

On Consistency of Bayesian Inference with Mixtures of Logistic Regression

Yang Ge and Wenxin Jiang

Department of Statistics, Northwestern University

Evanston, IL 60208, USA

Abstract

This is a theoretical study of the consistency properties of Bayesian inference using mixtures of logistic regression models. When standard logistic regression models are combined in a ‘mixtures of experts’ set-up, a flexible model is formed to model the relationship between a binary (yes-no) response y and a vector of predictors \mathbf{x} . Bayesian inference conditional on the observed data can then be used for regression and classification. The paper gives conditions on choosing the number of experts (i.e., number of mixing components) k , or choosing a prior distribution for k , so that Bayesian inference is ‘consistent’, in the sense of ‘often approximating’ the underlying true relationship between y and \mathbf{x} . The resulting classification rule is also ‘consistent’, in the sense of having near-optimal performance in classification. We show these desirable consistency properties with a nonstochastic k growing slowly with the sample size n of the observed data, or with a

random k that takes large values with nonzero but small probabilities.

1 Introduction

Mixtures of experts (ME, Jacobs, Jordan, Nowlan and Hinton 1991) and Hierarchical mixtures of experts (HME, Jordan and Jacobs 1994) are popular techniques for regression and classification, and have attracted attentions in both areas of neural networks and statistics. ME and HME are a variety of neural networks that have an interpretation of probabilistic mixture, in contrast to the usual neural nets that are based on linear combinations. With mixture, instead of with linear combinations, simple models are combined in ME and HME for improved predictive capability. This structure of probabilistic mixture allows the use of convenient computing algorithms such as the EM (Expectation Maximization) algorithm (Jordan and Xu 1995) and the Gibbs sampler (Peng, Jacobs and Tanner 1996).

The ME and HME are flexible constructions that can allow various models or ‘experts’ to be mixed. For binary classification, simple and standard classifiers such as logistic regression models can be combined, to model the relationship between a binary response $y \in \{0, 1\}$ and a predictor vector \mathbf{x} . Such combined models can approximate arbitrary smooth relations between y and \mathbf{x} in the sense of Jiang and Tanner (1999a). Peng, Jacobs and Tanner (1996) applies mixtures of binary and multinomial logistic regression for pattern recognition. They found that a Bayesian approach based on simulating the posterior distribution gives better performance than frequentist approach based on likelihood. Recently, Wood,

Kohn, Jiang and Tanner (2005) study binary regression where probit-transformed spline models with different smoothing parameters are mixed, and Markov Chain Monte Carlo methods are used to simulate the posterior distributions for both the model parameters and the number of mixing components. Their extensive simulations and empirical studies demonstrate excellent performance of the Bayesian approach and local adaptivity of the mixing paradigm. These empirical successes have motivated us to study the theoretical reasons behind: Why does the Bayesian procedure work well in such mixture models of binary regression?

The purpose of the current paper is to study the ‘consistency’ properties of Bayesian inference for mixtures of binary logistic regression. *Will inferential results based on the posterior distribution be reliable?* In a Bayesian approach, the posterior distribution will propose various relationships between \mathbf{x} and y , based on some observed data. We will investigate the conditions under which the proposed relationships are ‘consistent’ or ‘often close’ to the underlying true relationship. This will also imply that the resulting ‘plug-in’ classification rule have near-optimal performance.

There are several senses of consistency. The precise formulation of these problems are given in Section 2. We assume that the true model possesses some unknown smooth mean function $E(y|\mathbf{x})$, which can be *outside* of the proposed ME family, and that the observed data $(y_i, \mathbf{x}_i)_{i=1}^n$ are n independent and identical copies of (y, \mathbf{x}) .

In Section 3 we will first study the consistency problem for a sequence of ME models, where the number of experts (or mixture components) $K = k_n$ increases with sample size n . Such a construct allows a large number of experts eventually

and enables good functional approximation (Jiang and Tanner 1999a). We will show that the following condition on k_n leads to consistency: k_n increases to infinity at a rate slower than n^a for some $a \in (0, 1)$, as sample size n increases.

Later in Section 3 we will consider the case when the number of experts K is regarded to be random and follows a prior distribution. We will show that the critical conditions for consistency involve the prior on K : the prior is supported for all large values of K and has a sufficiently thin tail.

Our work parallels Lee (2000), who studies similar properties (without classification consistency) for ordinary neural networks (NN) based on linear combinations. Lee (2000)’s method involves truncating the space of all proposed models into a limited part and an unlimited part, and show that (i) the unlimited part has very small prior probability satisfying some ‘tail condition’; (ii) the limited part is not too complicated in the sense of satisfying an ‘entropy condition’; (iii) the prior is chosen to have not-too-small probability mass around the true model, which is an ‘approximation condition’.

Condition (iii) typically involves some approximation results, since the prior-proposed models have to be able to get as near as possible to the true model, or else over some neighborhood of the true model the prior mass would be zero.

We approach by implementing these conditions for mixtures of experts (ME). However, we note that there is a fundamental difficulty resulting from the mechanism of approximation with ME. In the known mechanism (Jiang and Tanner 1999a), to approximate a true relation arbitrarily well, ME with many experts ‘crowded together’ and with large parameter values are used. This results in large parameter values of the ME model (the components that describe the changing

of mixing weights) increasing with K . In Bayesian approach, typically very *small* prior is given to such ME configurations; they have large parameter values lying in the tail of the prior. Yet we would like to show that the resulting posterior of such configurations are *not too small*, since these configurations are close to the true relation.

In order to handle this difficulty, we characterize how large the ME parameter values are needed for good approximation: values of order $\ln(K)$ are sufficient, which are in fact not too far in the tail of the prior distribution. Such a result is established by embedding ME with $K^*(< K)$ -experts as a subset of ME with K -experts. (See Lemma 5 and its proof.)

When we consider the situation with random K , we face another difficulty: the usual priors of K , such as Geometric or Poisson, cannot satisfy both the conditions (i) and (ii). If the truncation occurs at a too-large K , the limited part of the proposed model space may become too complicated to satisfy the entropy condition. If the truncation occurs at a too-small K , the tails may be too thick to satisfy the tail condition. Such a dilemma was not discussed in Lee (2000), who did not consider the entropy condition for the case of random K .

In order to handle this situation, we introduce a ‘contraction sequence’ for the number of experts: $K = k(i)$ which grows to infinity as i increases but grows *slower* than i . Then the prior probability, e.g., $\lambda_i = (0.5)^i$ for the Geometric, is put on index i . Since $k(i)$ can stay unchanged for some i , this contraction sequence effectively groups the geometric probabilities together at the choice of number of experts and produces a thinner tail. We show that using suitably contracted Geometric or Poisson priors on K , all conditions hold to produce consistency.

2 Notation and Definitions

2.1 Models

We first define the single-layer mixture-of-experts models where logistic regression models are mixed.

The binary response variable is $y \in \{0, 1\}$ and \mathbf{x} is an s -dimensional predictor. As in Jiang and Tanner (1999a,b), we let $\Omega = [0, 1]^s = \otimes_{q=1}^s [0, 1]$ be the space of the predictor \mathbf{x} , and let \mathbf{x} have a uniform distribution on Ω . This is a convenient starting point and the results can be easily adapted to the case when \mathbf{x} has a positive density and is supported on a compact set.

This convenient formulation results in several simplified relations. The joint density of $p(y, \mathbf{x})$ is the same as the conditional density $p(y|\mathbf{x})$, which is completely determined by the conditional probability of a positive response $P(y = 1|\mathbf{x})$, which is equal to the condition mean or regression function $\mu(\mathbf{x}) = E(y|\mathbf{x})$, which is alternatively formulated in a transformed version $h(\mathbf{x}) = \log\{\mu(\mathbf{x})/(1 - \mu(\mathbf{x}))\}$, called the log-odds.

We consider a family Φ of ‘smooth relations’ between y and \mathbf{x} as defined in Jiang and Tanner (1999a,b): *Φ is the family of joint densities $p(y, \mathbf{x})$ such that the log-odds $h(\mathbf{x})$ has continuous derivatives up to the second order, which are all bounded above by a constant, uniformly over \mathbf{x} .*

Such a nonparametric family Φ can be approximated by mixtures of logistic regression (Jiang and Tanner 1999a,b). Define the family $\mathbf{\Pi}_k$, of mixtures of k logistic regression models, as follows: $\mathbf{\Pi}_k$ is the set of joint densities $f(y, \mathbf{x}|\theta)$,

such that the conditional densities have the form $f(y|\mathbf{x}, \theta) = \sum_{j=1}^k g_j H_j$, where $g_j = \frac{e^{u_j + \mathbf{v}_j^T \mathbf{x}}}{\sum_{l=1}^k e^{u_l + \mathbf{v}_l^T \mathbf{x}}}$; $H_j = \left(\frac{e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}}{1 + e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}} \right)^y \left(\frac{1}{1 + e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}} \right)^{1-y}$. The α , $\boldsymbol{\beta}$, u , \mathbf{v} 's are parameters of the model. Except that we restrict $u_1 = 0$ and $\mathbf{v}_1 = \mathbf{0}$ for the sake of the identifiability, we allow all components of the parameter vectors to vary in $(-\infty, \infty)$. We denote by θ the combined vector of parameters $\theta = (\alpha_1, \boldsymbol{\beta}_1^T, \dots, \alpha_k, \boldsymbol{\beta}_k^T, u_2, \mathbf{v}_2^T, \dots, u_k, \mathbf{v}_k^T)^T \in \Re^{\dim(\theta)}$, where $\dim(\theta) = (s+1)(2k-1)$.

2.2 Bayesian inference

The observed dataset is $(Y_1, X_1), \dots, (Y_n, X_n)$, which we simply denote as $(Y_i, X_i)^n$. Here n is the sample size. We assume $(Y_i, X_i)^n$ to be an iid (independent and identically distributed) sample of an unknown density f_0 from the *nonparametric* family of smooth relations Φ . The mixture of logistic regression approach involves estimating the nonparametric f_0 using *parametric* relations f from $\mathbf{\Pi}_k$, the family of mixtures of k logistic regression models.

We now describe Bayesian inference for uncovering f_0 based on $(Y_i, X_i)^n$. In the mixture of logistic regression approach, one first puts a prior distribution π_n to propose densities f from the k -mixture family $\mathbf{\Pi}_k$ (through the corresponding parameters θ). This prior will then produce a posterior distribution of f over $\mathbf{\Pi}_k$ (through the corresponding θ), *conditional on the observed data*: $\pi_n(d\theta | ((Y_i, X_i)^n)) = \prod_{i=1}^n f(Y_i, X_i | \theta) \pi_n(d\theta) / \int \prod_{i=1}^n f(Y_i, X_i | \theta) \pi_n(d\theta)$.

Then, the predictive density, which is the Bayes estimate of f_0 , is given by

$$\hat{f}_n(\cdot) = \int f(\cdot | \theta) \pi_n(d\theta | (Y_i, X_i)^n).$$

Let $\mu_0(\mathbf{x}) = E_{f_0}[Y|X = \mathbf{x}] = \sum_{y=0,1} y f_0(y|\mathbf{x})$ be the true regression function, then $\hat{\mu}_n(\mathbf{x}) = E_{\hat{f}_n}[Y|X = \mathbf{x}]$ is the estimated regression function.

For now we will let $k = k_n$ be nonstochastic and possibly depend on sample size n , which explains the dependence of prior on n . Later we will also consider the case when $k = K$ is itself regarded as a random component of the parameter; a prior randomly decides to use an f from $\mathbf{\Pi}_k$ with probability $P(K = k)$, $k = 1, 2, 3, \dots$

The prior densities on θ -components are assumed to be independent normal with zero mean and common positive variance σ^2 . (The results can be easily generalized to cases with different means and variances.)

2.3 Consistency

We first define consistency in regression function estimation, which we will call *R-consistency*.

Definition 1 (*R-Consistency*). $\hat{\mu}_n$ is asymptotically consistent for μ_0 if

$$\int (\hat{\mu}_n(\mathbf{x}) - \mu_0(\mathbf{x}))^2 d\mathbf{x} \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty.$$

Here and below, the convergence in probability of the form $q\{(Y_i, X_i)^n\} \xrightarrow{P} q_0$, for any quantity dependent on the observed data, means $\lim_{n \rightarrow \infty} P_{(Y_i, X_i)^n}[|q\{(Y_i, X_i)^n\} - q_0| \leq \epsilon] = 1$ for all $\epsilon > 0$, where $(Y_i, X_i)^n$ are an iid random sample from the true density f_0 . This definition describes a desirable property for the estimated regression function $\hat{\mu}_n$ to be *often* (with $P_{(Y_i, X_i)^n}$ tending to one) *close* (in L_2 sense) to the true μ_0 , for large n .

Now we define consistency in terms of the density function, which we will term as *D-consistency*. First, for any $\varepsilon > 0$, define a Hellinger ε -neighborhood by

$$A_\varepsilon = \{f : D_H(f, f_0) \leq \varepsilon\}$$

where $D_H(f, f_0) = \sqrt{\int \int (\sqrt{f} - \sqrt{f_0})^2 dx dy}$ is the Hellinger distance.

Definition 2 (*D-Consistency*). Suppose $(Y_i, X_i)^n$ is an iid random sample from density f_0 . The posterior is asymptotically consistent for f_0 over Hellinger neighborhood if for any $\varepsilon > 0$,

$$\Pr(A_\varepsilon | (Y_i, X_i)^n) \xrightarrow{P} 1 \quad \text{as } n \rightarrow \infty.$$

That is, the posterior probability of any Hellinger neighborhood of f_0 converges to 1 in probability.

This definition describes a desirable property for the posterior-proposed joint density f to be *often close* to the true f_0 , for large n .

Now we define the consistency in classification, which we will call *C-consistency*. Here we consider the use of the ‘plug-in’ classification rule $\hat{C}_n(\mathbf{x}) = I[\hat{\mu}_n(\mathbf{x}) > 1/2]$ in predicting Y . We are interested in how the misclassification error $E_{(Y_i, X_i)^n} P\{\hat{C}_n(X) \neq Y | (Y_i, X_i)^n\}$ approaches the minimal error $P\{C_o(X) \neq Y\} = \inf_{C: \text{Dom}(X) \rightarrow \{0,1\}} P\{C(X) \neq Y\}$, where $C_o(\mathbf{x}) = I[\mu_0(\mathbf{x}) > 1/2]$ is the ideal ‘Bayes rule’ based on the (unknown) true mean function μ_0 .

Definition 3 (*C-Consistency*). Let $\hat{B}_n : \text{Dom}(X) \mapsto \{0, 1\}$ be a classification rule that is computable based on the observed data $(Y_i, X_i)^n$. If $\lim_{n \rightarrow \infty} E_{(Y_i, X_i)^n} P\{\hat{B}_n(X) \neq Y | (Y_i, X_i)^n\} = P\{C_o(X) \neq Y\}$, then \hat{B}_n is called a consistent classification rule.

It is straightforward to show that three consistency concepts are related in our situation with binary data, where $\hat{\mu}_n$ and μ_0 are bounded between $[0, 1]$:

Proposition 1 (*Relations among three consistencies*). D -Consistency $\implies R$ -Consistency $\implies C$ -Consistency.

Proof: The first relation is due to Lemma 4. The second relation is due to Corollary 6.2 of Devroye, Györfi and Lugosi (1996). \square

In the paper we will first establish D-consistency, then R- and C- consistencies naturally follow.

3 Results and Conditions

We first consider the case when the number of experts $K = k_n$ is a nonstochastic sequence depending on sample size n .

Theorem 1 (*Nonstochastic K*) Let the prior for the parameters, $\pi_n(d\theta)$, be independent normal distributions with mean zero and fixed variance σ^2 for each parameter in the model. Let k_n be the number of experts in the model, such that

(i) $\lim_{n \rightarrow \infty} k_n = \infty$ and

(ii) $k_n \leq n^a$ for all sufficiently large n , for some $0 < a < 1$.

Then we have the following results.

a). The posterior distribution of the densities is D -consistent for f_0 , i.e.,

$\Pr(\{f : D_H(f, f_0) \leq \epsilon\} | (Y_i, X_i)^n) \xrightarrow{P} 1$ as $n \rightarrow \infty$, for all $\epsilon > 0$.

b). The estimated regression function $\hat{\mu}_n$ is R -consistent for μ_0 , i.e. $\int(\hat{\mu}_n - \mu_0)^2 d\mathbf{x} \xrightarrow{P} 0$ as $n \rightarrow \infty$.

c). The plug-in classification rule $\hat{C}_n(\mathbf{x}) = I[\hat{\mu}_n(\mathbf{x}) > 1/2]$ is C -consistent for the Bayes rule $C_o(\mathbf{x}) = I[\mu_0(\mathbf{x}) > 1/2]$, i.e., $\lim_{n \rightarrow \infty} E_{(Y_i, X_i)^n} P\{\hat{C}_n(X) \neq Y | (Y_i, X_i)^n\} = P\{C_o(X) \neq Y\}$.

Now we consider the case when the number of experts K is a random parameter. We will consider the possibility that $K = k(I)$ is constructed out of a more basic random index I , which, for example, can be the Geometric or the Poisson distribution. The function $k(\cdot)$ will be called a *contraction* function. We will see the reason to introduce the contraction: the sufficient condition we propose on K requires a very thin probability; common distributions such as geometric or Poisson can be used only after a tail-thinning contraction.

The densities $f(y, \mathbf{x}|k, \theta)$ are now indexed by both the parameter vector θ and the number of experts k . The prior is $\pi(k, d\theta) = \tilde{\lambda}_k \pi(d\theta|k)$, where $\tilde{\lambda}_k = P[k(I) = k]$, and $\pi(d\theta|k)$ is again chosen to be the independent $N(0, \sigma^2)$ distributions on all components of θ . The posterior distribution is then $\pi(k, d\theta | ((Y_i, X_i)^n) = \prod_{i=1}^n f(Y_i, X_i|k, \theta) \pi(k, d\theta) / \sum_{j=1}^{\infty} \prod_{i=1}^n f(Y_i, X_i|j, \theta') \pi(j, d\theta')$.

Then, the predictive density, which is the Bayes estimate of f_0 , is given by $\hat{f}_n(\cdot) = \sum_{k=1}^{\infty} \int f(\cdot|k, \theta) \pi(k, d\theta | (Y_i, X_i)^n)$. The corresponding estimated regression function is $\hat{\mu}_n(\mathbf{x}) = \sum_{y=0,1} y \hat{f}_n(y|\mathbf{x})$. The plug-in classification rule is $\hat{C}_n(\mathbf{x}) = I[\hat{\mu}_n(\mathbf{x}) > 1/2]$.

Theorem 2 (*Random K*) Suppose the priors $\pi(d\theta|k)$ conditional on the number of experts are independent normal with mean 0 and fixed variance σ^2 . Suppose

prior put on the number of experts $k(I)$ satisfies the following conditions:

(iii) $P[k(I) = k] > 0$ for all sufficiently large k ;

(iv) The tail probabilities decrease at a faster-than-geometric rate, i.e., there exists $q > 1$ such that fixing any $r > 0$, for all sufficiently large k , $P[k(I) \geq k] \leq \exp(-k^q r)$.

Then we have the following results.

d). The posterior distribution of the densities is D -consistent for f_0 , i.e., $\Pr(\{f : D_H(f, f_0) \leq \epsilon\} | (Y_i, X_i)^n) \xrightarrow{P} 1$ as $n \rightarrow \infty$, for all $\epsilon > 0$.

e). The estimated regression function $\hat{\mu}_n$ is R -consistent for μ_0 , i.e. $\int (\hat{\mu}_n - \mu_0)^2 d\mathbf{x} \xrightarrow{P} 0$ as $n \rightarrow \infty$.

f). The plug-in classification rule $\hat{C}_n(\mathbf{x}) = I[\hat{\mu}_n(\mathbf{x}) > 1/2]$ is C -consistent for the Bayes rule $C_o(\mathbf{x}) = I[\mu_0(\mathbf{x}) > 1/2]$, i.e., $\lim_{n \rightarrow \infty} E_{(Y_i, X_i)^n} P\{\hat{C}_n(X) \neq Y | (Y_i, X_i)^n\} = P\{C_o(X) \neq Y\}$.

The super-geometrically-thin tail condition (iv) cannot be directly satisfied, if the number of experts follows some common distributions such as Geometric or Poisson. However, if one applies a contraction $k(\cdot)$ to a Geometric or Poisson random variable, where $k(\cdot)$ grows very slowly, the probability of a large *contracted* $k(I)$ can be made sufficiently small.

Remark 1 For example, consider the contractions of the form $k(I) = \lceil \chi(I) \rceil + 1$, where $\lceil u \rceil$ represents the integer part of u , and $\chi(I)$ is a strictly and slowly increasing function. It is easy to confirm that when I is a Geometric random variable, taking $\chi(I) = I^{1/q^{1+\delta}}$ (for some $\delta > 0$ and $q > 1$) will make $k(I)$ satisfy condition (iv), after using the equation $P(I \geq B) = P(I > 0)^B$. When I is

a Poisson random variable, taking $\chi(I) = \{\ln(I + 1)\}^{1/q^{1+\delta}}$ (for some $\delta > 0$ and $q > 1$) will make $k(I)$ satisfy condition (iv), after applying a Chebyshev's inequality to obtain $P(I \geq B) \leq EI/B$.

Below we will first give the proofs of these main theorems. The lemmas used will be stated and proved later.

3.1 Proof of Theorem 1

The proof involves splitting the space $\mathbf{\Pi}_{k_n}$ of all k_n -expert densities into a limited part \mathcal{F}_n and an unlimited part \mathcal{F}_n^c and applying Proposition 2 below.

Let \mathcal{F}_n be the set of mixture-of-experts models with each parameter bounded by C_n in absolute value. That is,

$$|u_j| \leq C_n, |v_{jh}| \leq C_n, |\alpha_j| \leq C_n, |\beta_{jh}| \leq C_n, \quad j = 1, \dots, k_n, h = 1, \dots, s,$$

where C_n grows with n such that $n^{\frac{1}{2}+\eta} \leq C_n \leq \exp(n^{b-a})$ for some $\eta > 0$ and $0 < a < b < 1$ (a is the same a as in the bound of k_n).

Proposition 2 (Lee 2000, Theorem 2).

Suppose the following conditions hold:

Tail condition i. There exists an $r > 0$ and N_1 , such that $\pi_n(\mathcal{F}_n^c) < \exp(-nr) \quad \forall n \geq N_1$;

Entropy condition ii. There exists some constant $c > 0$ such that $\forall \varepsilon > 0$, $\int_0^\varepsilon \sqrt{H_{[\cdot]}(u)} du \leq c\sqrt{n\varepsilon^2}$ for all sufficiently large n ;

Approximation condition iii. For all $\gamma, \nu > 0$, there exists an N_2 , such that $\pi_n(\text{KL}_\gamma) \geq \exp(-n\nu)$, $\forall n \geq N_2$.

Then the posterior is asymptotically consistent for f_0 over Hellinger neighborhoods, i.e., for any $\epsilon > 0$,

$$\Pr(\{f : D_H(f, f_0) \leq \epsilon\} | (Y_i, X_i)^n) \xrightarrow{P} 1 \quad \text{as } n \rightarrow \infty.$$

Here, for any $\gamma > 0$, define a Kullback-Leibler γ -neighborhood by

$$\text{KL}_\gamma = \{f : D_K(f, f_0) \leq \gamma\},$$

where $D_K(f, f_0) = \int \int f_0 \ln(f_0/f) dx dy$ is the Kullback-Leibler divergence.

This proposition was proved in Lee (2000), Theorem 2, where the entropy condition was used but not stated explicitly. Here $H_{[\cdot]}(\cdot)$ is the Hellinger bracketing entropy defined in the following steps, where the family of function is taken to be $\mathcal{F}^* = \{\sqrt{f} : f \in \mathcal{F}_n\}$, the set of square-roots of densities from \mathcal{F}_n , and the metric is the L_2 -norm so that $\|\sqrt{f} - \sqrt{g}\| = D_H(f, g)$ the Hellinger distance, for any two densities f and g .

Definition 4 (*Brackets and bracketing entropy*)

- i. For any two functions l and u , define the bracket $[l, u]$ as the set of all functions f such that $l \leq f \leq u$.
- ii. Let $\|\cdot\|$ be a metric. Define an ε -brackets as a bracket with $\|u - l\| \leq \varepsilon$.
- iii. Define the bracketing number of a set of functions \mathcal{F}^* as the minimum number of ε -brackets needed to cover \mathcal{F}^* , and denote it by $N_{[\cdot]}(\varepsilon, \mathcal{F}^*, \|\cdot\|)$.
- iv. The bracketing entropy, denoted by $H_{[\cdot]}(\cdot) = \ln N_{[\cdot]}(\cdot, \mathcal{F}^*, \|\cdot\|)$, is the natural logarithm of the bracketing number.

Now we prove Theorem 1. Lemma 1 guarantees the tail condition. Lemma 2 guarantees the entropy condition. Lemma 3 guarantees the approximation condition. Therefore we have the D-consistency due to Proposition 2. The R- and C-consistencies follow from Proposition 1. \square

3.2 Proof of Theorem 2

Let \mathcal{G}_m be a restricted set of mixtures of m -experts models, whose parameter components are all bounded by $C_n m$ in absolute value, with $n^{\frac{1}{2}+\eta} \leq C_n \leq \exp(n^{b-a})$ for some $\eta > 0$ and $0 < a < b < 1$, where $a = 1/q$. Such restricted sets \mathcal{G}_m are nested due to Proposition 3 (later).

We let $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{G}_k$, where $k_n = \lceil (cn)^a \rceil$, $a = 1/q \in (0, 1)$ and $c \in (0, 1]$.

Tail condition i:

Note that $\pi(\mathcal{F}_n^c) = 1 - \pi(\mathcal{F}_n) \leq \sum_{k=k_n+1}^{\infty} \tilde{\lambda}_k + \sum_{k=1}^{k_n} \tilde{\lambda}_k \pi(\|\theta\|_{\infty} > C_n k | k)$, where $\|\theta\|_{\infty}$ is the maximum absolute value of all the θ components. For all sufficiently large n and all $r > 0$, the tail probability $\sum_{k=k_n+1}^{\infty} \tilde{\lambda}_k$ is less than $e^{-nr}/2$ due to Condition (iv). All tail probabilities $\pi(\|\theta\|_{\infty} > C_n k | k)$ are less than $e^{-nr}/2$ also, due to the choice of C_n and the normality of $\pi(d\theta | k)$ (using the Mill's Ratio for normal tail probabilities). Therefore $\pi(\mathcal{F}_n^c) \leq e^{-nr}$ for all sufficiently large n and all $r > 0$, showing the tail condition i.

Entropy condition ii:

Note that the $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{G}_k = \mathcal{G}_{k_n}$ since the sets of density functions represented by \mathcal{G}_k are increasing with k (Proposition 3). Then the entropy condition

can be computed for \mathcal{G}_{k_n} , where the bounds of the parameter values are now $C_n k_n$ instead of the previous bound C_n . Repeating the proof of the entropy condition as before shows that the condition still holds.

Approximation condition iii:

Fix any $\gamma > 0$. Then $\pi(KL_\gamma) = \sum_{k=1}^{\infty} P(K = k)\pi(KL_\gamma|k) \geq P(K = k_n)\pi(KL_\gamma|k_n) > 0$, due to the positive $P(K = k_n)$ (guaranteed by condition (iii) of Theorem 2) and that $\pi(KL_\gamma|k_n) > e^{-n\nu}$, fixing any $\nu > 0$, for all large enough n , which was proved for nonstochastic k_n before. (Here $k_n = \lceil (cn)^a \rceil$ is less than n^a and increases to ∞ .) Therefore $\pi(KL_\gamma) \geq e^{-n\nu}$ for all sufficiently large n , fixing any $\nu > 0$, since the left hand side is positive and not dependent on n . This shows the approximation condition iii.

So all conditions of Proposition 2 hold and the D-consistency holds, which further implies the R- and C- consistency. \square

4 Lemmas Used for Proving the Theorems

In the first three lemmas, the number of experts k_n satisfies conditions (i) and (ii) of Theorem 1. The prior π_n for the parameters is such that each parameter is an independent normal with mean 0 and fixed variance σ^2 . The dimension of the parameters $\dim(\theta)$ will be denoted as $d_n = (s + 1)(2k_n - 1)$ both here and later in the proofs.

Lemma 1 (for tail condition) *There exists a constant $r > 0$, such that*

$$\pi_n(\mathcal{F}_n^c) < \exp(-nr)$$

for all sufficiently large n . Here \mathcal{F}_n is the limited part of the k_n -experts family defined in Section 3.1.

Lemma 2 (for entropy condition) *Consider the family of square-root densities $\mathcal{F}^* = \{\sqrt{f} : f \in \mathcal{F}_n\}$ defined in Section 3.1. Then the following relations hold for the Hellinger bracket entropy $H_{[\]}(\cdot)$ for \mathcal{F}^* :*

a). $H_{[\]}(u) \leq \ln \left[\left(\frac{4C_n^2 d_n}{u} \right)^{d_n} \right]$

b). *There exists a constant $c > 0$, such that $\forall \varepsilon > 0$,*

$$\int_0^\varepsilon \sqrt{H_{[\]}(u)} du \leq c\sqrt{n}\varepsilon^2$$

for all sufficiently large n .

Lemma 3 (for approximation condition) *For all $\gamma, \nu > 0$, there exists an N_2 , such that $\pi_n(\text{KL}_\gamma) \geq \exp(-n\nu)$, $\forall n \geq N_2$. Here KL_γ is the Kullback-Leibler neighborhood defined in Section 3.1.*

The following lemma holds whether or not the number of experts is random.

Using notation in Sections 2.2 and 2.3, we have:

Lemma 4 (regression function vs. density function)

a). $\int (\hat{\mu}_n - \mu_0)^2 d\mathbf{x} \leq 4D_H^2(\hat{f}_n, f_0);$

b). $D_H^2(\hat{f}_n, f_0) \leq \varepsilon^2 + 4\pi_n[\{f : D_H(f, f_0) > \varepsilon\} | (Y_i, X_i)^n], \forall \varepsilon > 0.$

The next proposition is used to form the nested sequence of restricted models in Section 3.2, for proving consistency with random number of experts.

Proposition 3 (*Nesting*). *Let $\mathcal{G}_m = \mathbf{\Pi}_m \cap \{f : |\theta_l| < Cm, \forall 1 \leq l \leq \dim(\theta)\}$ for some $C \geq 1$ not dependent on m , which is a restricted set of m -expert models with parameters bounded by Cm . If $m' \geq m$, then $\mathcal{G}_m \subseteq \mathcal{G}_{m'}$. Here $\mathbf{\Pi}_m$ is the m -expert family defined in Section 2.1.*

The proofs of these results are contained in the Appendix.

5 Conclusions

Our work shows that Bayesian inference based on mixtures of logistic regression can be a reliable tool for estimating the regression function and the joint density, as well as for binary classification. We expect that analogous properties may be studied in multiway classification, where multinomial logistic regression models are mixed. This, as well as Bayesian inference based on mixtures of generalized linear models (such as mixtures of Poisson and Gamma regression), form natural topics for future research. So far, we have focused on classification rules of the form $\hat{C}_n(\mathbf{x}) = I[\hat{\mu}_n(\mathbf{x}) > 1/2]$. However, as a referee points out, the concept of C-consistency can also be extended to rules of the form $\hat{C}_n(\mathbf{x}) = I[\hat{\mu}_n(\mathbf{x}) > r]$ for some $r \in (0, 1)$, which may be useful in situations with asymmetric costs, e.g., when misclassifying $Y = 1$ as 0 costs more than misclassifying $Y = 0$ as 1.

A long-standing question in mixtures-of-experts theory is the selection of number of experts (or mixing components) K . The current work provides insight from

the view of Bayesian inference, either from choosing a nonstochastic sequence $K = k_n$ dependent on sample size n , or from treating K as random and placing a suitable prior on K . The latter approach is especially interesting, since it can generate a posterior distribution on K conditional on the observed data: $\pi(K|data) = \int_{\theta} \pi(K, d\theta|data)$. This method of inference on K is in some sense robust and protective against model misspecification: it does not need to assume a true model with number of experts k_0 . The true model is a nonparametric one with arbitrary smooth relation in family Φ . In general there is no ‘true number of experts’ for K . What are proposed by $\pi(K|data)$ are ‘good K ’s instead of ‘true K ’— they are the K ’s for some good approximating models from the mixtures of experts family.

It may also be interesting to consider random K with a *finite* prior distribution, with support increasing with n . This in some sense is combining the approach of the two theorems. The motivation is that we would like the number of experts K to be random in order to search over a range of values. On the other hand, we would like K to be not too large, in order to reduce computation. (Large K would correspond to a high-dimensional parametric model.) There are several possibilities leading to consistent Bayesian inference. One can use a truncated prior $P[K = k] = P[k(I) = k]I[k \leq B_n]/P[k(I) \leq B_n]$, $k = 1, 2, \dots$. Here $k(I)$ satisfies conditions (iii) and (iv) of Theorem 2 and can be, e.g., the contracted Poisson or contracted Geometric random variables described in Remark 1. The truncation bound can be taken to be, e.g., $B_n = 2\lceil(cn)^{1/q}\rceil + 1$, where $q > 1$ is the same as in condition (iv) and $c \in (0, 1]$. One can also use a uniform prior, e.g., $P[K = k] = \lceil(cn)^a\rceil^{-1}I[k \leq \lceil(cn)^a\rceil]$, $k = 1, 2, \dots$, for some $a \in (0, 1)$, $c \in (0, 1]$.

Both can easily be shown to lead to consistent Bayesian inference, by adapting the proof of Theorem 2.

Appendix: Secondary Propositions and Proofs

Denote $f = f(y|\mathbf{x}; k, \theta)$ for a mixture-of- k -experts (conditional) density. Then the following two propositions hold for any (k, θ) and for any $(y, \mathbf{x}) \in \{0, 1\} \times [0, 1]^s$, which will be useful later.

Proposition 4 $\sqrt{f} \leq 1$.

Proposition 5 $\left| \frac{\partial \ln f}{\partial \theta_l} \right| \leq 1$, where θ_l is the l^{th} element of θ , for each $l = 1, \dots, \dim(\theta)$.

The following lemma will be used for proving Lemma 3.

Lemma 5 *Let f be the mixture-of-experts model with parameters $(\theta_1, \dots, \theta_{d_n})$ and \tilde{f} be another mixture-of-experts model with parameters $(\tilde{\theta}_1, \dots, \tilde{\theta}_{d_n})$. Suppose that the number of experts of f and \tilde{f} are both k_n , where k_n grows to infinity with n and $k_n \leq n^a$ for some $0 < a < 1$, for all large enough n . Define a δ -neighborhood of f*

$$M_\delta^n(f) = \{\tilde{f} : |\theta_i - \tilde{\theta}_i| \leq \delta, i = 1, 2, \dots, d_n\}.$$

Then the following holds for any $\gamma > 0$: Given any $f_0 \in \Phi$ (the ‘smooth nonparametric’ family defined in Section 2.1), for all sufficiently large n , there exist δ and f such that $M_\delta^n(f) \subseteq KL_\gamma$ (i.e., for any $\tilde{f} \in M_\delta^n(f)$, we have $D_K(\tilde{f}, f_0) \leq \gamma$),

where $\delta = \frac{\gamma}{4(s+1)n^a}$ and f is a k_n -expert density with parameter components satisfying $\max_{k=1}^{d_n} |\theta_k| \leq c(\gamma) + \ln k_n$, for some constant $c(\gamma)$ depending on γ but not on n .

Proof of Proposition 4

$$\sqrt{f} = \sqrt{\sum_{j=1}^k g_j H_j} \leq \sqrt{\sup_j H_j (\sum_j g_j)} = \sup_j \sqrt{H_j} \leq 1,$$

since $H_j = \left(\frac{e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}}{1 + e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}} \right)^y \left(\frac{1}{1 + e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}} \right)^{1-y} \leq 1.$ □

Proof of Proposition 5

Note that for each $l = 1, \dots, \dim(\theta)$,

$$\left| \frac{\partial \ln f}{\partial \theta_l} \right| = \left| \frac{\partial}{\partial \theta_l} \ln \left(\sum_{j=1}^k g_j H_j \right) \right| = \left| \frac{\sum_j \left[\frac{\partial}{\partial \theta_l} \ln(g_j H_j) \right] (g_j H_j)}{\sum_j g_j H_j} \right| \leq \sup_j \left| \frac{\partial \ln(g_j H_j)}{\partial \theta_l} \right|.$$

Since for each $j = 1, \dots, k$,

$$\ln(g_j H_j) = (u_j + \mathbf{v}_j^T \mathbf{x}) - \ln \left(\sum_j e^{u_j + \mathbf{v}_j^T \mathbf{x}} \right) + y(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}) - \ln(1 + e^{\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}}),$$

it is easy to show that $\left| \frac{\partial \ln(g_j H_j)}{\partial \theta_l} \right| \leq \max\{|x_1|, \dots, |x_s|, 1\} = 1$. So, $\left| \frac{\partial \ln f}{\partial \theta_l} \right| \leq 1.$ □

Proof of Lemma 1

$$\begin{aligned} \pi_n(\mathcal{F}_n^c) &= \Pr(\text{at least one } |\theta_l| > C_n, l = 1, \dots, d_n) \\ &\leq \sum_{l=1}^{d_n} \Pr(|\theta_l| > C_n) = 2 \sum_{l=1}^{d_n} \Pr\left(\frac{\theta_l}{\sigma} > \frac{C_n}{\sigma}\right) \\ &\leq \frac{2d_n \sigma}{C_n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{C_n^2}{2\sigma^2}\right\} \quad \text{by Mill's ratio} \\ &\leq \exp\left\{-\frac{n^{1+2\eta}}{2\sigma^2} + \ln[(s+1)(2n^a - 1)(2\sigma/\sqrt{2\pi})]\right\} \quad \text{noting } k_n \leq n^a \\ &\leq \exp(-nr) \end{aligned}$$

for any $r > 0$, for all sufficiently large n , since $\eta > 0$. \square

Proof of Lemma 2a) Use $f_t = f(y, \mathbf{x}|k, t)$ to simplify the notation, while showing the dependence on a parameter valued at t . Use $\|t\|_\infty = \sup_{j=1}^{\dim(t)} |t_j|$ to denote the L_∞ norm for a vector t .

$$\begin{aligned} |\sqrt{f_t} - \sqrt{f_s}| &= \left| \sum_{l=1}^{d_n} \frac{\partial}{\partial \theta_l} \sqrt{f_\theta} \cdot (t_l - s_l) \right| \quad (\theta \text{ is an intermediate point between } t \text{ and } s) \\ &= \left| \sum_{l=1}^{d_n} (t_l - s_l) \left(\frac{\partial}{\partial \theta_l} \ln \sqrt{f_\theta} \right) \sqrt{f_\theta} \right| \leq \sum_{l=1}^{d_n} \sup_l |t_l - s_l| \cdot \left| \frac{1}{2} \frac{\partial \ln f_\theta}{\partial \theta_l} \right| \cdot |\sqrt{f_\theta}| \\ &\leq \frac{1}{2} d_n \|t - s\|_\infty \quad \text{by Propostions 4 and 5.} \end{aligned}$$

Since C_n grows with n such that $n^{\frac{1}{2}+\eta} \leq C_n \leq \exp(n^{b-a})$, so $C_n \geq 1$. Then, $|\sqrt{f_t} - \sqrt{f_s}| \leq \|t - s\|_\infty \cdot \frac{C_n d_n}{2}$. Let $F(x, y) = C_n d_n / 2$. By Theorem 3 and Eqn (15) of Lee (2000), we have

$$N_{[\cdot]}(2\varepsilon \|F\|_2, \mathcal{F}^*, \|\cdot\|_2) \leq N(\varepsilon, \mathcal{F}_n, L_\infty) \leq \left(\frac{C_n + 1}{\varepsilon} \right)^{d_n}$$

Here, $N(\varepsilon, \mathcal{F}_n, \|\cdot\|)$ is the covering number, i.e. the minimal number of balls of radius ε that are required to cover the set \mathcal{F}_n under a specified metric. Now,

$2\varepsilon \|F\|_2 = 2\varepsilon \sqrt{\sum_{y=0}^1 \int_{\Omega} (C_n d_n / 2)^2 d\mathbf{x}} = \sqrt{2} \varepsilon C_n d_n$, replace $2\varepsilon \|F\|_2$ with ε , then

$$N_{[\cdot]}(\varepsilon, \mathcal{F}^*, \|\cdot\|_2) \leq \left(\frac{C_n + 1}{\varepsilon / (\sqrt{2} C_n d_n)} \right)^{d_n} = \left(\frac{\sqrt{2} C_n (C_n + 1) d_n}{\varepsilon} \right)^{d_n} \leq \left(\frac{4C_n^2 d_n}{\varepsilon} \right)^{d_n}.$$

Therefore, $H_{[\cdot]}(u) = \ln N_{[\cdot]}(u, \mathcal{F}^*, \|\cdot\|_2) \leq \ln \left(\frac{4C_n^2 d_n}{u} \right)^{d_n}$.

Proof of Lemma 2b)

By the result of Lemma 2a),

$$\int_0^\varepsilon \sqrt{H_{[\cdot]}(u)} du \leq \int_0^\varepsilon \sqrt{\ln \left(\frac{4C_n^2 d_n}{u} \right)^{d_n}} du$$

$$\begin{aligned}
&= \sqrt{d_n} \int_{\infty}^{v(\varepsilon)} \frac{v}{\sqrt{2}} \left(-4C_n^2 d_n v e^{-v^2/2} \right) dv \quad \text{where } v(u) = \sqrt{2 \ln \frac{4C_n^2 d_n}{u}} \\
&= 4C_n^2 d_n \sqrt{d_n/2} \left[\frac{\varepsilon}{4C_n^2 d_n} v(\varepsilon) + \int_{v(\varepsilon)}^{\infty} e^{-v^2/2} dv \right] \\
&\leq C_n^2 d_n \sqrt{d_n/2} \left[\frac{\varepsilon}{C_n^2 d_n} \sqrt{2 \ln(4C_n^2 d_n/\varepsilon)} + 4\sqrt{2\pi} \frac{\phi(\sqrt{2 \ln(4C_n^2 d_n/\varepsilon)})}{\sqrt{2 \ln(4C_n^2 d_n/\varepsilon)}} \right] \\
&= \varepsilon \sqrt{d_n/2} \sqrt{2 \ln(4C_n^2 d_n/\varepsilon)} \left[1 + \frac{1}{2 \ln(4C_n^2 d_n/\varepsilon)} \right] \\
&\leq 2\varepsilon \sqrt{d_n} \sqrt{\ln C_n^2 + \ln(4d_n) - \ln \varepsilon} \quad \text{for all large enough } n.
\end{aligned}$$

Noting that $d_n = (s+1)(2k_n - 1) \leq 2(s+1)n^a$, and $C_n \leq \exp(n^{b-a})$, we have

$$l.h.s. \leq 2\varepsilon \sqrt{2(s+1)n^a} \sqrt{2n^{b-a} + \ln 8(s+1) + \ln n^a - \ln \varepsilon}.$$

Since $0 < a < b < 1$, then $\exists t$ such that $0 < a < t < b < 1$ and $b - a < 1 - t$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \int_0^\varepsilon \sqrt{H_{[\cdot]}(u)} du \leq 2\varepsilon \sqrt{n^{-t}} \sqrt{2(s+1)n^a} \sqrt{n^{-(1-t)}} \sqrt{2n^{b-a} + \ln 8(s+1) + \ln n^a - \ln \varepsilon} \\
&\rightarrow 0 \quad \text{as } n \rightarrow \infty.
\end{aligned}$$

So, $\exists c > 0$ such that $\forall \varepsilon > 0$, $\int_0^\varepsilon \sqrt{H_{[\cdot]}(u)} du \leq c\sqrt{n}\varepsilon^2$ for all sufficiently large n . \square

Proof of Lemma 3

We use the neighborhood $M_\delta = M_\delta^n(f)$ in Lemma 5 to prove the result. By Lemma 5, we have $M_\delta \subseteq KL_\gamma$ for all sufficiently large n . Then,

$$\begin{aligned}
\pi_n(KL_\gamma) &\geq \pi_n(M_\delta) = \Pr_{\theta} \left\{ \bigcap_{l=1}^{d_n} |\theta_l - \tilde{\theta}_l| \leq \delta \right\} = \prod_{l=1}^{d_n} \int_{\theta_l - \delta}^{\theta_l + \delta} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u^2}{2\sigma^2}\right) du \\
&\geq \prod_{l=1}^{d_n} \frac{2\delta}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(|\theta_l| + \delta)^2}{2\sigma^2}\right) \geq \prod_{l=1}^{d_n} \frac{2\delta}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(c(\gamma) + \ln k_n + \delta)^2}{2\sigma^2}\right]
\end{aligned}$$

$$\begin{aligned}
&= \exp \left[-d_n \left\{ \frac{(c(\gamma) + \ln k_n + \delta)^2}{2\sigma^2} + \ln \left(\frac{\sqrt{2\pi\sigma^2}}{2\delta} \right) \right\} \right] \\
&\geq \exp \left[-n^a 2(s+1) \left\{ \frac{(2 \ln n^a)^2}{2\sigma^2} \right\} \right] \quad \text{for all large enough } n, \text{ using } k_n \leq n^a, \delta = \frac{\gamma}{4(s+1)n^a} \\
&\geq \exp(-n\nu) \quad \text{for all large enough } n, \text{ fixing any } \nu > 0, \text{ since } a \in (0, 1).
\end{aligned}$$

□

Proof of Lemma 4a)

$$\begin{aligned}
&\int (\hat{\mu}_n - \mu_0)^2 d\mathbf{x} = \int \left[\int y(\hat{f}_n - f_0) dy \right]^2 d\mathbf{x} = \int \left[\int y(\sqrt{\hat{f}_n} + \sqrt{f_0})(\sqrt{\hat{f}_n} - \sqrt{f_0}) dy \right]^2 d\mathbf{x} \\
&\leq \int \left[\int y^2(\sqrt{\hat{f}_n} + \sqrt{f_0})^2 dy \right] \left[\int (\sqrt{\hat{f}_n} - \sqrt{f_0})^2 dy \right] d\mathbf{x}
\end{aligned}$$

since

$$\begin{aligned}
&\int y^2(\sqrt{\hat{f}_n} + \sqrt{f_0})^2 dy = \int y^2(\hat{f}_n + f_0 + 2\sqrt{\hat{f}_n f_0}) dy \\
&\leq 2 \int y^2(\hat{f}_n + f_0) dy = 2 \sum_{y=0}^1 y^2(\hat{f}_n + f_0) \leq 4 \quad \text{by Proposition 4.}
\end{aligned}$$

$$\text{Then, } \int (\hat{\mu}_n - \mu_0)^2 d\mathbf{x} \leq 4 \int \int (\sqrt{\hat{f}_n} - \sqrt{f_0})^2 dy d\mathbf{x} = 4D_H^2(\hat{f}_n, f_0). \quad \square$$

Proof of Lemma 4b)

Denote $\hat{f}_n = \int f \pi_n(d\theta | (Y_i, X_i)^n) = E_{\theta| \cdot} f$. Denote $A_\epsilon = \{f : D_H(f, f_0) \leq \epsilon\}$ as in Section 2.3. Then,

$$\begin{aligned}
&D_H^2(\hat{f}_n, f_0) = \int \int (\sqrt{E_{\theta| \cdot} f} - \sqrt{f_0})^2 dy d\mathbf{x} \\
&= \int \int (f_0 + E_{\theta| \cdot} f - 2\sqrt{f_0 E_{\theta| \cdot} f}) dy d\mathbf{x} = 2 - 2 \int \int \sqrt{E_{\theta| \cdot} (f f_0)} dy d\mathbf{x} \\
&\leq 2 - 2 \int \int E_{\theta| \cdot} (\sqrt{f f_0}) dy d\mathbf{x} \quad \text{by Jensen's inequality} \\
&= E_{\theta| \cdot} (2 - 2 \int \int \sqrt{f f_0} dy d\mathbf{x}) \quad \text{by Fubini's Theorem}
\end{aligned}$$

$$\begin{aligned}
&= E_{\theta_1} \int (f + f_0 - 2\sqrt{ff_0}) dy d\mathbf{x} = E_{\theta_1} \int \int (\sqrt{f} - \sqrt{f_0})^2 dy d\mathbf{x} = E_{\theta_1} D_H^2(f, f_0) \\
&= \int D_H^2(f, f_0) \pi_n(d\theta | (Y_i, X_i)^n) \\
&= \int_{A_\varepsilon} D_H^2(f, f_0) \pi_n(d\theta | (Y_i, X_i)^n) + \int_{A_\varepsilon^c} D_H^2(f, f_0) \pi_n(d\theta | (Y_i, X_i)^n) \\
&\leq \varepsilon^2 + \int_{A_\varepsilon^c} \left[\int \int (\sqrt{f} - \sqrt{f_0})^2 dy d\mathbf{x} \right] \pi_n(d\theta | (Y_i, X_i)^n) \\
&= \varepsilon^2 + \int_{A_\varepsilon^c} \left(\int \int (f + f_0 - 2\sqrt{ff_0}) dy d\mathbf{x} \right) \pi_n(d\theta | (Y_i, X_i)^n) \\
&\leq \varepsilon^2 + 2 \int_{A_\varepsilon^c} \left[\int \int (f + f_0) dy d\mathbf{x} \right] \pi_n(d\theta | (Y_i, X_i)^n) \\
&= \varepsilon^2 + 4 \int_{A_\varepsilon^c} \pi_n(d\theta | (Y_i, X_i)^n) \\
&= \varepsilon^2 + 4\pi_n(\{f : D_H(f, f_0) > \varepsilon\} | (Y_i, X_i)^n)
\end{aligned}$$

It is easy to see that Lemmas 4(a,b) actually hold also for the case with random number of experts k , by augmenting the integration over $d\theta$ with sum over k . \square

Proof of Lemma 5

Jiang and Tanner (1999b, Theorem 2) state that $\sup_{f_0 \in \Phi} \inf_{f \in \Pi_k} D_K(f, f_0) \leq \frac{c}{k^{4/s}}$ for some $c > 0$ independent of k , for each $k = 1, 2, 3, \dots$. Here, $s = \dim(\mathbf{x})$. Therefore, given any $\gamma > 0$, for any true model $f_0 \in \Phi$, there exists a k^* -experts model $f^* \in \Pi_{k^*}$, with k^* large enough, such that

$$D_K(f^*, f_0) \leq \frac{c}{(k^*)^{4/s}} + \frac{\gamma}{4} < \gamma/2.$$

This k^* -experts density f^* can be written as a k_n -experts density f ($k_n > k^*$ for all large enough n) if we let

$$\begin{cases} u_j = u_j^* & , \quad j = 1, \dots, k^* - 1 \\ u_j = u_{k^*}^* - \ln(k_n - k^* + 1) & , \quad j = k^*, \dots, k_n \end{cases}$$

and

$$\begin{cases} (\alpha_j, \boldsymbol{\beta}_j, \mathbf{v}_j) = (\alpha_j^*, \boldsymbol{\beta}_j^*, \mathbf{v}_j^*) & , \quad j = 1, \dots, k^* - 1 \\ (\alpha_j, \boldsymbol{\beta}_j, \mathbf{v}_j) = (\alpha_{k^*}^*, \boldsymbol{\beta}_{k^*}^*, \mathbf{v}_{k^*}^*) & , \quad j = k^*, \dots, k_n. \end{cases}$$

Here u 's, \mathbf{v} 's, α 's and $\boldsymbol{\beta}$'s are components of parameter θ for density f ; u^* 's, \mathbf{v}^* 's, α^* 's and $\boldsymbol{\beta}^*$'s are components of parameter θ^* for density f^* . This parametrization for embedding is explained in the proof of Proposition 3. This implies that there also exists a k_n -experts model f such that

$$D_K(f, f_0) = D_K(f^*, f_0) < \gamma/2,$$

for all sufficiently large n . Let θ^* and θ denote the vectors of parameters in the k^* -experts and k_n -experts model, respectively. From the above parameter settings, we have

$$\|\theta\|_\infty \leq \max\{\|\theta^*\|_\infty, |u_{k^*}^*| + \ln(k_n - k^* + 1)\} \leq c(\gamma) + \ln k_n$$

for some constant $c(\gamma)$ possibly dependent on γ .

Now consider any k_n -expert model $\tilde{f} \in M_\delta^n(f)$. Note that

$$D_K(\tilde{f}, f_0) = \int f_0 \ln \frac{f_0}{\tilde{f}} = \int f_0 \ln \left(\frac{f_0}{f} \cdot \frac{f}{\tilde{f}} \right) = \int f_0 \ln \frac{f_0}{f} + \int f_0 \ln \frac{f}{\tilde{f}}.$$

The first term $\int f_0 \ln \frac{f_0}{f} = D_K(f, f_0) < \frac{\gamma}{2}$ for all sufficiently large n . For the second part,

$$\begin{aligned} \ln \frac{f}{\tilde{f}} &= \ln f - \ln \tilde{f} \\ &\leq \sum_{l=1}^{d_n} \left| \frac{\partial}{\partial u_l} \ln f_u \right| |\theta_l - \tilde{\theta}_l| \quad (u \text{ is an intermediate point between } \theta \text{ and } \tilde{\theta}) \\ &\leq \delta \sum_{l=1}^{d_n} \left| \frac{\partial \ln f_u}{\partial u_l} \right| \quad \text{since } \tilde{f} \in M_\delta^n(f) \end{aligned}$$

$$\begin{aligned}
&\leq d_n \delta \quad \text{by Proposition 5} \\
&\leq \gamma/2 \quad \text{noting that } \delta = \frac{\gamma}{4(s+1)n^a} < \frac{\gamma}{2d_n}.
\end{aligned}$$

Then $\int f_0 \ln \frac{f}{f_0} \leq \gamma/2$. Therefore, $D_K(\tilde{f}, f_0) \leq \gamma$. Therefore, $M_\delta^n(f) \subseteq KL_\gamma$. \square

Proof of Proposition 3

We need to show $\forall f \in \mathcal{G}_m, f \in \mathcal{G}_{m'}$.

$\forall f \in \mathcal{G}_m$,

$$\begin{aligned}
f(y, \mathbf{x}) &= \sum_{j=1}^{m-1} g_j H_j + \frac{e^{u_m + \mathbf{v}_m^T \mathbf{x}}}{\sum_{l=1}^{m-1} e^{u_l + \mathbf{v}_l^T \mathbf{x}} + e^{u_m + \mathbf{v}_m^T \mathbf{x}}} H_m \\
&= \sum_{j=1}^{m-1} g_j H_j + \frac{\sum_{l=m}^{m'} \frac{1}{(m'-m+1)} e^{u_m + \mathbf{v}_m^T \mathbf{x}}}{\sum_{l=1}^{m-1} e^{u_l + \mathbf{v}_l^T \mathbf{x}} + \sum_{l=m}^{m'} \frac{1}{(m'-m+1)} e^{u_m + \mathbf{v}_m^T \mathbf{x}}} H_m \\
&= \sum_{j=1}^{m-1} g_j H_j + \sum_{l=m}^{m'} \frac{e^{\tilde{u}_m + \tilde{\mathbf{v}}_m^T \mathbf{x}}}{\sum_{l=1}^{m'} e^{\tilde{u}_l + \tilde{\mathbf{v}}_l^T \mathbf{x}}} H_m
\end{aligned}$$

This is an m' -experts model ($m' \geq m$) with parameters

$$\begin{cases} \tilde{u}_j = u_j & , \quad j = 1, \dots, m-1 \\ \tilde{u}_j = u_m - \ln(m' - m + 1) & , \quad j = m, \dots, m' \end{cases}$$

and

$$\begin{cases} (\tilde{\alpha}_j, \tilde{\beta}_j, \tilde{\mathbf{v}}_j) = (\alpha_j, \beta_j, \mathbf{v}_j) & , \quad j = 1, \dots, m-1 \\ (\tilde{\alpha}_j, \tilde{\beta}_j, \tilde{\mathbf{v}}_j) = (\alpha_m, \beta_m, \mathbf{v}_m) & , \quad j = m, \dots, m'. \end{cases}$$

Note that for $j = m, \dots, m'$,

$$\begin{aligned}
|\tilde{u}_j| &= |u_m - \ln(m' - m + 1)| \\
&\leq |u_m| + \ln(m' - m + 1) \quad m' - m + 1 \geq 1 \\
&\leq Cm + (m' - m) \\
&\leq Cm + C(m' - m) \quad \text{since } C \geq 1 \\
&= Cm'
\end{aligned}$$

The other parameters are bounded in absolute value by Cm and hence bounded by Cm' . Therefore, $f \in \mathcal{G}_{m'}$, proving $\mathcal{G}_m \subseteq \mathcal{G}_{m'}$. \square

Acknowledgments

The authors wish to thank the two referees for useful suggestions on improving the paper.

References

- Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*. **3**, 79-87.
- Jiang, W. and Tanner, M. A. (1999a). On the approximation rate of hierarchical mixtures-of-experts for generalized linear models', *Neural Computation*. **11**, 1183-1198.
- Jiang, W. and Tanner, M. A. (1999b). 'Hierarchical Mixtures-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation'. *Annals of Statistics*. **27**, 987-1011.
- Jordan, M. I., and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*. **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.
- Lee, H. K. H. (2000). Consistency of posterior distributions for neural networks. *Neural Networks*. **13**, 629-642.
- 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. *Journal of American Statistical Association*. **91**, 953-960.
- 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. *Journal of American Statistical Association*. **91**, 953-960.
- Wood, S. A., Kohn, R., Jiang, W. and Tanner, M. A. (2005). Spatially adaptive nonparametric binary regression using a mixture of probits. *Technical Report, Department of Statistics, Northwestern University*. (Downloadable at <http://newton.stats.northwestern.edu/~jiang/report/binary.probit.pdf>.)