The Annals of Statistics
1999, Vol. 27, No. 3, 987–1011

HIERARCHICAL MIXTURES-OF-EXPERTS FOR EXPONENTIAL FAMILY REGRESSION MODELS: APPROXIMATION AND MAXIMUM LIKELIHOOD ESTIMATION

BY WENXIN JIANG AND MARTIN A. TANNER[1]

Northwestern University

We consider 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 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, where m is the number of experts, and s is the dimension of the predictor $\mathbf{x}$. We also provide conditions under which the mean-square error of the estimated mean response obtained from the maximum likelihood method converges to zero, as the sample size and the number of experts both increase.

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), have received considerable attention due to flexibility in modeling, appealing interpretation and the availability of convenient computational algorithms. In contrast to the single-layer ME model, the HME model has a tree-structure and can summarize the data at multiple scales of resolution due to its use of nested predictor regions. By the way they are constructed, ME and HME models are natural tools for likelihood-based inference using the expectation maximization (EM) algorithm [Jordan and Jacobs (1994) and Jordan and Xu (1995)], as well as for Bayesian analysis based on data augmentation [Peng, Jacobs and Tanner (1996)]. 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

Received March 1998; revised March 1999.
[1]Supported in part by NIH Grant CA35464.
AMS 1991 *subject classifications*. Primary 62G07; secondary 41A25.
Key words and phrases. Approximation rate, exponential family, generalized linear models, Hellinger distance, Hierarchical mixtures-of-experts, Kullback–Leibler divergence, maximum likelihood estimation, mean square error.

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 the denseness and consistency of these models in the generalized linear model context. 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 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 can be parameterized 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 a link function. The inverse link function $\psi(\cdot)$ is often 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 nonnegative. In general, the GLM1 probability density function (pdf) of y conditional on $\mathbf{x}$ is parameterized by the conditional mean $\mu(\mathbf{x})$, 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 Remark 3 (at the end of Section 3), we will discuss the situation when the dispersion parameter is also estimated.

A mixtures-of-experts model assumes that the total conditional density of the response is a local mixture of the conditional densities 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 mixing 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 $\mu(\mathbf{x}) = \sum_J g_J(\mathbf{x})\mu_J(\mathbf{x})$. 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 probabilistic decision tree, where the gating networks determine the branching probabilities as a function of the covariate $\mathbf{x}$, and the experts are identified with the leaves of the decision tree. A simple mixtures-of-experts model takes the expert label J to be an integer. An HME model takes J as an integer vector, with dimension equal to the number of layers in the HME decision tree.

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; g_i and $g_{j|i}$ $(i, j \in \{1, 2\})$ are logistic-type local weights [see Section 2.3 and (2.6)] which are identified with the probabilities of decisions at the branches of the tree. Note that the product $g_i g_{j|i}$ gives a probability $g_J(\mathbf{x}) = g_i g_{j|i}$ for the path $J = (i, j)$. At the leaves of the tree, each expert $J = (i, j)$ proposes a conditional density, $p_J = p_{ij}$, say, with a corresponding conditional mean μ_{ij}. Summing over all path J's gives the total conditional density of the conditional mixture model: $p = \sum_J g_J p_J$. The corresponding summation of the conditional mean functions, $\mu = \sum_J g_J \mu_J$, is presented in a recursive way in the original article of Jordan and Jacobs (1994), by $\mu = \sum_{i=1}^{2} g_i \mu_i$ and $\mu_i = \sum_{j=1}^{2} g_{j|i} \mu_{ij}$, and is illustrated in Figure 1.

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. An interesting proposition is to extend this result to HME for GLM1s with nonlinear link functions and to consider the consistency of

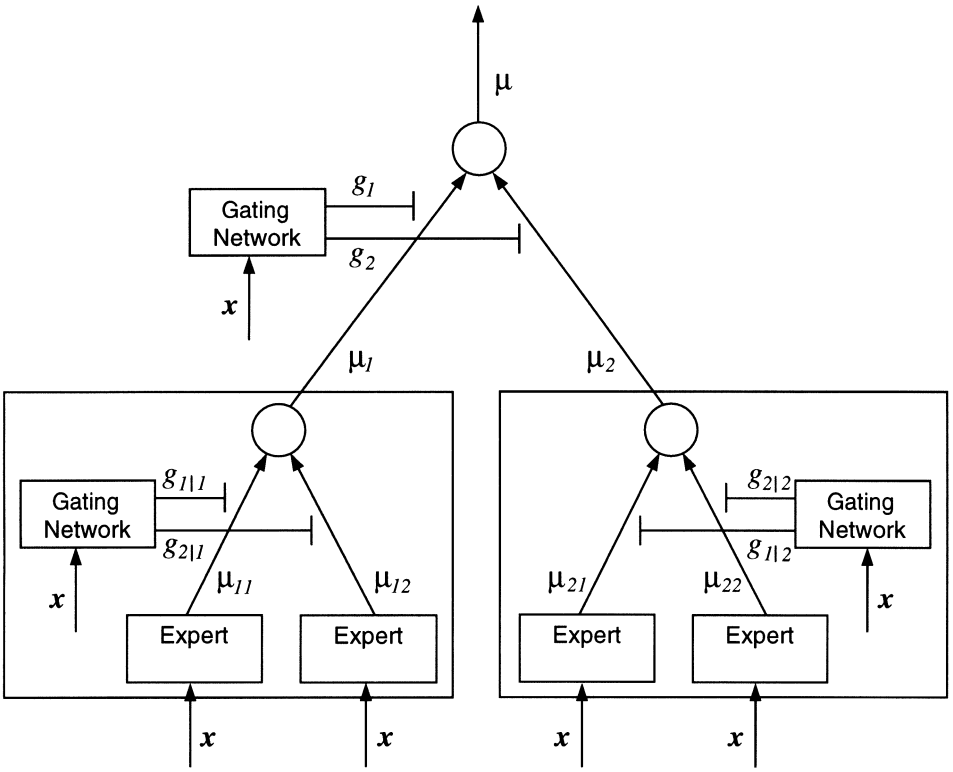

FIG. 1. *A two-layer hierarchical mixtures-of-experts model.*

maximum likelihood estimation. Jiang and Tanner (1999) show that the HME for generalized linear *mean* functions can be used to approximate arbitrary smooth functions in a transformed Sobolev class. The present paper, in contrast, focuses on the denseness of HME *density functions* and on the *consistency* of the maximum likelihood learning. The maximum likelihood (ML) approach has two advantages over the conventional least-squares approach. (1) The maximum likelihood approach gives the smallest asymptotic variance for the estimator of the mean response, in the case of correct model specification. (2) 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. (1) The maximum likelihood method deals with density functions rather than with mean functions. A result on the denseness of mean functions, such as the ones stated in Zeevi, Meir and Maiorov (1998) and Jiang and Tanner (1999), 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. (2) 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 generalized linear mean function (GLM1).

2.1. *GLM1.* We first describe the one-parameter exponential family in a way that best fits the purpose of this paper. Let $(A, \mathscr{F}_A, \lambda)$ 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

$$\pi(h, y) = \exp\{a(h)y + b(h) + c(y)\} \quad \text{for } y \in A, \tag{2.1}$$

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-$\mathscr{F}_A$.

We list some well-known properties of the one-parameter exponential models, which will be useful later.

1. The moment generating function exists in some neighborhood of the origin, and thus moments of all orders exist. See Lehmann (1991), Theorem 1.4.2, page 31.

2. For each positive integer k, $\mu_{(k)}(h) = \int_A y^k \pi(h, y)\, d\lambda$ is differentiable in h up to any order, due to the analyticity of a, b and Theorem 1.4.1 of Lehmann (1991), page 29. In particular, we denote $\mu_{(1)}(h) = \psi(h) = \mu$ and $\mu_{(2)}(h) = v(h)$ as the first two moments.
3. 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. ψ: $\Re \mapsto \psi(\Re)$ forms a $\mathscr{C}^\infty$-diffeomorphism.

Note that the parameterization of $a(h)$, $b(h)$ and $c(h)$ is not unique. For our purpose we require these functions to be defined on the entire real line.

Some examples follow.

Poisson. $\mathscr{P}(\mu)$ 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

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

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

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 $\pi(h, y)$. In this context, the inverse of $\psi(\cdot)$ is called a link function [McCullagh and Nelder (1989)]. For a specific probability model of y (say, Poisson), there could be different choices of the link function. Our paper does not restrict the choice of the link function, as long as it is "smooth" on $\Re$ and invertible. Note that the functions a, b and c in (2.1) 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 GLM1 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 $\Omega = [0, 1]^s = \bigotimes_{q=1}^s [0, 1]$, 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 $(A, \mathscr{F}_A, \lambda)$ be a general measure space, $(\Omega, \mathscr{F}_\Omega, \kappa)$ be a probability space such that κ has a positive continuous density with respect to the Lebesgue measure on Ω and $(\Omega\otimes A, \mathscr{F}_\Omega\otimes\mathscr{F}_A, \kappa\otimes\lambda)$ 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

$$\varphi(\mathbf{x}, y) = \pi(h(\mathbf{x}), y). \tag{2.2}$$

Here $\pi(\cdot,\cdot)$: $\Re \otimes A \mapsto \Re$ has the one-parameter exponential form (2.1). In contrast to a GLM1 model, we allow a more flexible $h(\mathbf{x})$ in (2.2). 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 $D^{\mathbf{k}}h \equiv (\partial^{|\mathbf{k}|}h/\partial x_1^{k_1}\cdots\partial x_s^{k_s})$. In other words, $h \in W_{2;K_0}^{\infty}$, where $W_{2;K_0}^{\infty}$ 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

$$\mu(\mathbf{x}) = \int_A y\varphi(\mathbf{x}, y)\, d\lambda(y) = \psi(h(\mathbf{x})) \tag{2.3}$$

for all $\mathbf{x}$ in Ω. Sobolev classes of mean functions similar to $W_{2;K_0}^{\infty}$ are also considered in Mhaskar (1996) and Zeevi, Meir and Maiorov (1998). Our family of mean functions is a transformed class $\psi(W_{2;K_0}^{\infty})$, 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 proof. We also note that in the situation when Ω is the direct product of s closed intervals, suitable recentering and rescaling 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 $\mathbf{\Phi}$. This is the set of target functions that we will consider 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 $\mathbf{\Phi}$.

2.3. *The family of HME of GLM1s.* An approximator f in the HME family is assumed to have the following form:

$$f = f_\Lambda(\mathbf{x}, y; \theta) = \sum_{J\in\Lambda} g_J(\mathbf{x}; \mathbf{v})\pi(h_J(\mathbf{x}),\ y), \tag{2.4}$$

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 $\alpha_J \in \Re$ and $\boldsymbol{\beta}_J \in \Re^s$, as well as $\mathbf{v}$ which is some parameter for the gating function g_J's. 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 (2.4), 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 $l = \dim(\Lambda)$, which is the number of layers and the cardinality $m = \mathrm{card}(\Lambda)$, which is the number of experts. A HME of l-layers has a structure of the form $\Lambda = \bigotimes_{k=1}^{l} A_k$ where $A_k = \{1, \ldots, w_k\}$, $w_k \in \mathcal{N}$ and $k = 1, \ldots, l$. (We use $\mathcal{N}$ to denote the set of all positive integers.) Graphically, $w_k = \mathrm{card}(A_k)$ represents the number of expert branches or the number of "splits" at layer k, $k = 1, \ldots, l$. Note that in this paper we restrict attention to "rectangular-shaped" structures (corresponding to balanced trees). A generic

expert label J in Λ can then be expressed as $J = (j_1, \ldots, j_l)$ 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, $\mathscr{J}$, $\mathscr{J}_m$ and $\mathscr{S}$, which will be used later when formulating the results. The set of all possible HME structures under consideration is $\mathscr{J} = \{\Lambda\colon \Lambda = \bigotimes_{k=1}^{l}\{1, \ldots, w_k\}; w_k \in \mathscr{N};\ k = 1, \ldots, l;\ l \in \mathscr{N}\}$. The set of all HME structures containing no more than m experts is denoted as $\mathscr{J}_m = \{\Lambda\colon \Lambda \in \mathscr{J}, \operatorname{card}(\Lambda) \le m\}$. We also introduce a symbol $\mathscr{S}$ to denote a generic subset of $\mathscr{J}$. 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 $\mathscr{S}$ of HME structures. A trivial example of $\mathscr{S}$ is $\mathscr{J}$. Another example of $\mathscr{S}$ is $\mathscr{S}_L = \{\Lambda\colon \Lambda \in \mathscr{J}, \dim(\Lambda) \le L\}$, which includes all structures with L or less layers. In particular, $\mathscr{S}_1$ represents the set of single-layer structures. A third example of $\mathscr{S}$ is $\mathscr{S}_B = \{\Lambda\colon \Lambda = \bigotimes_{k=1}^{l}\{1, 2\};\ l \in \mathscr{N}\}$, which represent 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 $\mathscr{G} = \{G_{\mathbf{v},\Lambda}\colon \mathbf{v} \in V_\Lambda, \Lambda \in \mathscr{S}\}$ be the family of gating vectors defined on the set of structures $\mathscr{S}$, which will be referred to as a *gating class* defined on $\mathscr{S}$.

In the following we define the *logistic gating class* $\mathscr{G} = \mathscr{L}$ on the set of all structures $\mathscr{J}$. This class has been commonly used in literature [see Jordan and Jacobs (1994)]. Here, for each structure Λ in $\mathscr{J}$ 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 $(j_1, j_2, \ldots, j_l)$. Then,

$$g_J \equiv g_{j_1 j_2 \ldots j_l} = g_{j_1} g_{j_2|j_1} \cdots g_{j_l|j_1 j_2 \ldots j_{l-1}}. \tag{2.5}$$

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

$$g_{j_q|j_1\ldots j_{q-1}} = \frac{\exp(\xi_{j_q|j_1\ldots j_{q-1}})}{\sum_{k=1}^{w_q}\exp(\xi_{k|j_1\ldots j_{q-1}})}, \tag{2.6}$$

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

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

since otherwise a "translation" of the parameters,

$$\begin{aligned}
&\phi_{k|j_1\ldots j_{q-1}} \to \phi_{k|j_1\ldots j_{q-1}} + \phi_0,\\
&\gamma_{k|j_1\ldots j_{q-1}} \to \gamma_{k|j_1\ldots j_{q-1}} + \gamma_0 \quad \text{all } k = 1, \ldots, w_q,
\end{aligned}$$

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 $(\phi_{j_q|j_1\ldots j_{q-1}}, \gamma^T_{j_q|j_1\ldots j_{q-1}})$, where $(j_1, \ldots, j_{q-1}) \in \bigotimes_{r=1}^{q-1}\{1, \ldots, w_r\}$ and $j_q \in \{1, \ldots, w_q - 1\}$ for all $q = 1, \ldots, l$. It is easy to see that

$$\begin{aligned}\dim(\mathbf{v}) &= (s+1)\big\{(w_1-1) + w_1(w_2-1) + \ldots + w_1\ldots w_{l-1}(w_l-1)\big\}\\ &= (s+1)(m-1),\end{aligned}$$

and the parameter space V_Λ for $\mathbf{v}$ is $\Re^{(s+1)(m-1)}$, where $m = w_1\cdots w_l = \text{card}(\Lambda)$. 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 $\Re^{(s+1)m}$. The space of grand parameter θ's, corresponding to structure Λ, is $\tilde{\Theta}_\Lambda = \Re^{(s+1)(2m-1)}$. 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 (2.4), specific to a structure Λ, which can be denoted as $\Pi_\Lambda = \{f_\Lambda(\cdot, \cdot; \theta)\colon \theta \in \tilde{\Theta}_\Lambda\}$. 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 [see Fritsch, Finke and Waibel (1997) for an adaptive approach]. We consider in this paper choosing among the set of structures $\mathscr{J}_m \cap \mathscr{S}$. Denote

$$\mathbf{\Pi}_{m,\mathscr{S}} = \big\{f\colon f \in \Pi_\Lambda; \Lambda \in \mathscr{J}_m \cap \mathscr{S}\big\}. \tag{2.7}$$

This set, $\mathbf{\Pi}_{m,\mathscr{S}}$, 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 $\mathscr{S}$ of HME structures, which will be specified later. We do not explicitly require that $\mathbf{\Pi}_{m,\mathscr{S}}$ 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

$$\mu_\Lambda(\mathbf{x}; \theta) = \int_A y f_\Lambda(\mathbf{x}, y; \theta)\, d\lambda(y) = \sum_{J\in\Lambda} g_J(\mathbf{x}; \mathbf{v})\psi(\alpha_J + \mathbf{x}^T\boldsymbol{\beta}_J), \tag{2.8}$$

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 $\mathbf{\Pi}_{m,\mathscr{S}}$. 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, that is, 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 nonidentifiability of some components of 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 $(\Omega, \mathscr{F}_\Omega, \kappa)$ and $(A, \mathscr{F}_A, \lambda)$ be as introduced in Section 2.2. 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 (2.2)] with respect to the measure λ, for all $\mathbf{x} \in \Omega$.

The log-likelihood function based on the HME model is

$$\text{Ł}_{n,\Lambda}(\theta; \omega) = n^{-1} \sum_{i=1}^{n} \log\{f_\Lambda(\mathbf{X}_i, Y_i; \theta)/\varphi_0(\mathbf{X}_i, Y_i)\}, \tag{2.9}$$

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 $\varphi_0(\mathbf{X}_i, Y_i) = \exp(c(Y_i))$ where $c(\cdot)$ is from (2.1). 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 $\tilde{B}_\Lambda \subset \tilde{\Theta}_\Lambda$, that is,

$$\hat{\theta}_{n,\Lambda}(\omega) = \arg\max_{\theta \in \tilde{B}_\Lambda} \{\text{Ł}_{n,\Lambda}(\theta; \omega)\}. \tag{2.10}$$

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

$$\text{KL}(f, g) \equiv \int\int_{\Omega \otimes A} g(\mathbf{x}, y) \log\left\{\frac{g(\mathbf{x}, y)}{f(\mathbf{x}, y)}\right\} d\kappa(\mathbf{x})\, d\lambda(y). \tag{2.11}$$

It turns out that the KL divergence $\text{KL}(f_\Lambda, \varphi)$ is always well defined (see Corollary 1 in Section 2.5). Due to the nonidentifiability 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

$$\Theta_\Lambda = \{\theta \in \tilde{B}_\Lambda \colon \theta = \arg\min_{\theta^* \in \tilde{B}_\Lambda} \text{KL}(f_\Lambda(\cdot, \cdot; \theta^*), \varphi)\}. \tag{2.12}$$

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

$$(\text{MSE})_{n,\Lambda} = \mathrm{E}\int\{\mu_\Lambda(\mathbf{x};\hat{\theta}_{n,\Lambda}) - \mu(\mathbf{x})\}^2\, d\kappa(\mathbf{x}), \tag{2.13}$$

where E is the expectation taken on the MLE $\hat{\theta}_{n,\Lambda}$, and μ_Λ and μ are defined in (2.8) and (2.3), 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 $\mathbf{Q}^{(\nu)} = \{Q_J^{(\nu)}\}_{J\in\Lambda^{(\nu)}}$, $\Lambda^{(\nu)} \in \mathscr{J}$, 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 $p_\nu = \text{card}(\Lambda^{(\nu)})$, $(p_\nu \in \mathscr{N})$.

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 $\{\mathbf{Q}^{(\nu)}\colon \nu = 1, 2, \ldots\}$ is called a sequence of *fine partitions* with *structure sequence* $\{\Lambda^{(\nu)}\}$, *cardinality sequence* $\{p_\nu\}$ and *bounding constant* c_0.

DEFINITION 2 (Subgeometric). A sequence $\{a_\nu\}$ is *subgeometric with rate bounded by* M_2, if $a_\nu \in \mathscr{N}$, $a_\nu \to \infty$ 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 discrepancy between a pdf f in Π_Λ [of the form (2.4)] and a pdf φ in $\mathbf{\Phi}$ [of the form (2.2)]. One of them is the KL distance $\text{KL}(f, \varphi)$ [see (2.11)]. Another is the Hellinger distance

$$d_H(f,\varphi) = \left\{\int\int\left(\sqrt{f} - \sqrt{\varphi}\right)^2 d\lambda\, d\kappa\right\}^{1/2}. \tag{2.14}$$

This is a true distance and is invariant under rescaling of the measures λ and κ [see Devroye and Gyoerfi (1985)]. A third description is the L_2 distance between the means,

$$d_2(\mu_f, \mu_\varphi) = \left\|\mu_\Lambda(\cdot,\theta) - \mu(\cdot)\right\|_{2,\kappa} = \left\{\int(\mu_f - \mu_\varphi)^2\, d\kappa\right\}^{1/2}, \tag{2.15}$$

where $\mu_f = \int yf\, d\lambda$ 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.

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

$$\mathscr{D}(f,\varphi) = \int \sum_{J\in\Lambda} g_J(\mathbf{x};\mathbf{v})\{h_J(\mathbf{x}) - h(\mathbf{x})\}^2\, d\kappa, \tag{2.16}$$

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 the quality of this softly partitioned linear approximation. The name "upper divergence" is due to the following lemma, which implies that $\mathscr{D}$ is stronger than the other divergence measures, that is, KL, d_H and $d_2(\mu_f, \mu_\varphi)$ (the proof is in Section 4).

LEMMA 1 (Strength of divergence measures). *Consider any structure* Λ, *any HME density* $f = \sum_{J\in\Lambda} g_J \pi(h_J, \cdot)$ *in* Π_Λ *and any target density* $\varphi = \pi(h, \cdot)$ *in* $\mathbf{\Phi}$. *We have*:

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

(b) $d_H^2(f, \varphi) \le KL(f, \varphi)$. [*This lemma appeared in*, *e.g.*, *Haussler and Opper* (1995) *and Zeevi and Meir* (1997).]

(c) $\mathrm{KL}(f, \varphi) \le M_{II} \mathscr{D}(f, \varphi)$.

Here, $M_I = \sup_{|h|\le K}\{\int y^2 \pi(h, y)\, d\lambda\}$, *and*

$$M_{II} = \tfrac{1}{2}\left\{ \sup_{|h|\le K} \left| \int y\pi(h, y)\, d\lambda \right| \sup_{|h|\le K} |a''(h)| + \sup_{|h|\le K} |b''(h)| \right\},$$

where $a(\cdot)$ *and* $b(\cdot)$ *are defined as in* (2.1), *and* K *is an upperbound for* $|h_J(\mathbf{x})|$ *and* $|h(\mathbf{x})|$ *for all* $\mathbf{x}$ *in* Ω *and all* J *in* Λ.

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* $\mathscr{D}$ *are finite*.

The proof is obvious from Lemma 1. Note that $\mathscr{D}$ is finite, since it involves an integration of a continuous function over the compact space Ω of input $\mathbf{x}$.

In Section 4, we will use Lemma 1(c) on upper divergence to prove the denseness property of the HME densities in KL divergence.

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

CONDITION 1. $(A_{\mathscr{S}, p})$. For a subset $\mathscr{S} \subset \mathscr{J}$, there is a fine partition sequence $\{\{Q_J^{(\nu)}\}_{J\in\Lambda_0^{(\nu)}}: \Lambda_0^{(\nu)} \in \mathscr{S},\ \nu = 1, 2, \ldots\}$ with a bounding constant c_0 and a cardinality sequence $\{p_\nu: \nu = 1, 2, \ldots\}$, such that $\{p_\nu^{1/s}\}$ is subgeometric 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

$$G_{\mathbf{v}_\varepsilon, \Lambda_0^{(\nu)}} = \{g_J(\mathbf{x}; \mathbf{v}_\varepsilon)\}_{J\in\Lambda_0^{(\nu)}} \in \mathscr{G},$$

$$\Lambda_0^{(\nu)} \in \mathscr{S}, \text{ such that } \sup_{J\in\Lambda_0^{(\nu)}} \|g_J(\cdot; \mathbf{v}_\varepsilon) - \chi_{Q_J^{(\nu)}}(\cdot)\|_{p,\sigma} \le \varepsilon. \tag{3.1}$$

Here, $\|f(\cdot)\|_{p,\sigma} \equiv \{\int_\Omega |f(\mathbf{x})|^p \, d\sigma(\mathbf{x})\}^{1/p}$, 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, that is, $\chi_B(\mathbf{x}) = 1$ 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.

THEOREM 1 (Approximation rate in Hellinger distance). *Under Condition* $A_{\mathcal{S},1}$,

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

for some positive constant c independent of m. Here d_H *is the Hellinger distance defined in* (2.14).

REMARK 2. This theorem implies the same approximation rate in the L_1 distance $d_1(f, \varphi) = \int\int |f - \varphi| \, d\lambda \, d\kappa$, by the well-known result $d_1(f, \varphi) \leq 2d_H(f, \varphi)$ [see Devroye and Gyoerfi (1985)]. The same rate in general L_p distances ($p \geq 2$) is derived in Jiang and Tanner (1998) (under Condition $A_{\mathcal{S},p}$), where an extra boundedness condition is required.

THEOREM 2 (Approximation rate in KL divergence). *Under Condition* $A_{\mathcal{S},1}$,

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

for some positive constant c independent of m. Here KL *is the KL divergence defined in* (2.11).

All these results depend on a major condition $A_{\mathcal{S},1}$. The following remark claims that it is satisfied by certain gating functions defined on certain structures.

REMARKS. (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 $(1 + \exp(-\beta(z - z_0)))^{-1}$, which can approximate a step function $H(z - z_0)$ as β increases, for any location of jump z_0. The gating functions in a binary tree involve 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), Lemma, Section 3, show that Condition $A_{\mathscr{L},p}$ is satisfied (for any $p \in \mathscr{N}$) by the logistic gating class $\mathscr{L}$ defined on $\mathscr{S}_s$, which is the set of structures with no more than s layers, where $s = \dim(\mathbf{x})$. Moreover, due to Jiang and Tanner (1999), Remark (i), Section 5, Condition $A_{\mathscr{L},p}$ is also satisfied by the logistic gating class defined on the set of single-layer structures $\mathscr{S}_1$ (corresponding to the MEs), which implies that the approximation rates in the above theorems apply to ME of GLM1s.

(c) Another class of gating functions can be defined only on the binary trees (in $\mathscr{S}_B$). There, the logistic gating functions in (2.6) are replaced by continuous cumulative distribution functions (cdfs). One example is to use the normal cdf. Then the gating factor $g_{j_q|j_1\cdots j_{q-1}}$ of (2.6) 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 $\xi_{j_1\cdots j_{q-1}} = \phi_{j_1\cdots j_{q-1}} + \gamma_{j_1\cdots j_{q-1}}^T \mathbf{x}$. A similar argument as in part (a) of this remark shows that Condition $A_{\mathscr{L},p}$ is satisfied for this new gating class for any $p \in \mathscr{N}$.

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

CONDITION 2 (Scope of maximum likelihood searching). The scope of the maximum likelihood (ML) searching, $\tilde{B}_\Lambda$, is a compact subset of $[-C, C]^{\dim(\theta)}$ for some large positive constant C, and the scope is so large that it contains a point θ_{KL} which minimizes the KL divergence 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

$$\mathrm{KL}\big(f_\Lambda(\cdot,\cdot,\theta_{\mathrm{KL}}),\ \varphi(\cdot,\cdot)\big) = \inf_{\theta\in\tilde{\Theta}_\Lambda} \mathrm{KL}\big(f_\Lambda(\cdot,\cdot,\theta),\varphi(\cdot,\cdot)\big) = \inf_{f\in\Pi_\Lambda} \mathrm{KL}(f,\varphi).$$

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 KL 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 $\psi(W_{2;K_0}^\infty)$.

THEOREM 3 (Consistency of the maximum likelihood method). *Let* $(\mathrm{MSE})_{n,\Lambda}$ *be as defined in* (2.13). *Under regularity conditions* $A_{\mathscr{L},1}$ *and* 2,

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

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

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

where $s = \dim(\mathbf{x})$ and c is a positive constant independent of n, m and the structure Λ.

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

REMARK 3 (Unknown shape parameter). Up to now, we have been assuming that any 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 u is unknown and needs also to be estimated. We can either assume the same shape parameter for all experts, or allow it to differ. In either case, the grand parameter θ will be expanded to include additional component(s) from the u('s). We assume that the parameter space U of each component u is a compact subset of the positive real line. Lemma 1(c) requires a small modification, that is, 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(s) in f. However, using similar techniques, it is straightforward to show that all theorems on denseness and consistency are still valid.

4. Proofs and secondary lemmas. In this section we will often use the following shorthand notation: $g_J = g_J(\mathbf{x};\mathbf{v})$, $h_J = h_J(\mathbf{x})$, $\pi_J = \pi(h_J(\mathbf{x}), y)$, $f = \sum_J g_J\pi_J$ [for the HME density (2.4)], $\sum_J = \sum_{J\in\Lambda}$, $\sup_J = \sup_{J\in\Lambda}$, $h = h(\mathbf{x})$, $\varphi = \pi(h(\mathbf{x}), y)$ [for the true density (2.2)], $\mu = \mu_\varphi = \int y\varphi\, d\lambda$, $\mu_f = \int yf\, d\lambda$ and $\mu_J = \int y\pi_J\, d\lambda$.

4.1. *Proofs of Theorems* 1 *and* 2.

PROOF OF LEMMA 1. Here (b) is due to Lemma 3.2 of Zeevi and Meir (1997), which is a corollary to Lemma 5 of Haussler and Opper (1995).

The following observations prove (a):

$$\begin{aligned} d_2^2(\mu_f, \mu_\varphi) &= \int\left\{\int yf\, d\lambda - \int y\varphi\, d\lambda\right\}^2 d\kappa \\ &= \int\left\{\int y\left(\sqrt{f} - \sqrt{\varphi}\right)\left(\sqrt{f} + \sqrt{\varphi}\right) d\lambda\right\}^2 d\kappa \\ &\le \int\left\{\int |y|\left|\sqrt{f} - \sqrt{\varphi}\right|\left(\sqrt{f} + \sqrt{\varphi}\right) d\lambda\right\}^2 d\kappa \end{aligned}$$

$$\overset{\text{(i)}}{\leq} 2\int\left\{\int |y||\sqrt{f}-\sqrt{\varphi}|\left(\sqrt{f+\varphi}\right)d\lambda\right\}^2 d\kappa$$

$$\overset{\text{(ii)}}{\leq} 2\int\left[\left\{\int y^2(f+\varphi)\,d\lambda\right\}\left\{\int\left(\sqrt{f}-\sqrt{\varphi}\right)^2 d\lambda\right\}\right] d\kappa$$

$$\overset{\text{(iii)}}{\leq} 4M_I\int\int\left(\sqrt{f}-\sqrt{\varphi}\right)^2 d\lambda\, d\kappa.$$

Inequality (i) above uses $\sqrt{f}+\sqrt{\varphi}\leq\sqrt{2}\sqrt{f+\varphi}$; (ii) uses the Cauchy–Schwarz inequality. To obtain (iii), note that

$$\begin{aligned}\int y^2(f+\varphi)\,d\lambda &= \sum_J g_J\int y^2\pi_J\,d\lambda+\int y^2\varphi\,d\lambda\\ &\leq \sup_{J\in\Lambda}\sup_{\mathbf{x}\in\Omega}\left(\int y^2\pi_J\,d\lambda\right)\left(\sum_J g_J\right)+\sup_{\mathbf{x}\in\Omega}\int y^2\varphi\,d\lambda\\ &= \sup_{J\in\Lambda}\sup_{\mathbf{x}\in\Omega}\left(\int y^2\pi_J\,d\lambda\right)+\sup_{\mathbf{x}\in\Omega}\int y^2\varphi\,d\lambda\\ &= \sup_{J\in\Lambda}\sup_{\mathbf{x}\in\Omega}\left(\int y^2\pi(h_J,y)\,d\lambda\right)+\sup_{\mathbf{x}\in\Omega}\int y^2\pi(h,y)\,d\lambda\\ &\leq 2\sup_{|h|\leq K}\left\{\int y^2\pi(h,y)d\lambda\right\}=2M_I.\end{aligned}$$

The last inequality is true since K is an upperbound for $|h_J(\mathbf{x})|$ and $|h(\mathbf{x})|$ for all $\mathbf{x}$ in Ω and all J in Λ.

The following observations prove (c):

$$\begin{aligned}\mathrm{KL}(f,\varphi) &= \int\int\varphi\log(\varphi/f)\,d\lambda\,d\kappa\\ &= -\int\int\varphi\log\left[\sum_J g_J\exp\left(\{a(h_J)-a(h)\}y+\{b(h_J)-b(h)\}\right)\right]d\lambda\,d\kappa\\ &\overset{\text{(i)}}{\leq} -\int\int\varphi\sum_J g_J\left[\{a(h_J)-a(h)\}y+\{b(h_J)-b(h)\}\right]d\lambda\,d\kappa\\ &= -\int\sum_J g_J\left[\{a(h_J)-a(h)\}\mu+\{b(h_J)-b(h)\}\right]d\kappa\\ &\overset{\text{(ii)}}{=} -\tfrac{1}{2}\int\sum_J g_J\{\mu a''(h_{J*})+b''(h_{J*})\}(h_J-h)^2\,d\kappa\\ &\leq \tfrac{1}{2}\int\sum_J g_J|\mu a''(h_{J*})+b''(h_{J*})|(h_J-h)^2\,d\kappa\end{aligned}$$

$$\overset{\text{(iii)}}{=} M_{II} \int \sum_J g_J (h_J - h)^2 \, d\kappa.$$

Inequality (i) above is due to the convexity of $-\log(\cdot)$; (ii) is from a Taylor expansion for h_J around h,

$$-\{\mu a(h_J) + b(h_J)\} = -\{\mu a(h) + b(h)\} - \tfrac{1}{2}\{\mu a''(h_{J*}) + b''(h_{J*})\}(h_J - h)^2,$$

where h_{J*} is between h_J and h. [The "linear term" disappears due to property (3) in Section 2.1.] To prove inequality (iii), note that $|h_{J*}| \le \max\{|h_J|, |h|\} \le K$ since K is an upperbound for $|h_J(\mathbf{x})|$ and $|h(\mathbf{x})|$ for all $\mathbf{x}$ in Ω and all J in Λ. Note also that $\mu = \int y\pi(h, y)\, d\lambda$ where $|h| \le K$. Then

$$\begin{aligned} \left|\mu a''(h_{J*}) + b''(h_{J*})\right| &\le |\mu|\left|a''(h_{J*})\right| + \left|b''(h_{J*})\right| \\ &\le \sup_{|h|\le K}\left|\int y\pi(h, y)\, d\lambda\right| \sup_{|h|\le K}|a''(h)| + \sup_{|h|\le K}|b''(h)|, \end{aligned}$$

which proves (iii). □

PROOF OF THEOREM 2. Denote $\|f(\cdot)\|_{p,\kappa} = \{\int |f(\mathbf{x})|^p \, d\kappa\}^{1/p}$ $(p > 0)$ and $\|f(\cdot)\|_\infty = \sup_{\mathbf{x}\in\Omega} |f(\mathbf{x})|$. Consider a fine partition sequence $\{\{Q_J^{(\nu)}\}_{J\in\Lambda_0^{(\nu)}}: \Lambda_0^{(\nu)} \in \mathscr{S}, \nu = 1, 2, \ldots\}$ in Condition $A_{\mathscr{S},1}$, with a cardinality sequence $\{p_\nu\}$. We first show that for each $\nu \in \{1, 2, \ldots, \}$, for any $h \in W^\infty_{2;K_0}$,

$$(\S) \equiv \left\| \sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}}(\cdot)\hat{h}_J(\cdot) - h(\cdot)\right\|_{2,\kappa} \le \frac{c_1}{p_\nu^{2/s}}, \tag{4.1}$$

for some finite constant $c_1 > 0$, where, for each $\mathbf{x} \in \Omega$ and each $J \in \Lambda_0^{(\nu)}$,

$$\hat{h}_J(\mathbf{x}) \equiv \hat{\alpha}_J + \mathbf{x}^T\hat{\boldsymbol{\beta}}_J \equiv \{h(\boldsymbol{\xi}_J) - \boldsymbol{\xi}_J^T\nabla h(\boldsymbol{\xi}_J)\} + \mathbf{x}^T\nabla h(\boldsymbol{\xi}_J), \tag{4.2}$$

$\boldsymbol{\xi}_J$ being some point in the interior of $Q_J^{(\nu)}$. Here ∇h is the $s \times 1$ gradient column vector of a scalar function h. For any $h \in W^\infty_{2;K_0}$ and any $\hat{h}_J$ defined in (4.2), it is obvious that $\|h\|_\infty \le K_0$ and $\|\hat{h}_J\|_\infty \le K_0$.

To prove (4.1) note that

$$\begin{aligned} (\S) &= \left\| \sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}}(\cdot)\{\hat{h}_J(\cdot) - h(\cdot)\}\right\|_{2,\kappa} \\ &\le \|1\|_{2,\kappa} \sup_{J\in\Lambda_0^{(\nu)}} \left\|\hat{h}_J(\cdot) - h(\cdot)\right\|_\infty = \sup_{J\in\Lambda_0^{(\nu)}} \left\|\hat{h}_J(\cdot) - h(\cdot)\right\|_\infty, \end{aligned} \tag{4.3}$$

because $\sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}}(\mathbf{x}) = 1$ and κ is a probability measure.

By a second-order Taylor expansion of $h(\mathbf{x})$ around $\boldsymbol{\xi}_J$ and the definition of $\hat{h}_J(\mathbf{x})$ in (4.2), we have, for all $\mathbf{x} \in Q_J^{(\nu)}$, $J \in \Lambda_0^{(\nu)}$,

$$|\hat{h}_J(\mathbf{x}) - h(\mathbf{x})| \le \frac{1}{2}\Bigg(\sum_{|\mathbf{k}|=2} \|D^{\mathbf{k}} h\|_\infty\Bigg)\{\rho(\mathbf{x}, \boldsymbol{\xi}_J)\}^2 \le \frac{1}{2} K_0 \frac{c_0^2}{p_\nu^{2/s}}, \tag{4.4}$$

where $\rho(\mathbf{x}, \boldsymbol{\xi}_J) = \max_{1\le q\le s} |(\mathbf{x}-\boldsymbol{\xi}_J)_q|$, c_0 is the bounding constant from Definition 1 and Condition $A_{\mathscr{P},1}$. The latter inequality holds since (1) $\mathbf{x}$, $\boldsymbol{\xi}_J \in Q_J^{(\nu)}$, leading to $\rho(\mathbf{x}, \boldsymbol{\xi}_J) \le c_0/p_\nu^{1/s}$ by Condition $A_{\mathscr{P},1}$ and the definition of a "fine partition"; (2) $h \in W^\infty_{2;K_0}$. Then, (4.4) and (4.3) lead to (4.1) (take $c_1 = \frac{1}{2}K_0 c_0^2$).

Now, by Condition $A_{\mathscr{P},1}$, for all $\varepsilon > 0$, there exists $\mathbf{v}_\varepsilon \in V_{\Lambda_0^{(\nu)}}$ such that (3.1) holds for the norm $\|(\cdot)\|_{1,\sigma}$. This implies that

$$\sup_{J\in\Lambda_0^{(\nu)}} \|g_J(\cdot; \mathbf{v}_\varepsilon) - \chi_{Q_J^{(\nu)}}(\cdot)\|_{1,\kappa} \le \left\|\frac{d\kappa}{d\sigma}\right\|_\infty \varepsilon, \tag{4.5}$$

where $\|d\kappa/d\sigma\|_\infty = \sup_{\mathbf{x}\in\Omega} |(d\kappa/d\sigma_0)/(d\sigma/d\sigma_0)| < \infty$, since both σ and κ have positive continuous densities with respect to the Lebesgue measure σ_0. Now consider

$$\begin{aligned}(*) &\equiv \Bigg\| \sum_{J\in\Lambda_0^{(\nu)}} g_J(\cdot; \mathbf{v}_\varepsilon)\{\hat{h}_J(\cdot) - h(\cdot)\}^2 \Bigg\|_{1,\kappa} \\ &\le \underbrace{\Bigg\| \sum_{J\in\Lambda_0^{(\nu)}} \{g_J(\cdot; \mathbf{v}_\varepsilon) - \chi_{Q_J^{(\nu)}}(\cdot)\}\{\hat{h}_J(\cdot) - h(\cdot)\}^2 \Bigg\|_{1,\kappa}}_{(\dagger)} \\ &\quad + \underbrace{\Bigg\| \sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}}(\cdot)\{\hat{h}_J(\cdot) - h(\cdot)\}^2 \Bigg\|_{1,\kappa}}_{(\ddagger)},\end{aligned}$$

due to the triangular inequality.

Note that

$$\begin{aligned}(\ddagger) &= \int \sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}}(\cdot)\{\hat{h}_J(\cdot) - h(\cdot)\}^2 \, d\kappa \\ &\overset{\text{(i)}}{=} \int \Bigg\{ \sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}}(\cdot)\hat{h}_J(\cdot) - h(\cdot)\Bigg\}^2 d\kappa \overset{\text{(ii)}}{\le} \frac{c_1^2}{p_\nu^{4/s}}.\end{aligned}$$

Here (ii) follows from (4.1), and (i) uses the properties $\sum_{J\in\Lambda_0^{(\nu)}} \chi_{Q_J^{(\nu)}} = 1$ and $\chi_{Q_J^{(\nu)}}\chi_{Q_I^{(\nu)}} = \chi_{Q_J^{(\nu)}}\delta_{IJ}$, where $\delta_{IJ} = 1$ if $I = J$, and 0 otherwise.

Note also that

$$(\dagger) \overset{\text{(iii)}}{\leq} \sum_{J\in\Lambda_0^{(\nu)}} \left\| \{g_J(\cdot;\mathbf{v}_\varepsilon) - \chi_{Q_J^{(\nu)}}(\cdot)\}\{\hat{h}_J(\cdot) - h(\cdot)\}^2 \right\|_{1,\kappa}$$

$$\leq \left\{ \sup_{J\in\Lambda_0^{(\nu)}} \left\| \hat{h}_J(\cdot) - h(\cdot) \right\|_\infty \right\}^2 \sum_{J\in\Lambda_0^{(\nu)}} \left\| g_J(\cdot;\mathbf{v}_\varepsilon) - \chi_{Q_J^{(\nu)}}(\cdot) \right\|_{1,\kappa}$$

$$\overset{\text{(iv)}}{\leq} \left(\frac{c_1}{p_\nu^{2/s}}\right)^2 p_\nu \left\| \frac{d\kappa}{d\sigma} \right\|_\infty \varepsilon.$$

Here (iii) is due to the triangular inequality, and (iv) is due to (4.3)–(4.5).

Combining the results for (†) and (‡), we get

$$(*) \leq \left(\frac{c_1}{p_\nu^{2/s}}\right)^2 p_\nu \left\| \frac{d\kappa}{d\sigma} \right\|_\infty \varepsilon + \left(\frac{c_1}{p_\nu^{2/s}}\right)^2. \tag{4.6}$$

Denote $\varphi = \pi(h(\cdot), y)$ and $\hat{f} = \sum_{J\in\Lambda_0^{(\nu)}} g_J(\cdot, \mathbf{v}_\varepsilon)\pi(\hat{h}_J(\cdot), y)$. Then obviously, $\hat{f} \in \mathbf{\Pi}_{p_\nu,\mathscr{S}}$ and $\varphi \in \mathbf{\Phi}$, and $(*) = \mathscr{D}(\hat{f}, \varphi)$. Using Lemma 1(c), we have

$$\mathrm{KL}(\hat{f}, \varphi) \leq M_{II}\left(\frac{c_1}{p_\nu^{2/s}}\right)^2 p_\nu \left\| \frac{d\kappa}{d\sigma} \right\|_\infty \varepsilon + M_{II}\left(\frac{c_1}{p_\nu^{2/s}}\right)^2. \tag{4.7}$$

The upperbound K in M_{II} can be taken as K_0, since both $|h|$ and $|\hat{h}_J|$ are bounded above by K_0. This makes M_{II} to be a constant that does not depend on φ.

By (4.7) and the arbitrariness of ε, we have

$$\inf_{f\in\mathbf{\Pi}_{p_\nu,\mathscr{S}}} \mathrm{KL}(f, \varphi) \leq \frac{M_{II}c_1^2}{p_\nu^{4/s}}. \tag{4.8}$$

Since $\{p_\nu^{1/s}\}$ is subgeometric, for all $m \in \mathscr{N}$, there exists p_ν, such that $p_\nu \leq m < p_{\nu+1}$, and

$$1/p_\nu^{4/s} \geq 1/m^{4/s} > 1/p_{\nu+1}^{4/s}. \tag{4.9}$$

By the definition in (2.7), $\mathbf{\Pi}_{m,\mathscr{S}}$ is monotone nondecreasing in m, and hence

$$\mathbf{\Pi}_{p_\nu,\mathscr{S}} \subset \mathbf{\Pi}_{m,\mathscr{S}} \subset \mathbf{\Pi}_{p_{\nu+1},\mathscr{S}}.$$

Hence, for all $m \in \mathscr{N}$,

$$\inf_{f\in\mathbf{\Pi}_{m,\mathscr{S}}} \mathrm{KL}(f, \varphi) \leq \inf_{f\in\mathbf{\Pi}_{p_\nu,\mathscr{S}}} \mathrm{KL}(f, \varphi) \leq \frac{M_{II}c_1^2}{p_\nu^{4/s}} \quad \text{[by (4.8)]}$$

$$\leq \frac{M_2^4 M_{II}c_1^2}{p_{\nu+1}^{4/s}} \leq \frac{M_2^4 M_{II}c_1^2}{m^{4/s}} = c/m^{4/s},$$

by noting that $\{p_\nu^{1/s}\}$ is subgeometric with rate bounded by M_2, and using (4.9).

By construction, c does not depend on φ in $\mathbf{\Phi}$. Hence,

$$\sup_{\varphi\in\mathbf{\Phi}} \inf_{f\in\mathbf{\Pi}_{m,\mathscr{S}}} \mathrm{KL}(f,\varphi) \le c/m^{4/s} \quad \text{for all } m\in\mathscr{M}.$$

[Tracing the constant c leads to $c = (\frac{1}{2}K_0c_0^2M_2^2\sqrt{M_{II}})^2$, which does not have explicit dependence on $s = \dim(\mathbf{x})$.] □

The proof of Theorem 1 is obvious from Lemma 1(b) and Theorem 2.

4.2. *Proof of Theorem* 3. We first bound the MSE defined in (2.13) by the sum of a "stochastic" part $2S_{n\Lambda}$ and a "deterministic" part $2D_{n\Lambda}$. Then we will prove that each part goes to zero as the sample size and the number of experts increase:

$$\begin{aligned}
\mathrm{MSE}_{n,\Lambda} &\equiv \mathrm{E}\int\left\{\mu_\Lambda(\mathbf{x};\,\hat\theta_{n,\Lambda}) - \mu(\mathbf{x})\right\}^2 d\kappa \\
&\le 2\,\mathrm{E}\int\left\{\mu_\Lambda(\mathbf{x};\,\theta_{n,\Lambda}) - \mu(\mathbf{x})\right\}^2 d\kappa \\
&\quad + 2\,\mathrm{E}\int\left\{\mu_\Lambda(\mathbf{x};\,\hat\theta_{n,\Lambda}) - \mu_\Lambda(\mathbf{x};\theta_{n,\Lambda})\right\}^2 d\kappa \\
&\equiv 2D_{n,\Lambda} + 2S_{n,\Lambda}
\end{aligned} \tag{4.10}$$

since $(a+b)^2 \le 2a^2 + 2b^2$ for all $a,\ b\in\Re$. Here $\theta_{n,\Lambda}$, is a point in Θ_Λ defined in (2.12).

Note that by the definition of $\theta_{n,\Lambda}$,

$$\begin{aligned}
\theta_{n,\Lambda} &= \arg\max_{\theta\in\tilde B_\Lambda} \int\!\!\int_{\Omega\otimes A} \varphi(\mathbf{x},\ y)\log\left\{\frac{f_\Lambda(\mathbf{x},\ y;\ \theta)}{\varphi_0(\mathbf{x},\ y)}\right\} d\lambda(y)\,d\kappa(\mathbf{x}) \\
&\equiv \arg\max_{\theta\in\tilde B_\Lambda} \text{Ł}_{\infty,\Lambda}(\theta),
\end{aligned}$$

where we choose $\varphi_0(\mathbf{x},\ y) = e^{c(y)}$ [$c(\cdot)$ is from (2.1)]. There could be more than one such maximizer and the set of such maximizers is just Θ_Λ.

We will treat $\theta_{n,\Lambda}$, which is used in (4.10), as a random variable depending on sample size n, since our choice of $\theta_{n,\Lambda}$ out of the set Θ_Λ may depend on the MLE $\hat\theta_{n,\Lambda}(\omega)$ that we adopt. Note that $\theta_{n,\Lambda}$ is a minimizer of the KL divergence, when the parameter varies in the scope $\tilde B_\Lambda$. However, when Condition 2 holds, $\theta_{n,\Lambda}$ also minimizes the KL divergence over the entire parameter space $\tilde\Theta_\Lambda$, that is,

$$\mathrm{KL}\left\{f_\Lambda(\cdot,\cdot;\theta_{n,\Lambda}),\varphi(\cdot,\cdot)\right\} = \mathrm{KL}\left\{f_\Lambda(\cdot,\cdot;\theta_{\mathrm{KL}}),\varphi(\cdot,\cdot)\right\} = \inf_{f\in\Pi_\Lambda}\mathrm{KL}(f,\varphi). \tag{4.11}$$

In the following, we formulate the convergence of MLE in a setting with nonidentifiable parameterization. We first state and prove a proposition on the uniform convergence of the log-likelihood function.

PROPOSITION 1 (Uniform convergence of log-likelihood). *Let* $Ł_{n,\Lambda}$ *be the log-likelihood function defined in* (2.9) [*with* $\varphi_0(\mathbf{X}_i, Y_i) = \exp(c(Y_i))$, $c(\cdot)$ *is from* (2.1)]. *We have*

$$\sup_{\theta\in\tilde{B}_\Lambda}\left|Ł_{n,\Lambda}(\theta;\omega) - Ł_{\infty,\Lambda}(\theta)\right| \to 0$$

for almost all stochastic sequences ω, *and*

$$Ł_{\infty,\Lambda}(\theta) = \int_{\Omega\otimes A} \varphi(\mathbf{x}, y)\log\{f_\Lambda(\mathbf{x}, y;\theta)/\varphi_0(\mathbf{x}, y)\}\,d\kappa\,d\lambda$$

is a continuous function of θ *on* $\tilde{B}_\Lambda$.

PROOF. Choose $\varphi_0(\mathbf{x}, y) = e^{c(y)}$ [see (2.1)]. Denote $f_* = f_\Lambda(\mathbf{x}, y;\theta)/\varphi_0(\mathbf{x}, y)$. Then $f_* = \sum_J g_J e_J$, where $e_J = \exp(ya(h_J) + b(h_J))$. By construction, f_* is a measurable function of $(\mathbf{x}, y)$ for each θ, and a continuous function of θ for each $(\mathbf{x}, y)$.

Note that

$$\begin{aligned}\sup_J|\log e_J| &\geq \sup_J(\log e_J) = \log\Big(\sup_J e_J\Big)\\ &\geq \log\Big(\sum_J g_J e_J\Big) \geq \sum_J g_J(\log e_J)\\ &\geq -\sup_J\left|\log e_J\right|,\end{aligned}$$

where we have used the monotonicity and concavity of $\log(\cdot)$ and the fact that g_J's are nonnegative and have a unity sum.

Hence, for all $\theta\in\tilde{B}_\Lambda$, $\mathbf{x}\in\Omega$ and $y\in A$,

$$\begin{aligned}\left|\log\{f_\Lambda(\mathbf{x}, y;\theta)/\varphi_0(\mathbf{x}, y)\}\right| = |\log f_*| &= \left|\log\Big(\sum_J g_J e_J\Big)\right|\\ &\leq \sup_J|\log e_J| = \sup_J\left|ya(h_J) + b(h_J)\right|\\ &\leq M_a|y| + M_b \equiv M(y),\end{aligned}$$

with $M_a = \sup_J\sup_{\mathbf{x}\in\Omega}\sup_{\theta\in\tilde{B}_\Lambda}|a(h_J)|$ and $M_b = \sup_J\sup_{\mathbf{x}\in\Omega}\sup_{\theta\in\tilde{B}_\Lambda}|b(h_J)|$ being finite, since $a(h_J)$ and $b(h_J)$ are continuous functions of $(\mathbf{x}^T, \theta^T)$ on the compact set $\Omega\otimes\tilde{B}_\Lambda$. Next we show that $\mathrm{E}\{M(Y)\} = \int\int M(y)\varphi\,d\lambda\,d\kappa < \infty$. Note that $\int_A y^2\pi(h, y)\,d\lambda$ is a continuous function in h, due to property (2) of Section 2.1. By the continuity of $h(\cdot)$, $\int_A y^2\pi(h(\mathbf{x}), y)\,d\lambda$ is a continuous function of $\mathbf{x}$ in the compact set Ω, leading to the finiteness of $\int_\Omega\{\int_A y^2\cdot\pi(h(\mathbf{x}), y)\,d\lambda\}^{1/2}\,d\kappa$. Therefore, $\int\int|y|\varphi\,d\lambda\,d\kappa$, being bounded above by $\int_\Omega\{\int_A y^2\pi(h(\mathbf{x}),\ y)\,d\lambda\}^{1/2}\,d\kappa$, is also finite, and so is $\mathrm{E}\{M(Y)\}$.

Therefore, $\log\{f_\Lambda(\mathbf{x}, y;\theta)/\varphi_0(\mathbf{x}, y)\}$ is bounded above by an integrable function. By the uniform law of large numbers [Jennrich (1969), Theorem 2, or White (1994), Theorem A.2.1], we obtain the strong uniform convergence of $Ł_{n,\Lambda}(\theta;\omega)$, as well as the continuity of $Ł_{\infty,\Lambda}(\theta)$. □

This proposition leads to the following lemma for the convergence of MLE.

LEMMA 2. *Let* $\hat{\theta}_{n,\Lambda}(\omega)$, $n = 1, 2, \ldots$, *be a sequence of global maximizers of* $Ł_{n,\Lambda}(\theta, \omega)$ *in* $\tilde{B}_\Lambda$, *as defined in* (2.10). *Let* Θ_Λ *be the set of minimizers of the KL divergence between the true density and the HME density*, *as defined in* (2.12). *We have*

$$\inf_{\theta \in \Theta_\Lambda} \rho_E\big(\hat{\theta}_{n,\Lambda}(\omega), \theta\big) \to 0 \quad \text{as } n \to \infty$$

for almost all ω, *where* $\rho_E(\cdot, \cdot)$ *is the Euclidean metric*.

PROOF. Denote

$$d(\phi, \Theta_\Lambda) \equiv \inf_{\theta \in \Theta_\Lambda} \rho_E(\phi, \theta) \quad \text{for } \phi \in \tilde{B}_\Lambda.$$

For any $\varepsilon > 0$, denote

$$B_\varepsilon^c(\Theta_\Lambda) \equiv \big\{\phi \in \tilde{B}_\Lambda \colon d(\phi, \Theta_\Lambda) \geq \varepsilon\big\}.$$

We show

(∗) For almost all ω, there exists $N(\omega, \varepsilon) \in \mathscr{N}$ such that $n > N(\omega, \varepsilon)$ implies $\hat{\theta}_{n,\Lambda}(\omega) \notin B_\varepsilon^c(\Theta_\Lambda)$. Then $d(\hat{\theta}_{n,\Lambda}(\omega), \Theta_\Lambda) \leq \varepsilon$.

To show (∗), note that by Proposition 1, $Ł_{n,\Lambda}(\cdot; \omega)$ is uniformly convergent to $Ł_{\infty,\Lambda}(\cdot)$ for almost all ω. Hence, for any $\delta > 0$, there exists $N(\omega)$ such that $n > N(\omega)$ implies that, for almost all ω,

$$Ł_{n,\Lambda}(\theta_0; \omega) > Ł_{\infty,\Lambda}(\theta_0) - \delta/2,$$

$$Ł_{n,\Lambda}(\phi; \omega) < Ł_{\infty,\Lambda}(\phi) + \delta/2,$$

for all $\theta_0 \in \Theta_\Lambda$ and all $\phi \in B_\varepsilon^c(\Theta_\Lambda)$. Hence,

$$Ł_{n,\Lambda}(\theta_0; \omega) - Ł_{n,\Lambda}(\phi; \omega) > Ł_{\infty,\Lambda}(\theta_0) - Ł_{\infty,\Lambda}(\phi) - \delta. \tag{4.12}$$

We can choose $\delta > 0$ so small that

$$\delta < \inf_{\phi \in B_\varepsilon^c(\Theta_\Lambda)} \big[Ł_{\infty,\Lambda}(\theta_0) - Ł_{\infty,\Lambda}(\phi)\big] \equiv I_\varepsilon. \tag{4.13}$$

Note that $I_\varepsilon > 0$. This is because $B_\varepsilon^c(\Theta_\Lambda)$ is compact and $Ł_{\infty,\Lambda}$ is continuous (by Proposition 1), and the infimum in (4.13) is achieved for some $\phi \in B_\varepsilon^c(\Theta_\Lambda)$, where $Ł_{\infty,\Lambda}(\theta_0) - Ł_{\infty,\Lambda}(\phi) > 0$. Hence, for all $\phi \in B_\varepsilon^c(\Theta_\Lambda)$, when $n > N(\omega)$, where $N(\omega)$ is as chosen above,

$$Ł_{n,\Lambda}(\theta_0; \omega) - Ł_{n,\Lambda}(\phi; \omega) > 0 \quad \text{by (4.12) and (4.13).}$$

So $\hat{\theta}_{n,\Lambda}(\omega) = \arg\max_{\phi \in \tilde{B}_\Lambda} Ł_{n,\Lambda}(\phi; \omega) \notin B_\varepsilon^c(\Theta_\Lambda)$, leading to the proof of (∗) and the lemma. □

The next lemma shows that the "stochastic part" of the MLE goes to zero when the sample size increases.

LEMMA 3. *There is a sequence of* $\theta_{n,\Lambda} \in \Theta_\Lambda$, $n = 1, 2, \ldots$, *possibly random, making* $S_{n,\Lambda} \to 0$ *as* $n \to 0$, *for any structure* Λ.

PROOF. Note that Lemma 2 implies that there exists a sequence $\theta_{n,\Lambda}(\omega) \in \Theta_\Lambda$, $n = 1, \ldots,$ such that $\rho_E(\hat{\theta}_{n,\Lambda}(\omega), \theta_{n,\Lambda}(\omega)) \to 0$ for almost all ω. By the definition in (2.8), $\mu_\Lambda(\cdot;\cdot)$ is continuous on $\Omega \otimes \tilde{B}_\Lambda$, due to the continuity of the $g_J(\cdot;\cdot)$'s and $\psi(\cdot)$. Hence, we have

$$\left\{\mu_\Lambda(\mathbf{x};\, \hat{\theta}_{n,\Lambda}(\omega)) - \mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}(\omega))\right\}^2 \to 0$$

for all $\mathbf{x}$ and almost all ω.

Next we show that $\{\mu_\Lambda(\mathbf{x};\, \hat{\theta}_{n,\Lambda}(\omega)) - \mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}(\omega))\}^2$ is bounded above by an integrable function. This is because

$$\begin{aligned}
&\left\{\mu_\Lambda(\mathbf{x};\, \hat{\theta}_{n,\Lambda}(\omega)) - \mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}(\omega))\right\}^2 \\
&\quad \le 2\left\{\mu_\Lambda(\mathbf{x};\, \hat{\theta}_{n,\Lambda}(\omega))\right\}^2 + 2\left\{\mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}(\omega))\right\}^2 \\
&\quad \le 4M_\Lambda^2,
\end{aligned}$$

where $M_\Lambda \equiv \sup_{\mathbf{x}\in\Omega} \sup_{\theta\in\tilde{B}_\Lambda} |\mu_\Lambda(\mathbf{x};\, \theta)| < \infty$.

Therefore, by the Lebesgue's dominated convergence theorem,

$$\mathrm{E}\int \left[\mu_\Lambda(\mathbf{x};\, \hat{\theta}_{n,\Lambda}) - \mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda})\right]^2 d\kappa(\mathbf{x}) \to 0. \qquad \square$$

Now we provide an upper-bound for the "deterministic part" $2D_{n,\Lambda}$.

LEMMA 4. *Let* $D_{n,\Lambda} = E\int_\Omega \{\mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}) - \mu(\mathbf{x})\}^2\, d\kappa(\mathbf{x})$ *be as defined in* (4.10). *If Conditions* $A_{\mathscr{S},1}$ *and* 2 *hold, then we have*

$$\inf_{\Lambda\in\mathscr{J}_m\cap\mathscr{S}} D_{n,\Lambda} \le c^*/m^{4/s},$$

for some finite positive constant c^* *independent of the number of experts* m.

PROOF. Consider a sequence of maximizers $\theta_{n,\Lambda} = \theta_{n,\Lambda}(\omega)$ in Θ_Λ. Note that for each ω, n and Λ, we have

$$\begin{aligned}
\int \left\{\mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}(\omega)) - \mu(\mathbf{x})\right\}^2 d\kappa &\overset{\text{(i)}}{\le} 4M_I \mathrm{KL}\left\{f_\Lambda(\cdot,\cdot;\, \theta_{n,\Lambda}(\omega)), \varphi(\cdot,\cdot)\right\} \\
&\overset{\text{(ii)}}{=} 4M_I \inf_{f\in\Pi_\Lambda} \mathrm{KL}(f, \varphi).
\end{aligned}$$

Then $D_{n,\Lambda} = \mathrm{E}\int \{\mu_\Lambda(\mathbf{x};\, \theta_{n,\Lambda}(\omega)) - \mu(\mathbf{x})\}^2\, d\kappa \le 4M_I \inf_{f\in\Pi_\Lambda} \mathrm{KL}(f, \varphi)$, and

$$\begin{aligned}
\inf_{\Lambda\in\mathscr{J}_m\cap\mathscr{S}} D_{n,\Lambda} &\le 4M_I \inf_{\Lambda\in\mathscr{J}_m\cap\mathscr{S}} \inf_{f\in\Pi_\Lambda} \mathrm{KL}(f, \varphi) \\
&= 4M_I \inf_{f\in\mathbf{\Pi}_{m,\mathscr{S}}} \mathrm{KL}(f, \varphi) \overset{\text{(iii)}}{\le} 4M_I c/m^{4/s}.
\end{aligned}$$

Here (i) is due to Lemma 1(a), where the constant M_I can be made independent of φ and $f_\Lambda(\cdot, \cdot; \theta_{n,\Lambda}(\omega))$ by using an upperbound $K = \max\{(s+1)C, K_0\}$, where C is the constant in Condition 2 and K_0 is the radius of the Sobolev ball in Section 2.2. [Note that for all parameters in $\tilde{B}_\Lambda$, $|h_J(\mathbf{x})|$ is less than $(s+1)C$, and for all φ in $\mathbf{\Phi}$, $|h(\mathbf{x})|$ is less than K_0.] (ii) is due to Condition 2, and equation (4.11); (iii) is due to Theorem 2 and Condition $A_{\mathscr{S},1}$. □

Now we are ready to prove the consistency theorem.

PROOF OF THEOREM 3. Note that $(\text{MSE})_{n,\Lambda} \le 2D_{n,\Lambda} + 2S_{n,\Lambda}$ from (4.10). For each Λ in $\mathscr{J}_m \cap \mathscr{S}$, find a (possibly random) sequence $\theta_{n,\Lambda}$ as in Lemma 3 such that $S_{n,\Lambda} \to 0$. Then $\sup_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} S_{n,\Lambda} \to 0$ as $n \to \infty$, since the cardinality of $\mathscr{J}_m \cap \mathscr{S}$ is finite. Then we have

$$\begin{aligned}\inf_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} (\text{MSE})_{n,\Lambda} &\le \inf_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} \left(2S_{n,\Lambda} + 2D_{n,\Lambda}\right) \\ &\le 2 \sup_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} S_{n,\Lambda} + 2 \inf_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} D_{n,\Lambda} \\ &\le o_m(n^0) + 2c^*/m^{4/s},\end{aligned}$$

due to Lemma 4. [We denote the term $\sup_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} S_{n,\Lambda}$ as $o_m(n^0)$, since it is $o(1)$ as $n \to \infty$, and is possibly dependent on m.]

Therefore

$$0 \le \limsup_{n \to \infty} \inf_{\Lambda \in \mathscr{J}_m \cap \mathscr{S}} (\text{MSE})_{n,\Lambda} \le 2c^*/m^{4/s}$$

and

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

5. Conclusions. We investigated the power of the HME for GLM1 experts in terms of approximating a flexible class of density functions with conditional mean functions belonging to a transformed Sobolev class. We demonstrated that the approximation rate of HME density functions is $O(m^{-2/s})$ in Hellinger distance and $O(m^{-4/s})$ 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 $O(m^{-2/s})$ rate is optimal. In fact, for the special case of mixing linear model experts in a single layer, Zeevi, Meir and Maiorov (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, Meir and Maiorov (1998) in the following aspects: (1) We deal with mixtures of *generalized* linear models instead of the mixtures of ordinary linear models. (2) We consider the setup of the HME networks instead of the single-layer mixtures of experts. (3) We consider the maximum likelihood method instead of the least-squares approach for model fitting. (4) In relation to the use of the maximum likelihood method, we obtained the approximation rate in terms of probability density functions instead of in terms of the mean response. (5) We have formulated the conditions and proofs of our results in a way that is protective of the inherent nonidentifiability problems of the parameterization.

Acknowledgments. The authors thank John Kolassa and the referees for helpful comments and Assaf J. Zeevi for suggesting a reference for Lemma 1(b).

REFERENCES

BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, NJ.

BISHOP, C .M. (1995). *Neural Networks for Pattern Recognition*. Oxford Univ. Press.

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

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

FRITSCH, J., FINKE, M. and WAIBEL, A. (1997). Adaptively growing hierarchical mixtures of experts. In *Advances in Neural Informations Processing Systems* **9** (M. C. Mozer, M. I. Jordan and T. Petsche, eds.). MIT Press.

GHAHRAMANI, Z. and HINTON, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Dept. Computer Science, Univ. Toronto.

HAUSSLER, D. and OPPER, M. (1995). General bounds on the mutual information between a parameter and n conditionally independent observations. In *Proceedings of the Eighth Annual Computational Learning Theory Conference (COLT), 1995, Santa Cruz, CA*. ACM Press, New York.

HAYKIN, S. (1994). *Neural Networks*. Macmillan, New York.

JAAKKOLA, T. S. and JORDAN, M. I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models* (M. I. Jordan, ed.). Kluwer, Dordrecht.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp.* **3** 79–87.

JENNRICH, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* **40** 633–643.

JIANG, W. and TANNER, M. A. (1998). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. Technical report, Dept. Statistics, Northwestern Univ., Evanston, IL.

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 1567, Artificial Intelligence Lab., Massachusetts Institute Technology.

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, Dept. Computer Science and Applied Mathematics, Aston Univ., Birmingham, UK.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

WHITE, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge Univ. Press.

ZEEVI, A. and MEIR, R. (1997). Density estimation through 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.

DEPARTMENT OF STATISTICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60208
E-MAIL: wjiang@nwu.edu
tanm@neyman.stats.nwu.edu