(\alpha + \mathbf{x}^T\boldsymbol{\beta})$ are mixed. Here $\psi(\cdot)$ is the inverse link function. Suppose the true response y follows an exponential family regression model with mean function belonging to a class of smooth functions of the form $\psi(h(\mathbf{x}))$ where $h(\cdot) \in W_{2;K_0}^{\infty}$ (a Sobolev class over $[0,1]^s$). It is shown that the HME mean functions can approximate the true mean function, at a rate of $O(m^{-2/s})$ in L_p norm. Moreover, the HME probability density functions can approximate the true density, at a rate of $O(m^{-2/s})$ in Hellinger distance, and at a rate of $O(m^{-4/s})$ in Kullback-Leibler divergence. These rates can be achieved within the family of HME structures with a tree of binary splits, or within the family of structures with a single layer of experts. Here s is the dimension of the predictor $\mathbf{x}$. It is also shown that likelihood-based inference based on HME is consistent in recovering the truth, in the sense that as the sample size n and the number of experts m both increase, the mean square error of the estimated mean response goes to zero. Conditions for such results to hold are stated and discussed.

1 Introduction

Both the Mixtures-of-Experts (ME) model, introduced by Jacobs, Jordan, Nowlan and Hinton (1991), and the Hierarchical Mixtures-of-Experts (HME) model, introduced by Jordan and Jacobs (1994), provide important paradigms for learning from data, and are of mutual interest to researchers in artificial intelligence and in statistics. The fundamental problem is to learn a mapping in which the structure of the mapping varies for different regions of the input space. The ME and HME approach is to assign an "expert" network to each of these different regions and to then use a "gating" network to decide which experts should be used to determine the output. As part of the learning process, one needs to discover how to assign experts to the various regions and how to train the experts to adapt to their assigned task. The HME model has a tree-structure and can summarize the data at multiple scales of resolution due to its use of nested input regions. An introduction and application of mixing experts for generalized linear models (GLMs) are presented in Jordan and Jacobs (1994) and Peng, Jacobs and Tanner (1996).

Both ME and HME have been empirically shown to be powerful and general frameworks for examining relationships among variables in a variety of settings [Cacciatore and Nowlan (1994), Meilă and Jordan (1995), Ghahramani and Hinton (1996), Tipping and Bishop (1997) and Jaakkola and Jordan (1998)]. Despite the fact that ME and HME have been incorporated into neural network textbooks [e.g., Bishop (1995) and Haykin (1994) which features an HME design on the cover], there has been very little formal statistical justification [see Zeevi, Meir and Maiorov (1998)] of the methodology. In this paper we consider a fundamental question regarding ME and HME: *Given that we train an ME or HME network using noisy data, under what conditions are the inferences and predictions based on this system valid?* To answer this question we consider the denseness and consistency of the ME and HME networks. Before proceeding we present some notation regarding mixtures and hierarchical mixtures of generalized linear models and one-parameter exponential family regression models.

Generalized linear models, which are natural extensions of the usual linear model, are widely used in statistical practice [McCullagh and Nelder (1989)]. One-

parameter exponential family regression models [see Bickel and Doksum (1977), page 67] with generalized linear mean functions (GLM1) are special examples of the generalized linear models, where the probability distribution is totally determined by the mean function. In the regression context, a GLM1 model proposes that the conditional expectation $\mu(\mathbf{x})$ of a real response variable y (the output) is related to a vector of predictors (or inputs) $\mathbf{x} \in \Re^s$ via a generalized linear function $\mu(\mathbf{x}) = \psi(\alpha + \boldsymbol{\beta}^T\mathbf{x})$, with $\alpha \in \Re$ and $\boldsymbol{\beta} \in \Re^s$ being the regression parameters and $\psi^{-1}(\cdot)$ being the link function. Examples include the log link where $\psi(\cdot) = \exp(\cdot)$, the logit link where $\psi(\cdot) = \exp(\cdot)/\{1+\exp(\cdot)\}$, and the identity link which recovers the usual linear model. The inverse link function $\psi(\cdot)$ is used to map the entire real axis to a restricted region which contains the mean response. For example, when y follows a Poisson distribution conditional on $\mathbf{x}$, a log link is often used so that the mean is non-negative. In general, the GLM1 probability density function of y conditional on $\mathbf{x}$ is totally determined by the conditional mean function $\mu(\mathbf{x})$, having the form $p(y;\mathbf{x}) = \exp\{a_*(\mu)y + b_*(\mu) + c_*(y)\}$, where $\mu = \mu(\mathbf{x}) = \psi(\alpha + \boldsymbol{\beta}^T\mathbf{x})$, and $a_*(\cdot)$, $b_*(\cdot)$ and $c_*(\cdot)$ are some fixed functions. Such models include Poisson, binomial and exponential regression models, as well as the normal and gamma regression models with dispersion parameters regarded as known. In Section 4, we will discuss the situation when the dispersion parameter is also estimated. Before then, we focus on GLM1 exclusively.

A Mixtures-of-Experts model assumes that the total output is a locally-weighted average of the output of several GLM1 experts. It is important to note that such a model differs from standard mixture models [e.g., Titterington, Smith and Makov (1985)] in that the weights depend on the input. A generic expert labeled by an index J, proposes that the response y, conditional on the input $\mathbf{x}$, follows a probability distribution with density $p_J(y;\mathbf{x}) = \pi(h_J(\mathbf{x}), y) = \exp\{a_*(\mu_J)y + b_*(\mu_J) + c_*(y)\}$, where $\mu_J = \psi(h_J(\mathbf{x}))$ and $h_J(\mathbf{x}) = \alpha_J + \boldsymbol{\beta}_J^T\mathbf{x}$. The total probability density of y, after combining several experts, has the form $p(y;\mathbf{x}) = \sum_J g_J(\mathbf{x})p_J(y;\mathbf{x})$, where the local weight $g_J(\mathbf{x})$ depends on the input $\mathbf{x}$, and is often referred to as a gating function. The total mean response then becomes $\mu(\mathbf{x}) = \sum_J g_J(\mathbf{x})\mu_J(\mathbf{x})$. A simple Mixtures-of-Experts model takes J to be an integer. An HME model takes J as an integer vector, with dimension equal to the number of layers in the expert network.

An example of the HME model with two layers is given in Jordan and Jacobs (1994), as illustrated in Figure 1. Note that the HME is a graphical model with a probabilistic decision tree, where the weights of experts reflect a recursive stochastic decision process. In Figure 1, adapted from Jordan and Jacobs (1994), the expert label J is a two-component vector with each component taking either value 1 or 2. The total mean response μ is recursively defined by $\mu = \sum_{i=1}^2 \mathrm{g}_i\mu_i$ and $\mu_i = \sum_{j=1}^2 \mathrm{g}_{j|i}\mu_{ij}$, where g_i and $\mathrm{g}_{j|i}$ are logistic-type local weights associated with the "gating networks" for the choice of experts or expert groups at each stage of the decision tree, conditional on the previous history of decisions. Note that the product $\mathrm{g}_i\mathrm{g}_{j|i}$ gives a weight $g_J(\mathbf{x}) = \mathrm{g}_i\mathrm{g}_{j|i}$ for the entire decision history $J = (i, j)$. At the top of the tree is the mean response μ, which is dependent on the entire history of probabilistic decisions and also on the input $\mathbf{x}$.

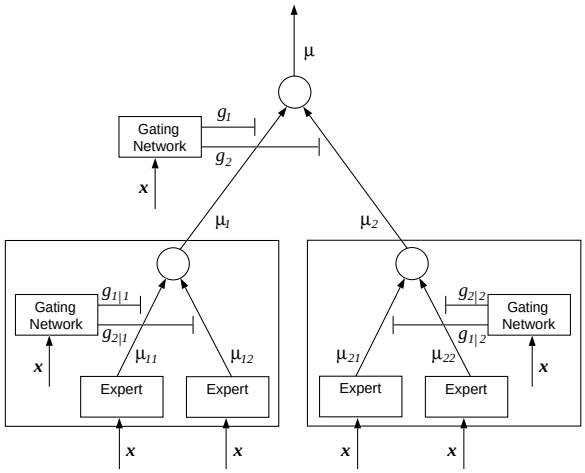

Figure 1: A Two-Layer Hierarchical Mixtures-of-Experts Model

One important issue is the approximation power of the HME models. Is the family of mean functions of the form $\sum_J g_J(\mathbf{x})\mu_J(\mathbf{x})$ proposed by HME rich enough to approximate an arbitrary smooth mean function of a certain family to any degree of accuracy? What precision, in a certain norm, can the approximation based on a specific number of experts achieve? Such problems of denseness and complexity are well-described and studied in the neural network literature [see Mhaskar (1996)]. A different question is the consistent learning property of HME with respect to a specific learning procedure. An HME model, as we will see later, is characterized by a parameter vector, which can be estimated based on a training data set consisting of n pairs of $(\mathbf{x}, y)$'s, following a learning procedure (or fitting method) such as least-squares or maximum likelihood approach. The consistency problem centers on whether the learning procedure will produce an estimated mean function which is close to the

true mean function, when the size of the training data set is sufficiently large. Various methods of measuring the closeness include the convergence in probability and the convergence in mean square error of the estimated mean function. The latter is a stronger mode of convergence due to Chebyshev's inequality [see Bickel and Doksum (1977), page 463] and is the mode of convergence we will consider in this paper.

Regarding these important theoretical questions, it is demonstrated by Zeevi, Meir and Maiorov (1998) that one-layer mixtures of linear model experts can be used to approximate a class of smooth functions as the number of experts increases, and the least-squares method can be used to estimate the mean response consistently when the sample size increases. One goal of this paper is to extend this result to HME for GLM1s with nonlinear link functions, and to consider the consistency of maximum likelihood estimation. The maximum likelihood (ML) approach has two advantages over the conventional least-squares approach. (i) The maximum likelihood approach gives the smallest asymptotic variance for the estimator of the mean response, in the case of correct model specification. (ii) The convenient EM algorithm can be used naturally for maximizing the likelihood, just as in the case of ordinary mixture models. However there are two difficulties for studying the consistency properties of a likelihood-based approach. (i) The maximum likelihood method deals with density functions rather than with mean functions. A result on the denseness of mean functions, such as the one stated in Zeevi, Meir and Maiorov (1998), is not enough. We need to establish a similar result for the density functions. We show that HME for GLM1 density functions can be used to approximate density functions of the form $\pi(h(\mathbf{x}), y)$, where $h(\cdot)$ is an arbitrary smooth function in a Sobolev class. (ii) The maximum likelihood method minimizes the Kullback-Leibler (KL) divergence, while the consistency properties for the estimates of mean responses are usually investigated by showing that the mean square error (MSE) of the estimated mean responses converge to zero in some fashion. We need to establish a relationship between the KL divergence of the *density functions* and the MSE, or the L_2 distance of the *mean functions*.

We also note that the parameterization of the HME, as shown in the next section, is not identifiable. Care is needed for statements about the parameter estimates, which are not unique.

2 Notation and Definitions

In the following, we briefly review the one-parameter exponential family regression model with a generalized linear mean function (GLM1).

2.1 GLM1

We first describe the one-parameter exponential family. Let $(A, \mathcal{F}_A, \lambda)$ be a general measure space. A probability density function $\pi(h, \cdot)$ in the one-parameter exponential family is labeled by one real parameter h, and has the form

$$\pi(h, y) = \exp\{a(h)y + b(h) + c(y)\} \text{ for } y \in A, \quad (1)$$

such that $\int_A \pi(h, y) d\lambda(y) = 1$ for each $h \in \Re$. The functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ all have known forms; $a(\cdot)$ and $b(\cdot)$ are analytic and have nonzero derivatives on $\Re$; and $c(\cdot)$ is measurable -$\mathcal{F}_A$.

Note that the one-parameter exponential models have some well-known properties. For example:

(i) The moment generating function exists in some neighborhood of the origin, and thus moments of all orders exist–see Theorem 1.4.2, Lehmann (1991, p.31).

(ii) For each positive integer k, $\mu_{(k)}(h) = \int_A y^k \pi(h, y) d\lambda$ is differentiable in h up to any orders, due to the analyticity of a, b and Theorem 1.4.1 of Lehmann (1991, p.29). In particular, we denote $\mu_{(1)}(h) = \psi(h) = \mu$ and $\mu_{(2)}(h) = \upsilon(h)$ as the first two moments.

(iii) The first moment can be expressed as $\mu = \psi(h) \equiv \int_A y\pi(h, y) d\lambda = -b'(h)/a'(h)$ for all real h and is analytic. $\psi : \Re \mapsto \psi(\Re)$ forms a $\mathcal{C}^\infty$-diffeomorphism. The inverse of $\psi(\cdot)$ is called the link function (McCullagh and Nelder 1989).

Some examples are:

Poisson: $\mathcal{P}(\mu)$ where $\mu = e^h$, $y \in A = \{0, 1, 2, \ldots\}$. Then

$$\pi(h, y) = \frac{e^{-\mu}}{y!}\mu^y = \exp\{hy - e^h - \log(y!)\}.$$

Here we can take $a(h) = h$, $b(h) = -e^h$, $c(y) = -\log(y!)$.

Normal (σ^2 known, > 0): $N(\mu, \sigma^2)$ where $\mu = h$, $y \in A = \Re$. Then

$$\begin{aligned}\pi(h, y) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \\ &= \exp\left\{\left(\frac{h}{\sigma^2}\right) y - \frac{h^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}.\end{aligned}$$

Here we can take $a(h) = h/\sigma^2$, $b(h) = -h^2/(2\sigma^2)$, $c(y) = -y^2/(2\sigma^2) - (1/2)\log(2\pi\sigma^2)$.

Gamma (γ known, > 0): ? $(\gamma, \gamma\mu^{-1})$
where $\mu = e^h$, $y \in A = \Re^+ = (0, \infty)$. Then

$\pi(h,y) = \frac{(\gamma e^{-h})^\gamma y^{\gamma-1}}{\Gamma(\gamma)} e^{-\gamma e^{-h} y} = \exp\left\{-\gamma e^{-h} y - \gamma h + \gamma \log\gamma - \log ?\,(\gamma) + (\gamma-1)\log y\right\}$. Here we can take $a(h) = -\gamma e^{-h}$, $b(h) = -\gamma h$, $c(y) = \gamma\log\gamma - \log ?\,(\gamma) + (\gamma-1)\log y$.

Binomial: $Bin(\nu, p)$ where $p = \nu^{-1}\mu = e^h/(1+e^h)$, $y \in A = \{0, 1, 2, \ldots, \nu\}$. Then

$$\pi(h,y) = \binom{\nu}{y} p^y (1-p)^{\nu-y} = \binom{\nu}{y} \frac{e^{hy}}{(1+e^h)^\nu}$$
$$= \exp\left\{hy - \nu\log(1+e^h) + \log\binom{\nu}{y}\right\},$$
$$\text{where } \binom{\nu}{y} = \frac{\nu!}{y!(\nu-y)!}.$$

Here we can take $a(h) = h$, $b(h) = -\nu\log(1+e^h)$, $c(y) = \log\binom{\nu}{y}$.

The GLM1 assumes that $h = \alpha + \boldsymbol{\beta}^T\mathbf{x}$, which introduces the dependence of y on an s-dimensional predictor $\mathbf{x}$ through the density function $\pi(h,y)$. Note that the functions a, b and c in (1) correspond, respectively, to the functions $a_* \circ \psi$, $b_* \circ \psi$ and c_* in notation of Section 1, where $\circ$ stands for composition.

Now we introduce a target family of regression models which is more flexible than the family of GLM1s, by allowing $h(\cdot)$ to be an arbitrary smooth function (of $\mathbf{x}$) in a Sobolev class.

2.2 The Family of Target Functions

Let $\Omega = [0,1]^s = \otimes_{q=1}^s [0,1]$, the space of the predictor $\mathbf{x}$, where $\otimes$ stands for the direct product. Let $A \subset \Re$ be the space of the response y. Let $(A, \mathcal{F}_A, \lambda)$ be a general measure space, $(\Omega, \mathcal{F}_\Omega, \kappa)$ be a probability space such that κ has a positive continuous density with respect to the Lebesgue measure on Ω, and $(\Omega \otimes A, \mathcal{F}_\Omega \otimes \mathcal{F}_A, \kappa \otimes \lambda)$ be the product measure space. Consider a random predictor-response pair $(\mathbf{X}_{(s\times 1)}, Y_{(1\times 1)})$. Suppose $\mathbf{X}$ has a probability measure κ, and $(\mathbf{X}, Y)$ has a probability density function (pdf) φ with respect to $\kappa \otimes \lambda$, where φ is a target function of the form

$$\varphi(\mathbf{x}, y) = \pi(h(\mathbf{x}), y). \tag{2}$$

Here $\pi(\cdot,\cdot) : \Re \otimes A \mapsto \Re$ has the one-parameter exponential form (1). In contrast to a GLM1 model, we allow a more flexible $h(\mathbf{x})$ in (2). Here $h : \Omega \mapsto \Re$ is assumed to have continuous second derivatives, $\sum_{\mathbf{k}:0\le|\mathbf{k}|\le 2} ||D^{\mathbf{k}}h||_\infty \le K_0$, where $\mathbf{k} = (k_1, \ldots, k_s)$ is an s-dimensional vector of nonnegative integers between 0 and 2, $|\mathbf{k}| = \sum_{j=1}^s k_j$, $||h||_\infty \equiv \sup_{\mathbf{x}\in\Omega}|h(\mathbf{x})|$, and $D^{\mathbf{k}}h \equiv \frac{\partial^{|\mathbf{k}|}h}{\partial x_1^{k_1}\ldots\partial x_s^{k_s}}$. In other words, $h \in W^\infty_{2;K_0}$, where $W^\infty_{2;K_0}$ is a ball with radius K_0 in a Sobolev space with sup-norm and second-order continuous differentiability. The conditional mean function $\mu(\cdot)$, corresponding to $\varphi(\cdot,\cdot)$, is obviously

$$\mu(\mathbf{x}) = \int_A y\varphi(\mathbf{x}, y)d\lambda(y) = \psi(h(\mathbf{x})) \tag{3}$$

for all $\mathbf{x}$ in Ω. Sobolev classes of mean functions similar to $W^\infty_{2;K_0}$ are also considered in Mhaskar (1996) and Zeevi *et al.* (1998). Our family of mean functions is a transformed class $\psi(W^\infty_{2;K_0})$ where ψ^{-1} is the link function. We have restricted the predictor $\mathbf{x}$ to $\Omega = [0,1]^s$ to simplify the exposition. The theorems of this paper actually hold for Ω being any compact subset of $\Re^s$. The compactness of Ω is needed in the techniques of our proofs. We also note that in the situation when Ω is the direct product of s closed intervals, suitable re-centering and re-scaling of each of the s components of $\mathbf{x}$ can transform Ω into $[0,1]^s$.

Denote the set of all pdfs $\varphi(\cdot,\cdot) = \pi(h(\cdot),\cdot)$ defined this way as $\boldsymbol{\Phi}$. This is the set of target functions that we wish to approximate.

Now we define the hierarchical mixtures-of-experts (HME) for GLM1s. They are the functions which we use for approximating a function in $\boldsymbol{\Phi}$.

2.3 The Family of HME of GLM1s

An approximator f in the HME family is assumed to have the following form:

$$f = f_\Lambda(\mathbf{x}, y; \theta) = \sum_{J\in\Lambda} g_J(\mathbf{x}; \mathbf{v})\pi(h_J(\mathbf{x}),\, y), \tag{4}$$

where $h_J(\mathbf{x}) = \alpha_J + \boldsymbol{\beta}_J^T\mathbf{x}$, and $\pi(\cdot,\cdot)$ is as defined in Section 2.1. The parameters of this model include $\alpha_J \in \Theta_\alpha \subset \Re$ and $\boldsymbol{\beta}_J \in \Theta_{\boldsymbol{\beta}} \subset \Re^s$ with Θ_α and $\Theta_{\boldsymbol{\beta}}$ being some compact sets, as well as $\mathbf{v}$ which is some parameter for the gating function g_J's. For convenience, we assume that $h_J \in W^\infty_{2;K_1}$ with a bound K_1, parallel to the assumption of $h \in W^\infty_{2;K_0}$ for the target functions. We use the symbol θ to represent the grand vector of parameters containing all the components of the parameters $\mathbf{v}$, α_J and $\boldsymbol{\beta}_J$ for all $J \in \Lambda$. In (4), Λ is the set of labels of all the experts in a network, referred to as a *structure*. Two quantities are associated with a structure: the dimension $\ell = \dim(\Lambda)$, which is the number of layers; and the cardinality $m = \mathrm{card}(\Lambda)$, which is the number of experts. An HME of ℓ-layers has a structure of the form $\Lambda = \otimes_{k=1}^\ell A_k$ where $A_k = \{1, \ldots, w_k\}$, $w_k \in \mathcal{N}$, and $k = 1, \ldots, \ell$. (We use $\mathcal{N}$ to denote the set of all positive integers.) We call $w_k = \mathrm{card}(A_k)$ as the number of expert branches, or the number of choice-legs at layer k, $k = 1, \ldots, \ell$. Note that in this paper we restrict attention to "rectangular-shaped" structures. A

generic expert label J in Λ can then be expressed as $J = (j_1, \ldots, j_\ell)$ where $j_k \in A_k$ for each k.

To characterize a structure Λ, we often claim that it belongs to a certain *set of structures*. We now introduce three such sets of structures, $\mathcal{J}$, $\mathcal{J}_m$ and $\mathcal{S}$, which will be used later when formulating the results. The set of all possible HME structures under consideration is $\mathcal{J} = \{\Lambda : \Lambda = \otimes_{k=1}^{\ell}\{1, \ldots, w_k\};\ w_k \in \mathcal{N};\ k = 1, \ldots, \ell;\ \ell \in \mathcal{N}\}$. The set of all HME structures containing no more than m experts is denoted as $\mathcal{J}_m = \{\Lambda : \Lambda \in \mathcal{J}, \mathrm{card}(\Lambda) \leq m\}$. We also introduce a symbol $\mathcal{S}$ to denote a generic subset of $\mathcal{J}$. This is introduced in order to formulate a major condition for some results of this paper to hold. This condition, to be formulated in the next section, will be specific to a generic subset $\mathcal{S}$ of HME structures. A trivial example of $\mathcal{S}$ is $\mathcal{J}$. Another example of $\mathcal{S}$ is $\mathcal{S}_L = \{\Lambda : \Lambda \in \mathcal{J}, \dim(\Lambda) \leq L\}$, which includes all structures with L or less layers. In particular, $\mathcal{S}_1$ represents the set of single-layer structures. A third example of $\mathcal{S}$ is $\mathcal{S}_B = \{\Lambda : \Lambda = \otimes_{k=1}^{\ell}\{1, 2\};\ \ell \in \mathcal{N}\}$, which represents the set of trees with binary splits.

Associated with a structure Λ is a family of vectors of gating functions. Each member is called a *gating vector* and is labeled by a parameter vector $\mathbf{v} \in V_\Lambda$, V_Λ being some parameter space specific to the structure Λ. Denote a generic gating vector as $G_{\mathbf{v},\Lambda} \equiv (g_J(\cdot; \mathbf{v}))_{J \in \Lambda}$. We assume the $g_J(\mathbf{x}; \mathbf{v})$'s to be nonnegative, with sum equal to unity, and continuous in $\mathbf{x}$ and $\mathbf{v}$. Note that $\int_A f_\Lambda(\mathbf{x}, y; \theta) d\lambda(y) = 1$ is ensured. Let $\mathcal{G} = \{G_{\mathbf{v},\Lambda} : \mathbf{v} \in V_\Lambda, \Lambda \in \mathcal{S}\}$ be the family of gating vectors defined on the set of structures $\mathcal{S}$, which will be referred to as a *gating class* defined on $\mathcal{S}$.

In the following, we define the *logistic gating class* $\mathcal{G} = \mathcal{L}$ on the set of all structures $\mathcal{J}$. This class has been commonly used in the literature [see Jordan and Jacobs (1994)]. Here, for each structure Λ in $\mathcal{J}$ and each label J in Λ, a gating function $g_J = g_J(\cdot, \mathbf{v})$ is defined recursively. Suppose J is an ℓ-dimensional integer $(j_1, j_2, \ldots, j_\ell)$. Then,

$$g_J \equiv g_{j_1 j_2 \ldots j_\ell} = g_{j_1} g_{j_2|j_1} \cdots g_{j_\ell|j_1 j_2 \ldots j_{\ell-1}}. \qquad (5)$$

Here, for each q, the factor $g_{j_q|j_1 \ldots j_{q-1}}$ takes a multinomial logit form:

$$g_{j_q|j_1 \ldots j_{q-1}} = \frac{\exp(\xi_{j_q|j_1 \ldots j_{q-1}})}{\sum_{k=1}^{w_q} \exp(\xi_{k|j_1 \ldots j_{q-1}})}, \qquad (6)$$

where $\xi_{k|j_1 \ldots j_{q-1}} = \phi_{k|j_1 \ldots j_{q-1}} + \boldsymbol{\gamma}_{k|j_1 \ldots j_{q-1}}^T \mathbf{x}$, $(\phi_{k|j_1 \ldots j_{q-1}}, \boldsymbol{\gamma}_{k|j_1 \ldots j_{q-1}}^T) \in \Re^{s+1}$, $k = 1, \ldots, w_q$. Usually it is assumed that

$$\phi_{w_q|j_1 \ldots j_{q-1}} = \boldsymbol{\gamma}_{w_q|j_1 \ldots j_{q-1}} = \xi_{w_q|j_1 \ldots j_{q-1}} = 0,$$

since otherwise a transformation

$$\begin{cases} \phi_{k|j_1 \ldots j_{q-1}} & \to & \phi_{k|j_1 \ldots j_{q-1}} + \phi_0 \\ \boldsymbol{\gamma}_{k|j_1 \ldots j_{q-1}} & \to & \boldsymbol{\gamma}_{k|j_1 \ldots j_{q-1}} + \boldsymbol{\gamma}_0 \quad \text{all } k = 1, \ldots, w_q \end{cases}$$

would leave the probability density function $f_\Lambda(\mathbf{x}, y; \theta)$ unchanged. Note that the grand vector of "gating parameters" $\mathbf{v}$ includes all components of $(\phi_{j_q|j_1 \ldots j_{q-1}},\ \boldsymbol{\gamma}_{j_q|j_1 \ldots j_{q-1}}^T)$, where j_r take over all values $\{1, \ldots, w_r\}$ for $r = 1, \ldots, q-1$ and over all values $\{1, \ldots, w_r - 1\}$ for $r = q$; for all $q = 1, \ldots, \ell$. It is easy to see that $\dim(\mathbf{v}) = (s+1)(m-1)$, and the parameter space V_Λ for $\mathbf{v}$ is $\Re^{(s+1)(m-1)}$, where $m = w_1 \ldots w_\ell = \mathrm{card}(\Lambda)$. Note that the gating functions constructed in this way are analytic for $(\mathbf{v}^T, \mathbf{x}^T) \in \Re^{(s+1)(m-1)} \otimes \Re^s$. The space of regression parameters (or "expert parameters") $(\alpha_J, \boldsymbol{\beta}_J^T)$'s, corresponding to structure Λ, is $(\Theta_\alpha \otimes \Theta_{\boldsymbol{\beta}})^{\otimes m}$, which is a compact subset of $\Re^{(s+1)m}$. The space of grand parameters θ's, corresponding to structure Λ, is

$$\tilde{\Theta}_\Lambda = (\Theta_\alpha \otimes \Theta_{\boldsymbol{\beta}})^{\otimes m} \otimes \Re^{(s+1)(m-1)}. \qquad (7)$$

Here the $(2m-1)(s+1)$ dimensional grand parameter θ includes all components of the gating parameters from $\mathbf{v}$ and the expert parameters from $(\alpha_J, \boldsymbol{\beta}_J^T)_{J \in \Lambda}$.

Now we are ready to define the family of approximator functions. Let Π_Λ be the set of all function f_Λ's of the form (4), specific to a structure Λ, which can be denoted as $\Pi_\Lambda = \{f_\Lambda(\cdot, \cdot; \theta) : \theta \in \tilde{\Theta}_\Lambda\}$. This set Π_Λ is the set of HME functions from which an optimal function is chosen by the maximum likelihood method to approximate the truth. It is assumed that a structure Λ is chosen *a priori*. In practice, people often analyze data using different choices of structures and select the best fitting model. We consider in this paper choosing among the set of structures $\mathcal{J}_m \cap \mathcal{S}$. Denote

$$\mathbf{\Pi}_{m,\mathcal{S}} = \{f : f \in \Pi_\Lambda; \Lambda \in \mathcal{J}_m \cap \mathcal{S}\}. \qquad (8)$$

This set, $\mathbf{\Pi}_{m,\mathcal{S}}$, is the family of HME functions for which we examine the approximation rate in $\mathbf{\Phi}$, as $m \to \infty$. Note that this family of HME functions is specific to m, the maximum number of experts, as well as to some subset $\mathcal{S}$ of HME structures, which will be specified later. We do not explicitly require that $\mathbf{\Pi}_{m,\mathcal{S}}$ be a subset of $\mathbf{\Phi}$ in this paper.

Each HME density function $f_\Lambda(\mathbf{x}, y; \theta)$ generates a mean function $\mu_\Lambda(\mathbf{x}; \theta)$ by

$$\begin{aligned} \mu_\Lambda(\mathbf{x}; \theta) &= \int_A y f_\Lambda(\mathbf{x}, y; \theta) d\lambda(y) \\ &= \sum_{J \in \Lambda} g_J(\mathbf{x}; \mathbf{v}) \psi(\alpha_J + \mathbf{x}^T \boldsymbol{\beta}_J), \end{aligned} \qquad (9)$$

where $\psi(\cdot) = \int_A y \pi(\cdot, y) d\lambda(y)$.

The parameterization of the HME functions is not identifiable, in the sense that two different parameters θ in $\tilde{\Theta}_\Lambda$ can represent the same density function f in $\mathbf{\Pi}_{m,\mathcal{S}}$. For example,, the density functions are invariant under permutation of the expert label J's. Also, if two experts J and J' propose the same output, i.e., if $\alpha_J = \alpha_{J'}$ and $\boldsymbol{\beta}_J = \boldsymbol{\beta}_{J'}$, then the mixing proportions for these two experts can be arbitrary, as long as the sum of the two weights are unchanged. This can lead to the non-identifiability of some components of the parameter $\mathbf{v}$. Our description of the estimation procedure and the statement of the results will take these identifiability issues into account. The identifiability issues also suggest that it makes more sense to formulate the consistency problem in terms of the estimated mean response, rather than to look at the consistency of the parameter estimates.

2.4 The Method of Estimation

We will use the maximum likelihood method to train the architecture. Suppose we estimate the mean response $\mu(\mathbf{x})$ based on a data set of n predictor-response pairs $(\mathbf{X}_i, Y_i)$, $\mathbf{X}_i \in \Omega$, $Y_i \in A$, $i = 1, \ldots, n$. Let the measure spaces $(\Omega, \mathcal{F}_\Omega, \kappa)$ and $(A, \mathcal{F}_A, \lambda)$ be as introduced in Section 2.1. Assume that $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.) random vectors. The probability measure for $\mathbf{X}_i$ is κ. The probability measure of Y_i conditional on $\mathbf{X}_i = \mathbf{x}$ has a density $\varphi(\mathbf{x}, \cdot)$ [defined in (2)] with respect to the measure λ, for all $\mathbf{x} \in \Omega$.

The log-likelihood function based on the HME model is

$$\text{Ł}_{n,\Lambda}(\theta;\omega) = n^{-1} \sum_{i=1}^{n} \log\{f_\Lambda(\mathbf{X}_i, Y_i; \theta)/\varphi_0(\mathbf{X}_i, Y_i)\}, \tag{10}$$

where $f_\Lambda(\cdot, \cdot; \theta) \in \Pi_\Lambda$ is defined in Section 2.3, $\theta \in \tilde{\Theta}_\Lambda$, ω is the stochastic sequence of events $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots$, and $\varphi_0(\mathbf{X}_i, Y_i)$ can be any positive measurable function of the observed data that does not depend on the parameter θ. In this paper, we choose $\varphi_0(\mathbf{X}_i, Y_i) = e^{c(Y_i)}$, where $c(\cdot)$ is defined in (1). It turns out that such a choice makes the log-likelihood function uniformly convergent to its expectation, for almost all ω, in any compact subset of parameters, as $n \to \infty$. Define the maximum likelihood estimator (MLE) $\hat{\theta}_{n,\Lambda}(\omega)$ to be a maximizer (can be one out of many) of $\text{Ł}_{n,\Lambda}(\theta;\omega)$ over a compact set $\tilde{B}_\Lambda \subset \tilde{\Theta}_\Lambda$, i.e.,

$$\hat{\theta}_{n,\Lambda}(\omega) = \arg\max_{\theta \in \tilde{B}_\Lambda} \{\text{Ł}_{n,\Lambda}(\theta;\omega)\}. \tag{11}$$

The maximum likelihood method, in the large sample size limit, essentially searches for θ which minimizes the KL divergence $\text{KL}(f_\Lambda, \varphi)$ between $f_\Lambda = f_\Lambda(\cdot, \cdot; \theta) \in \Pi_\Lambda$ and $\varphi = \varphi(\cdot, \cdot) \in \mathbf{\Phi}$, where

$$\text{KL}(f, g) \equiv \int_{\Omega \otimes A} g(\mathbf{x}, y) \log\left\{\frac{g(\mathbf{x}, y)}{f(\mathbf{x}, y)}\right\} d\kappa(\mathbf{x}) d\lambda(y). \tag{12}$$

It turns out that the KL divergence $\text{KL}(f_\Lambda, \varphi)$ is always well defined (see Corollary 1 later). Due to the non-identifiability of the parameterization, there is *a set* of θ's in $\tilde{B}_\Lambda$ that minimize the KL divergence. Denote this set as Θ_Λ, which could be expressed as

$$\Theta_\Lambda = \{\theta \in \tilde{B}_\Lambda : \theta = \arg\min_{\theta^* \in \tilde{B}_\Lambda} \text{KL}(f_\Lambda(\cdot, \cdot; \theta^*), \varphi)\}. \tag{13}$$

Based on *any* MLE $\hat{\theta}_{n,\Lambda} = \hat{\theta}_{n,\Lambda}(\omega)$, an estimated mean response can be constructed as $\mu_\Lambda(\mathbf{x}; \hat{\theta}_{n,\Lambda})$. We do not explicitly require that for two different global MLEs the estimated mean responses be the same. The MSE of *an* estimated mean response is defined by

$$(\text{MSE})_{n,\Lambda} = \text{E} \int \{\mu_\Lambda(\mathbf{x}; \hat{\theta}_{n,\Lambda}) - \mu(\mathbf{x})\}^2 d\kappa(\mathbf{x}), \tag{14}$$

where E is the expectation taken on the MLE $\hat{\theta}_{n,\Lambda}$, μ_Λ and μ are defined in (9) and (3), respectively.

2.5 Technical Definitions

Some technical definitions are introduced below. We will use these definitions to formulate a major condition for our results to hold.

Definition 1 *(Fine Partition). For $\nu = 1, 2, \ldots$, let $\mathbf{Q}^{(\nu)} = \{Q_J^{(\nu)}\}_{J \in \Lambda^{(\nu)}}$, $\Lambda^{(\nu)} \in \mathcal{J}$, be a partition of $\Omega \subset \Re^s$. (This means that for fixed ν, the $Q_J^{(\nu)}$'s are mutually disjoint subsets of $\Re^s$ whose union is Ω.) Let $p_\nu = \text{card}(\Lambda^{(\nu)})$, $(p_\nu \in \mathcal{N})$.*

If $p_\nu \to \infty$, and if for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in Q_J^{(\nu)}$, $\rho(\boldsymbol{\xi}, \boldsymbol{\eta}) \equiv \max_{1 \le q \le s} |(\boldsymbol{\xi} - \boldsymbol{\eta})_q| \le c_0 / p_\nu^{1/s}$ for some constant c_0 independent of ν, J, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, then $\{\mathbf{Q}^{(\nu)} : \nu = 1, 2, \ldots\}$ is called a sequence of fine partitions *with* structure sequence $\{\Lambda^{(\nu)}\}$, cardinality sequence $\{p_\nu\}$, *and bounding constant c_0.*

Definition 2 *(Sub-Geometric). A sequence $\{a_\nu\}$ is* sub-geometric with rate bounded by M_2, *if $a_\nu \in \mathcal{N}$, $a_\nu \to \infty$ as $\nu \to \infty$, and $1 < |a_{\nu+1}/a_\nu| < M_2$ for all $\nu = 1, 2, \ldots$, for some finite constant M_2.*

In the following we introduce some measures of the discrepancies between a pdf f in Π_Λ [of the form (4)] and a pdf φ in $\mathbf{\Phi}$ [of the form (2)]. One of them is the KL distance $\text{KL}(f, \varphi)$ [see (12)]. Another is the Hellinger distance

$$d_H(f, \varphi) = \{\int (\sqrt{f} - \sqrt{\varphi})^2 d\lambda d\kappa\}^{1/2}. \tag{15}$$

This is a true distance, which is invariant under rescaling of the measures λ and κ. A third description is the L_2 distance between the means:

$$d_2(\mu_f, \mu_\varphi) = ||\mu_\Lambda(\cdot, \theta) - \mu(\cdot)||_{2,\kappa} = \{\int (\mu_f - \mu_\varphi)^2 d\kappa\}^{1/2}, \quad (16)$$

where $\mu_f = \int y f d\lambda$ and $\mu_\varphi = \int y \varphi d\lambda$, for f in Π_Λ and φ in $\mathbf{\Phi}$. This measure is used since it is closely related to the MSE defined in Section 2.4.

The fourth measure of discrepancy between f in Π_Λ and φ in $\mathbf{\Phi}$ is called the "upper divergence". For $f = \sum_{J \in \Lambda} g_J \pi(h_J, y)$ and $\varphi = \pi(h, y)$, the upper divergence is defined as

$$Q(f, \varphi) = \int \sum_{J \in \Lambda} g_J(\mathbf{x}; \mathbf{v}) \{h_J(\mathbf{x}) - h(\mathbf{x})\}^2 d\kappa, \quad (17)$$

where $h_J(\mathbf{x}) = \alpha_J + \boldsymbol{\beta}_J^T \mathbf{x}$. Note that the idea of HME approximation is to partition the input space "softly" according to the g_J's, and use a linear function $h_J(\mathbf{x})$ to approximate $h(\mathbf{x})$ in each partition, so as to approximate the (conditional) pdf $\pi(h(\mathbf{x}), \cdot)$ for all $\mathbf{x}$. The upper divergence measures how good is this softly-partitioned linear approximation. The name "upper divergence" is due to the following lemma, stated without proof, which implies that Q is stronger than the other divergence measures, i.e., KL, d_H and $d_2(\mu_f, \mu_\varphi)$.

Lemma 1 *(Strengths of Divergence Measures.) For any structure Λ, any f in Π_Λ and any φ in $\mathbf{\Phi}$, we have*

(a) $d_2^2(\mu_f, \mu_\varphi) \leq 4 M_I d_H^2(f, \varphi)$.

(b) [1] $d_H^2(f, \varphi) \leq KL(f, \varphi)$.

(c) $\mathrm{KL}(f, \varphi) \leq M_{II} Q(f, \varphi)$.

Here, $M_I = \sup_{|h| \leq K} \{\int y^2 \pi(h, y) d\lambda\}$, $K = \min\{K_0, K_1\}$ *where* K_0 *and* K_1 *are bounds of* $h(\cdot)$ *and* $h_J(\cdot)$ *in the Sobolev class* $W^\infty_{2;K_0}$ *and* $W^\infty_{2;K_1}$, *respectively.* $M_{II} = \frac{1}{2}\{\sup_{|h| \leq K} |\int y \pi(h, y) d\lambda| \cdot \sup_{|h| \leq K} |a''(h)| + \sup_{|h| \leq K} |b''(h)|\}$, *where* $a(\cdot)$ *and* $b(\cdot)$ *are defined as in (1).*

Remark 1 M_I and M_{II} are finite constants, due to the continuity of a'', b'', and $\int y^k \pi(h, y) d\lambda$ $(k = 1, 2)$, as functions of h.

Corollary 1 *All the divergence measures* $d_2(\mu_f, \mu_\varphi)$, d_H, KL *and* Q *are positive and finite.*

[1] This lemma appeared in, for example, Haussler and Opper (1995), and Zeevi and Meir (1997).

Proof: This is obvious since Q involves an integration of a continuous function over the compact space Ω of input $\mathbf{x}$. □

In the next section (Lemma 2), we will see that the HME functions related to a set of structures $\mathcal{S}$ are dense in $\mathbf{\Phi}$ in upper divergence, under a condition on the gating class defined on $\mathcal{S}$ (Condition $A_{\mathcal{S},1}$). This implies, in turn, the "denseness" in KL and d_H.

3 Results and Conditions

In the following, we state some regularity conditions, as well as some results which hold under these conditions.

Condition 1 *($A_{\mathcal{S},p}$). For a subset $\mathcal{S} \subset \mathcal{J}$, there is a fine partition sequence $\{\{Q_J^{(\nu)}\}_{J \in \Lambda_0^{(\nu)}} : \Lambda_0^{(\nu)} \in \mathcal{S}, \nu = 1, 2, \ldots\}$ with a bounding constant c_0 and a cardinality sequence $\{p_\nu : \nu = 1, 2, \ldots\}$, such that $\{p_\nu^{1/s}\}$ is sub-geometric with rate bounded by a constant M_2; and for all ν, for all $\varepsilon > 0$, there exists $\mathbf{v}_\varepsilon \in V_{\Lambda_0^{(\nu)}}$ and a gating vector*

$$G_{\mathbf{v}_\varepsilon, \Lambda_0^{(\nu)}} = \{g_J(\mathbf{x}; \mathbf{v}_\varepsilon)\}_{J \in \Lambda_0^{(\nu)}} \in \mathcal{G}, \quad \Lambda_0^{(\nu)} \in \mathcal{S}, \quad \textit{such that}$$

$$\sup_{J \in \Lambda_0^{(\nu)}} \|g_J(\cdot; \mathbf{v}_\varepsilon) - \chi_{Q_J^{(\nu)}}(\cdot)\|_{p,\sigma} \leq \varepsilon. \quad (18)$$

Here, $\|f(\cdot)\|_{p,\sigma} \equiv \left\{\int_\Omega |f(\mathbf{x})|^p d\sigma(\mathbf{x})\right\}^{1/p}$, *where* $p \in \mathcal{N}$; *σ is any probability measure on Ω which has a positive continuous density with respect to the Lebesgue measure; $\chi_B(\cdot)$ is the characteristic function for a subset B of Ω, i.e., $\chi_B(\mathbf{x}) = 1$ if $\mathbf{x} \in B$, 0 otherwise.*

This condition is a restriction on the gating class $\mathcal{G}$ defined on a set of structures $\mathcal{S}$. Loosely speaking, it indicates that the vectors of local gating functions in the parametric family should arbitrarily approximate the vector of characteristic functions for a partition of the predictor space Ω, as the cells of the partition become finer. Under this condition, the soft partitions are flexible enough to approximate a hard partition of the input space Ω, with the size of each cell having order $(1/m)^{1/s}$, m being the number of experts in the structure, $s = \dim(\mathbf{x})$. In each of these m small cells, a linear approximation $h_J(\mathbf{x}) = \alpha_J + \boldsymbol{\beta}_J^T \mathbf{x}$ for a second order continuously differentiable function $h(\mathbf{x})$ has an error bound of order $(1/m)^{2/s}$, by a second order Taylor expansion. It is not surprising that the HME mean functions, using the $\psi(h_J(\mathbf{x}))$'s as the building blocks, can approximate the mean functions in the target family of the form $\psi(h(\mathbf{x}))$, with an error bound of the same order. (Here ψ is the inverse link function.) This is summarized in Theorem 1 below.

Theorem 1 *(Approximation Rate of the Mean Functions.) Under the condition $A_{\mathcal{S},p}$,*

$$\sup_{\mu\in\psi(W^{\infty}_{2;K_0})}\inf_{f\in\mathbf{\Pi}_{m,\mathcal{S}}}\|\mu_f-\mu\|_{p,\sigma}\le\frac{c}{m^{2/s}}$$

for some constant $c>0$ independent of m. Here $s=\dim(\mathbf{x})$, m is the maximal number of experts in the HME family $\mathbf{\Pi}_{m,\mathcal{S}}$, $\mu_f=\int yfd\lambda$, $\mathbf{\Pi}_{m,\mathcal{S}}$ is defined in Section 2.3, $\psi(W^{\infty}_{2;K_0})$ is the set of all functions of the form (3), and $\|(\cdot)\|_{p,\sigma}$ is as defined in Condition $A_{\mathcal{S},p}$.

In the following we go one step further to discuss the denseness and approximation properties of the *density functions* of the HME of GLM1s, which is useful in investigating the consistency property of the maximum likelihood approach. From the discussions following Condition $A_{\mathcal{S},p}$, we also conclude that the upper divergence, consisting of the squares of the differences between h_J and h, should have an error bound of order $(1/m)^{4/s}$. This is summarized in Lemma 2.

Lemma 2 *(Approximation Rate in Upper Divergence.) If Condition $A_{\mathcal{S},1}$ holds, then we have*

$$\sup_{\varphi\in\mathbf{\Phi}}\inf_{f\in\mathbf{\Pi}_{m,\mathcal{S}}}Q(f,\varphi)\le\frac{c}{m^{4/s}},$$

where $s=\dim(\mathbf{x})$, m is the maximal number of experts in the HME family $\mathbf{\Pi}_{m,\mathcal{S}}$, and c is a positive constant independent of m.

From this lemma and Lemma 1, the following theorems on the approximation rates of the HME *density* functions in Hellinger distance and in KL divergence are obvious.

Theorem 2 *(Approximation Rate in Hellinger Distance.) Under the condition $A_{\mathcal{S},1}$,*

$$\sup_{\varphi\in\Phi}\inf_{f\in\mathbf{\Pi}_{m,\mathcal{S}}}d_H(f,\varphi)\le\frac{c}{m^{2/s}},$$

for some positive constant c independent of m. Here d_H is the Hellinger distance defined in (15).

Theorem 3 *(Approximation Rate in KL Divergence.) Under Condition $A_{\mathcal{S},1}$,*

$$\sup_{\varphi\in\Phi}\inf_{f\in\mathbf{\Pi}_{m,\mathcal{S}}}\mathrm{KL}(f,\varphi)\le c/m^{4/s},$$

for some positive constant c independent of m. Here KL *is the KL divergence defined in(12).*

The constant c's in the above theorems or lemmas can be different.

All these results depend on a major condition $A_{\mathcal{S},p}$. The following remark claims that it is satisfied by certain gating functions defined on certain structures.

Remark 2 (a). Condition $A_{\mathcal{S},p}$ is satisfied (for any $p\in\mathcal{N}$) by the logistic gating class $\mathcal{G}=\mathcal{L}$ defined on the set of structures $\mathcal{S}=\mathcal{S}_B$ for trees with binary splits (Section 2.3). This is because, roughly speaking, a logistic function from a binary split has the form $(1+e^{-\beta(z-z_0)})^{-1}$, which can approximate a step function $S(z-z_0)$ as β increases, for any location of jump z_0. The gating functions in a binary tree involves products of the logistic functions (and their complements), which can approximate products of step functions which form the characteristic functions of a fine partition. In this way Condition $A_{\mathcal{S},p}$ can be proved. This implies that the approximation rates in the theorems stated above apply to HME of GLM1s with binary trees.

(b). Jiang and Tanner (1999) [Section 5 Remark (i)] show that Condition $A_{\mathcal{S},p}$ is also satisfied (for any $p\in\mathcal{N}$) by the logistic gating class $\mathcal{L}$ defined on the set of single-layer structures $\mathcal{S}_1$, which correspond to the MEs. This implies that the approximation rates in the theorems stated above apply to ME of GLM1s.

(c). Another class of gating functions can be defined only on the binary trees (in $\mathcal{S}_B$). There, the logistic gating functions in (6) are replaced by continuous cumulative distribution functions (cdfs). One example is to use the normal cdf. In this way, the gating factor $g_{j_q|j_1\cdots j_{q-1}}$ of (6) becomes $\Phi(\xi_{j_1\cdots j_{q-1}})$ if $j_q=1$, or $1-\Phi(\xi_{j_1\cdots j_{q-1}})$ if $j_q=2$; where $\xi_{j_1\cdots j_{q-1}}=\phi_{j_1\cdots j_{q-1}}+\boldsymbol{\gamma}^T_{j_1\cdots j_{q-1}}\mathbf{x}$. A similar argument as in part (a) of this remark shows that Condition $A_{\mathcal{S},p}$ is satisfied for this new gating class for any $p\in\mathcal{N}$.

The next condition is useful for proving the consistency of the maximum (ML) likelihood learning method.

Condition 2 *(Scope of Maximum Likelihood Searching) The scope of the maximum likelihood (ML) searching, $\tilde{B}_\Lambda$, is a compact set which is so large that it contains a point θ^Q_Λ which minimizes the upper divergence $Q(f_\Lambda,\varphi)$ between $f_\Lambda(\cdot,\cdot;\theta)\in\Pi_\Lambda$ and $\varphi(\cdot,\cdot)\in\Phi$ among all choices of θ in $\tilde{\Theta}_\Lambda$, where*

$$\begin{aligned}&Q(f_\Lambda(\cdot,\cdot,\theta^Q_\Lambda),\ \varphi(\cdot,\cdot))\\&\quad=\inf_{\theta\in\tilde{\Theta}_\Lambda}Q(f_\Lambda(\cdot,\cdot,\theta),\ \varphi(\cdot,\cdot))=\inf_{f\in\Pi_\Lambda}Q(f,\varphi).\end{aligned}$$

This condition is similar to a usual condition under correct model specification, requiring that the scope of ML search should contain the true parameter so as to make the MLE consistent. The difference here is that there is no "true parameter", since the likelihood

functions are constructed based on the HME densities, which can only be used to *approximate* the true pdf in $\mathbf{\Phi}$. Condition 2 ensures that the ML searching area is big enough to contain an "optimal point" (instead of the true parameter) which minimizes the upper divergence between the true density and the HME density. This feature will be useful when proving the consistency result of the ML approach under model misspecification, when the likelihood function is constructed from the HME approximations, instead of a pdf from the true family $\mathbf{\Phi}$. Note that Condition 2 is hard to check in practice, although it looks plausible if a sufficiently large scope of ML search is used.

The next theorem states that the maximum likelihood method based on the HME of GLM1 models is consistent in estimating the mean functions in $\psi(W^{\infty}_{2;K_0})$.

Theorem 4 *(Consistency of the Maximum Likelihood Method). Let* $(\mathrm{MSE})_{n,\Lambda}$ *be as defined in (14). Under regularity conditions* $A_{\mathcal{S},2}$ *and 2,*

$$\lim_{m\to\infty} \limsup_{n\to\infty} \inf_{\Lambda\in\mathcal{S}\cap\mathcal{J}_m} (\mathrm{MSE})_{n,\Lambda} = 0.$$

Here $s = \dim(\mathbf{x})$*, n is the sample size,* $m = \sup_{\Lambda\in\mathcal{S}\cap\mathcal{J}_m}\{\mathrm{card}(\Lambda)\}$*, and* $\mathcal{J}_m = \{\Lambda : \Lambda \in \mathcal{J}, \mathrm{card}(\Lambda) \le m\}$ *is the set of all HME structures containing no more than m experts. Actually*

$$\limsup_{n\to\infty} \inf_{\Lambda\in\mathcal{S}\cap\mathcal{J}_m} (\mathrm{MSE})_{n,\Lambda} \le \frac{c}{m^{4/s}},$$

where c is a positive constant independent of n, m and the structure Λ*.*

The proof of this theorem starts with the decomposition of the MSE into two parts. One part characterizes the discrepancy between the *estimated mean function* and its *large sample limit*, which can be shown to converge to zero as the sample size increases. The second part comes from the discrepancy between the *large sample limit* of the estimated mean function and the *true mean function*, which is bounded by an approximation error by applying Lemmas 1 and 2, and converges to zero as the number of experts m increases. Combining these two parts leads to the proof for consistency. The details are included elsewhere.

4 Unknown Shape Parameter

Up to now, we have been assuming that the shape parameter u of a GLM1 expert is known, or fixed at a value which is equal to the shape parameter in the true pdf φ. An example of the shape parameter is $u = 1/\sigma^2$ for a normal expert. Now suppose the shape parameter is unknown and needs to be estimated also. We assume that the parameter space U of u is a compact subset of the positive real line. Lemma 1 (c) needs a little modification. A bound of the KL distance now requires an additional term proportional to the discrepancy between the true shape parameter in φ and the "proposed" shape parameter in f. Condition 2 needs to be modified. In addition to the condition on the scope $\tilde{B}_\Lambda$ for θ, we assume that the scope of the ML search in the u direction contains the true shape parameter. Using techniques similar to before, it is straightforward to show that, with this modification on Condition 2, all theorems on denseness and consistency are still valid.

5 Discussion

We investigated the power of the HME networks of one parameter exponential family regression models with generalized linear mean functions (GLM1 experts) in terms of approximating a certain class of relatively arbitrary density functions, namely, the density functions of one-parameter exponential family regression models with conditional mean functions belonging to a transformed Sobolev class. We demonstrated that the approximation rate of the HME *mean* functions is of order $O(m^{-2/s})$ in L_p norm. We also showed that the approximation rate of HME *density* functions is of order $O(m^{-2/s})$ in Hellinger distance, and of order $O(m^{-4/s})$ in KL divergence. Here s is the dimension of the predictor, and m is the maximal number of experts in the network. We also showed that the maximum likelihood (ML) approach, which is associated with some optimal statistical properties and a convenient maximization algorithm, is consistent in estimating the mean response from data, as the sample size and the number of experts both increase. Moreover, the approximation rates and the consistency result can be achieved within the family of HME structures with binary trees, or within the family of HME structures with one layer of experts (the MEs). We do not claim that the approximation rates obtained in this paper is optimal. In fact, for the special case of mixing linear model experts in a single layer, Zeevi *et al.* (1998) have shown that a better rate for approximation of mean functions can be achieved if higher-than second-order continuous differentiability of the target functions is assumed. Our work is different from Zeevi *et al.* (1998) regarding the following aspects: (i) We deal with mixtures of *generalized* linear models instead of the mixtures of ordinary linear models. (ii) We consider the set-up of the HME networks instead of the single-layer mixtures of experts. (iii) We consider the maximum likelihood method instead of the least-squares approach for model fitting. (iv) Related to the use of the maximum likelihood method, we obtained the approximation rate in terms of probability density

functions, as well as in terms of the mean response. (v) We have formulated the conditions and proofs of our results in a way that is protective of the inherent non-identifiability problems of the parameterization.

Acknowledgments

The authors wish to thank Assaf J. Zeevi for suggesting a reference for Lemma 1(b). Martin A. Tanner was supported in part by NIH Grant CA35464.

References

Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics*, Prentice-Hall, Englewood Cliffs, New Jersey.

Bishop, C .M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York.

Cacciatore, T. W. and Nowlan, S. J. (1994). Mixtures of controllers for jump linear and non-linear plants. In G. Tesauro, D.S. Touretzky, and T.K. Leen (Eds.), *Advances in Neural Informations Processing Systems 6*. Morgan Kaufmann, San Mateo, CA.

Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, Toronto, Ontario.

Haussler, D. and Opper, M. (1995). General Bounds on the Mutual Information Between a Parameter and n Conditionally Independent Observations. *Proceedings of the Eighth annual Computational Learning Theory Conference (COLT), 1995, Santa Cruz, CA*. ACM Press.

Haykin, S. (1994). *Neural Networks*. Macmillan College Publishing Company, New York.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In M.I. Jordan (Ed.), *Learning in Graphical Models*. Kluwer Academic Publishers.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp.* **3** 79-87.

Jiang, W. and Tanner, M. A. (1999). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models. *Neural Comp.* (To appear.)

Jordan, M. I., and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.* **6** 181-214.

Jordan, M. I., and Xu, L. (1995). Convergence results for the EM approach to mixtures-of-experts architectures. *Neural Networks* **8** 1409-1431.

Lehmann, E. L. (1991). *Theory of Point Estimation* Wadsworth, Monterey, CA.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Meilă, M. and Jordan, M. I. (1995). Learning fine motion by Markov mixtures of experts. A.I. Memo No. 1567, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Comp.* **8** 164-177.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *J. Amer. Statist. Assoc.* **91** 953-960.

Tipping, M. E. and Bishop, C. M. (1997). Mixtures of probabilistic principal component analysers. Technical Report NCRG-97-003, Department of Computer Science and Applied Mathematics, Aston University, Birmingham, UK.

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

Zeevi, A. and Meir, R. (1997). Density Estimation Thru Convex Combinations; Approximation and Estimation Bounds. *Neural Networks* **10** 99-106.

Zeevi, A., Meir, R., and Maiorov, V. (1998). Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Trans. Information Theory* **44**, 1010-1025.