NOTE **Communicated by Eduardo Mendes**

A Universal Approximation Theorem for Mixture-of-Experts Models

Hien D. Nguyen
hien1988@gmail.com
School of Mathematics and Physics, University of Queensland, Brisbane, Queensland 4072, Australia

Luke R. Lloyd-Jones
l.lloydjones@uq.edu.au
Centre for Neurogenetics and Statistical Genetics, Queensland Brain Institute, University of Queensland, Brisbane, Queensland 4072, Australia

Geoffrey J. McLachlan
g.mclachlan@uq.edu.au
School of Mathematics and Physics, University of Queensland, Brisbane, Queensland 4072, Australia

The mixture-of-experts (MoE) model is a popular neural network architecture for nonlinear regression and classification. The class of MoE mean functions is known to be uniformly convergent to any unknown target function, assuming that the target function is from a Sobolev space that is sufficiently differentiable and that the domain of estimation is a compact unit hypercube. We provide an alternative result, which shows that the class of MoE mean functions is dense in the class of all continuous functions over arbitrary compact domains of estimation. Our result can be viewed as a universal approximation theorem for MoE models. The theorem we present allows MoE users to be confident in applying such models for estimation when data arise from nonlinear and nondifferentiable generative processes.

1 Introduction

The mixture-of-experts (MoE) model is a neural network architecture for nonlinear regression and classification. The model was introduced in Jacobs, Jordan, Nowlan, and Hinton (1991) and Jordan and Jacobs (1994); reviews can be found in McLachlan and Peel (2000) and Yuksel, Wilson, and Gader (2012). Recent research includes Chamroukhi, Glotin, and Same (2013) and Nguyen and McLachlan (2016), where MoE models are used for curve classification and robust estimation, respectively.

Neural Computation **28, 2585–2593** (2016) © 2016 Massachusetts Institute of Technology
doi:10.1162/NECO_a_00892

Let $Y \in \mathbb{R}$ be a random variable and $\boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^p$ be a p-dimensional vector. Let the conditional probability density function of Y given $\boldsymbol{x}$ be

$$f_g(y|\boldsymbol{x}) = \sum_{i=1}^{g} \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g) h(y; \gamma_i + \boldsymbol{x}^T \boldsymbol{\delta}_i, \boldsymbol{\nu}_i), \tag{1.1}$$

where

$$\pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g) = \frac{\exp\left(\alpha_i + \boldsymbol{x}^T \boldsymbol{\beta}_i\right)}{\sum_{j=1}^{g} \exp\left(\alpha_j + \boldsymbol{x}^T \boldsymbol{\beta}_j\right)},$$

y is a realization of Y, and $h(y; \mu, \boldsymbol{\xi})$ is a univariate component probability density function (PDF) (in Y) with mean μ and nuisance parameter $\boldsymbol{\xi}$. Here $\alpha_i, \gamma_i \in \mathbb{R}$, $\boldsymbol{\beta}_i, \boldsymbol{\delta}_i \in \mathbb{R}^p$, and $\boldsymbol{\nu}_i \in \mathbb{R}^q$ for each $i = 1, \ldots, g$, and $\boldsymbol{\psi}_g^T = (\alpha_1, \boldsymbol{\beta}_1^T, \ldots, \alpha_g, \boldsymbol{\beta}_g^T)$. We say that equation 1.1 is a g-component MoE with mean function

$$\mu_g(\boldsymbol{x}; \boldsymbol{\theta}_g) = \sum_{i=1}^{g} \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g)(\gamma_i + \boldsymbol{x}^T \boldsymbol{\delta}_i), \tag{1.2}$$

where $\boldsymbol{\theta}_g^T = (\boldsymbol{\psi}_g^T, \gamma_1, \boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\psi}_g^T, \gamma_g, \boldsymbol{\delta}_g^T)$ is the function's parameter vector. The superscript T indicates matrix transposition.

Zeevi, Meir, and Maiorov (1998) showed that there exists a sequence of functions $\mu_g(\boldsymbol{x}; \boldsymbol{\theta}_g)$ that converges uniformly to any target function $m(\boldsymbol{x})$, in the index g, assuming that $m(\boldsymbol{x})$ belongs to a Sobolev class of functions and $\mathbb{X}$ is a closed unit hypercube. Their result was generalized to nonlinear mappings of the expression $\gamma_i + \boldsymbol{x}^T \boldsymbol{\delta}_i$ in the component PDFs of the MoE in Jiang and Tanner (1999b). The result from Jiang and Tanner (1999b) was expanded on in Jiang and Tanner (1999a), where it was shown that there exists a sequence of conditional PDFs $f_g(y|\boldsymbol{x})$ that converges in Kullback-Leibler divergence to any target-conditional PDF $f(y|\boldsymbol{x})$ in g, assuming that $f(y|\boldsymbol{x})$ belongs to the one-parameter exponential family of density functions; extensions to multivariate conditional density estimation are obtained in Norets (2010). Convergence results for MoE models with polynomial mean functions were obtained in Mendes and Jiang (2012). We note that the target mean function $m(\boldsymbol{x})$ is assumed to belong to a Sobolev class of functions in each of Jiang and Tanner (1999a, 1999b) and Mendes & Jiang (2012), as they are in Zeevi et al. (1998).

Define the class of all mean functions of form 1.2 as

$$\mathbb{M} = \left\{\mu_g(\boldsymbol{x}; \boldsymbol{\theta}_g) | g \in \mathbb{N}, \boldsymbol{\theta}_g \in \mathbb{R}^{g(2p+2)}\right\},$$

and let $\mathbb{C}(\mathbb{X})$ be the class of continuous functions on the domain $\mathbb{X}$. In this note, we prove that $\mathbb{M}$ is dense in the set $\mathbb{C}(\mathbb{X})$ under the assumption that

$\mathbb{X}$ is compact. Our result is obtained via the Stone-Weierstrass theorem (Stone, 1948; see also Cotter, 1990, for a discussion in the context of neural networks). Our result is a universal approximation theorem, similar in spirit to Cybenko (1989, theorem 2), where the linear combination of sigmoidal functions is proved dense in $\mathbb{C}(\mathbb{X})$. We show denseness of approximations to some conditional mean functions, whereas Cybenko (1989) targets a marginal multivariate mean function. Our results allow MoE users to be confident in applying such models for estimation when data arise from nonlinear and nondifferentiable generative processes as they improve on the guarantees of Zeevi et al. (1998).

2 Main Result

Define **0** to be a vector of zeros of an appropriate dimensionality. In order to facilitate the proofs, let

$$\mathbb{H} = \left\{\eta_g(\boldsymbol{x}; \boldsymbol{\omega}_g) | g \in \mathbb{N}, \boldsymbol{\omega}_g \in \mathbb{R}^{g(p+2)}\right\},$$

where

$$\eta_g(\boldsymbol{x}; \boldsymbol{\omega}_g) = \sum_{i=1}^{g} \gamma_i \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g)$$

and $\boldsymbol{\omega}_g^T = (\boldsymbol{\psi}_1^T, \gamma_1, \ldots, \boldsymbol{\psi}_g^T, \gamma_g)$ is the function's parameter vector. Note that

$$\mu_g(\boldsymbol{x}; \tilde{\boldsymbol{\theta}}_g) = \eta_g(\boldsymbol{x}; \boldsymbol{\omega}_g),$$

if $\tilde{\boldsymbol{\theta}}_g^T = (\boldsymbol{\psi}_g^T, \gamma_1, 0^T, \ldots, \boldsymbol{\psi}_g^T, \gamma_g, 0^T)$; thus $\mathbb{H} \subset \mathbb{M}$.

Theorem 1. *The class $\mathbb{H}$ is dense in $\mathbb{C}(\mathbb{X})$. Furthermore, the class $\mathbb{M}$ is dense in $\mathbb{C}(\mathbb{X})$, since $\mathbb{H} \subset \mathbb{M}$.*

3 Comparisons to Zeevi et al. (1998)

Zeevi et al. (1998, theorem 1) proved the class $\mathbb{H}$ dense within the Sobolev class $W_q^r(L)$ over the closed-unit hypercube domain (see Zeevi et al., 1998, for definitions). First, unlike Zeevi et al. (1998), we make no assumptions on the domain $\mathbb{X}$ other than compactness. Second, the target space $\mathbb{C}(\mathbb{X})$ makes no restrictions on differentiability, whereas $W_q^r(L)$ requires the rth partial derivatives to exist. Finally, we do not require the target function or its partial derivatives to be measurable or bounded, whereas Zeevi et al. (1998) require partial derivatives up to order r to be measurable and possess finite L_q norms bounded by L.

Unfortunately, by operating in $\mathbb{C}(\mathbb{X})$ rather than $W_q^r(L)$, we are unable to obtain convergence rates for functions from $\mathbb{H}$ to target functions in $\mathbb{C}(\mathbb{X})$. The convergence rates obtained in Zeevi et al. (1998) are conditional on the differentiability r and the norm order q.

4 Proof of Main Result

The Stone-Weierstrass theorem can be phrased as follows (cf. Cotter, 1990):

Theorem 2. *Let $\mathbb{X} \subset \mathbb{R}^p$ be a compact set and let $\mathbb{U}$ be a set of continuous real-valued functions on $\mathbb{X}$. Assume that*

i. The constant function $u(\boldsymbol{x}) = 1$ is in $\mathbb{U}$.
ii. For any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}$ such that $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, there exists a function $u \in \mathbb{U}$ such that $u(\boldsymbol{x}_1) \neq u(\boldsymbol{x}_2)$.
iii. If $a \in \mathbb{R}$ and $u \in \mathbb{U}$, then $au \in \mathbb{U}$.
iv. If $u, v \in \mathbb{U}$, then $uv \in \mathbb{U}$.
v. If $u, v \in \mathbb{U}$, then $u + v \in \mathbb{U}$.

If assumptions i to v are true, then $\mathbb{U}$ is dense in $\mathbb{C}(\mathbb{X})$. In other words, for any $\epsilon > 0$ and any $v \in \mathbb{C}(\mathbb{X})$, there exists a $u \in \mathbb{U}$ such that $\sup_{\boldsymbol{x}\in\mathbb{X}} |u(\boldsymbol{x}) - v(\boldsymbol{x})| < \epsilon$.

We note that $\mathbb{X}$ is compact if and only if it is bounded and closed in Euclidean spaces (see Dudley, 2004, chap. 2). We proceed to prove that $\mathbb{H}$ is dense in $\mathbb{C}(\mathbb{X})$.

Lemma 1. *The constant function $\eta(\boldsymbol{x}) = 1$ is in $\mathbb{H}$.*

Proof. Let $g = 1$ and $\tilde{\boldsymbol{\omega}}_1^T = (\boldsymbol{\psi}_1^T, 1)$. Set $\eta(\boldsymbol{x}) = \eta_1(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_1)$. For any choice of $\boldsymbol{\psi}_1$, $\eta(\boldsymbol{x}) = 1$. We obtain the result by noting that $\eta_1(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_1) \in \mathbb{H}$.

Lemma 2. *For any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}$ such that $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, there exists a function $\eta \in \mathbb{H}$ such that $\eta(\boldsymbol{x}_1) \neq \eta(\boldsymbol{x}_2)$.*

Proof. Let $g = 2$ and $\bar{\boldsymbol{\omega}}_2^T = (\boldsymbol{\psi}_1^T, 0, 0^T, \gamma_2)$, where $\gamma_2 \neq 0$. Set $\eta(\boldsymbol{x}) = \eta_2(\boldsymbol{x}; \bar{\boldsymbol{\omega}}_2)$, and assume that $\boldsymbol{x}_j^T = (x_{j1}, \ldots, x_{jp})$ for $j = 1, 2$, such that $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$. Let $\eta(\boldsymbol{x}_1) \neq \eta(\boldsymbol{x}_2)$; this is equivalent to

$$\eta_2(\boldsymbol{x}_1; \bar{\boldsymbol{\omega}}_2) \neq \eta_2(\boldsymbol{x}_2; \bar{\boldsymbol{\omega}}_2)$$
$$\frac{\gamma_2}{1 + \exp(\alpha_1 + \boldsymbol{x}_1^T \boldsymbol{\beta}_1)} \neq \frac{\gamma_2}{1 + \exp(\alpha_1 + \boldsymbol{x}_2^T \boldsymbol{\beta}_1)}$$

by substitution and reduces to

$$(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \boldsymbol{\beta}_1 \neq 0. \tag{4.1}$$

Equation 4.1 is violated if either $\boldsymbol{x}_1 = \boldsymbol{x}_2$, which causes a contradiction, or if $\boldsymbol{\beta}_1^T = (\beta_1, \ldots, \beta_p)$ is such that $\beta_k = 0$ whenever $x_{1k} \neq x_{2k}$, for $k = 1, \ldots, p$. To avoid violation of equation 4.1, we can set $\beta_k \neq 0$ for all k.

Thus, let $\tilde{\boldsymbol{\omega}}_2^T = (\tilde{\boldsymbol{\psi}}_1^T, 0, 0^T, \gamma_2)$ and $\eta(\boldsymbol{x}) = \eta_2(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_2)$, where $\tilde{\boldsymbol{\psi}}_1 \in (\mathbb{R}\backslash\{0\})^{p+1}$ and $\gamma_2 \neq 0$. If $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, then $\eta(\boldsymbol{x}_1) \neq \eta(\boldsymbol{x}_2)$. We obtain the result by noting that $\eta_2(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_2) \in \mathbb{H}$.

Lemma 3. *If $a \in \mathbb{R}$ and $\eta \in \mathbb{H}$, then $a\eta \in \mathbb{H}$.*

Proof. Let $a \in \mathbb{R}$ and $\eta(\boldsymbol{x}) = \eta_g(\boldsymbol{x}; \boldsymbol{\omega}_g)$. We can write

$$\begin{aligned} a\eta(\boldsymbol{x}) &= a\sum_{i=1}^{g} \gamma_i \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g) \\ &= \sum_{i=1}^{g} (a\gamma_i)\pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g) \\ &= \sum_{i=1}^{g} \tilde{\gamma}_i \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g), \end{aligned}$$

where $\tilde{\gamma}_i = a\gamma_i$ for $i = 1, \ldots, g$. Thus, $a\eta(\boldsymbol{x}) = \eta_g(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_g)$, where $\tilde{\boldsymbol{\omega}}_g^T = (\boldsymbol{\psi}_1^T, \tilde{\gamma}_1, \ldots, \boldsymbol{\psi}_g^T, \tilde{\gamma}_g)$. We obtain the result by noting that $\eta_g(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_g) \in \mathbb{H}$.

Lemma 4. *If $\eta, \lambda \in \mathbb{H}$, then $\eta\lambda \in \mathbb{H}$.*

Proof. Let $g, m \in \mathbb{N}$,

$$\boldsymbol{\omega}_g^{[\eta]T} = \left(\boldsymbol{\psi}_1^{[\eta]T}, \gamma_1^{[\eta]}, \ldots, \boldsymbol{\psi}_g^{[\eta]T}, \gamma_g^{[\eta]}\right)$$

and

$$\boldsymbol{\omega}_m^{[\lambda]T} = \left(\boldsymbol{\psi}_1^{[\lambda]T}, \gamma_1^{[\lambda]}, \ldots, \boldsymbol{\psi}_g^{[\lambda]T}, \gamma_g^{[\lambda]}\right),$$

and set $\eta(\boldsymbol{x}) = \eta_g(\boldsymbol{x}; \boldsymbol{\omega}_g^{[\eta]})$ and $\lambda(\boldsymbol{x}) = \eta_m(\boldsymbol{x}; \boldsymbol{\omega}_m^{[\lambda]})$. Here, the superscripts $[\eta]$ and $[\lambda]$ denote the parameter components belonging to the functions η and λ, respectively. We can write

$$\begin{aligned} \eta(\boldsymbol{x})\lambda(\boldsymbol{x}) &= \sum_{i=1}^{g} \gamma_i^{[\eta]} \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g^{[\eta]}) \sum_{j=1}^{m} \gamma_j^{[\lambda]} \pi_j(\boldsymbol{x}; \boldsymbol{\psi}_m^{[\lambda]}) \\ &= \sum_{i=1}^{g} \sum_{j=1}^{m} \gamma_i^{[\eta]} \gamma_j^{[\lambda]} \pi_i(\boldsymbol{x}; \boldsymbol{\psi}_g^{[\eta]}) \pi_j(\boldsymbol{x}; \boldsymbol{\psi}_m^{[\lambda]}). \end{aligned} \tag{4.2}$$

To simplify equation 4.2, for each $i = 1, \ldots, g$ and $j = 1, \ldots, m$, we can write

$$\begin{aligned} &\pi_i\big(\boldsymbol{x}; \boldsymbol{\psi}_g^{[\eta]}\big)\pi_j\big(\boldsymbol{x}; \boldsymbol{\psi}_m^{[\lambda]}\big) \qquad (4.3) \\ &= \frac{\exp\big(\alpha_i^{[\eta]} + \boldsymbol{x}^T\boldsymbol{\beta}_i^{[\eta]}\big)\exp\big(\alpha_j^{[\lambda]} + \boldsymbol{x}^T\boldsymbol{\beta}_j^{[\lambda]}\big)}{\sum_{k=1}^{g}\exp\big(\alpha_k^{[\eta]} + \boldsymbol{x}^T\boldsymbol{\beta}_k^{[\eta]}\big)\sum_{l=1}^{m}\exp\big(\alpha_l^{[\lambda]} + \boldsymbol{x}^T\boldsymbol{\beta}_l^{[\lambda]}\big)} \\ &= \frac{\exp\big(\big[\alpha_i^{[\eta]} + \alpha_j^{[\lambda]}\big] + \boldsymbol{x}^T\big[\boldsymbol{\beta}_i^{[\eta]} + \boldsymbol{\beta}_j^{[\lambda]}\big]\big)}{\sum_{k=1}^{g}\sum_{l=1}^{m}\exp\big(\big[\alpha_k^{[\eta]} + \alpha_l^{[\lambda]}\big] + \boldsymbol{x}^T\big[\boldsymbol{\beta}_k^{[\eta]} + \boldsymbol{\beta}_l^{[\lambda]}\big]\big)}. \end{aligned}$$

On performing the mapping from Table 1A, we can write the final line of equation 4.3 as $\pi_k(\boldsymbol{x}; \tilde{\boldsymbol{\psi}}_{gm})$, where $\tilde{\boldsymbol{\psi}}_k^T = (\tilde{\alpha}_k, \tilde{\boldsymbol{\beta}}_k^T)$ for $k = 1, \ldots, gm$. Furthermore, via the mapping from Table 1A, equation 4.2 can be simplified to

$$\begin{aligned} \eta(\boldsymbol{x})\lambda(\boldsymbol{x}) &= \sum_{k=1}^{gm}\tilde{\gamma}_k\pi_k\big(\boldsymbol{x}; \tilde{\boldsymbol{\psi}}_{gm}\big) \\ &= \eta_{gm}\big(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_{gm}\big), \end{aligned}$$

where $\tilde{\boldsymbol{\omega}}_{gm}^T = (\tilde{\boldsymbol{\psi}}_1^T, \tilde{\gamma}_1, \ldots, \tilde{\boldsymbol{\psi}}_{gm}^T, \tilde{\gamma}_{gm})$. We obtain the result by noting that $\eta_{gm}(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_{gm}) \in \mathbb{H}$.

Lemma 5. *If $\eta, \lambda \in \mathbb{H}$, then $\eta + \lambda \in \mathbb{H}$.*

Proof. Let $g, m \in \mathbb{N}$,

$$\boldsymbol{\omega}_g^{[\eta]T} = \big(\boldsymbol{\psi}_1^{[\eta]T}, \gamma_1^{[\eta]}, \ldots, \boldsymbol{\psi}_g^{[\eta]T}, \gamma_g^{[\eta]}\big),$$

and

$$\boldsymbol{\omega}_m^{[\lambda]T} = \big(\boldsymbol{\psi}_1^{[\lambda]T}, \gamma_1^{[\lambda]}, \ldots, \boldsymbol{\psi}_g^{[\lambda]T}, \gamma_g^{[\lambda]}\big),$$

and set $\eta(\boldsymbol{x}) = \eta_g(\boldsymbol{x}; \boldsymbol{\omega}_g^{[\eta]})$ and $\lambda(\boldsymbol{x}) = \eta_m(\boldsymbol{x}; \boldsymbol{\omega}_m^{[\lambda]})$. Here, the superscripts $[\eta]$ and $[\lambda]$ denote the parameter components belonging to the functions η and λ, respectively. We can write

$$\begin{aligned} &\eta(\boldsymbol{x}) + \lambda(\boldsymbol{x}) \\ &= \sum_{i=1}^{g}\gamma_i^{[\eta]}\pi_i\big(\boldsymbol{x}; \boldsymbol{\psi}_g^{[\eta]}\big) + \sum_{j=1}^{m}\gamma_j^{[\lambda]}\pi_j\big(\boldsymbol{x}; \boldsymbol{\psi}_m^{[\lambda]}\big) \\ &= \frac{\sum_{i=1}^{g}\gamma_i^{[\eta]}\exp\big(\alpha_i^{[\eta]} + \boldsymbol{x}^T\boldsymbol{\beta}_i^{[\eta]}\big)}{\sum_{k=1}^{g}\exp\big(\alpha_k^{[\eta]} + \boldsymbol{x}^T\boldsymbol{\beta}_k^{[\eta]}\big)} \end{aligned}$$

Table 1: Mapping of Parameter Components for Lemmas 4 and 5.

A. Lemma 4								
$\alpha_1^{[\eta]} + \alpha_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_1$	$\boldsymbol{\beta}_1^{[\eta]} + \boldsymbol{\beta}_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_1$	$\gamma_1^{[\eta]}\gamma_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_1$
$\alpha_1^{[\eta]} + \alpha_2^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_2$	$\boldsymbol{\beta}_1^{[\eta]} + \boldsymbol{\beta}_2^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_2$	$\gamma_1^{[\eta]}\gamma_2^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_2$
$\vdots$			$\vdots$			$\vdots$		
$\alpha_1^{[\eta]} + \alpha_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{m+1}$	$\boldsymbol{\beta}_1^{[\eta]} + \boldsymbol{\beta}_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{m+1}$	$\gamma_1^{[\eta]}\gamma_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{m+1}$
$\alpha_2^{[\eta]} + \alpha_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{m+2}$	$\boldsymbol{\beta}_2^{[\eta]} + \boldsymbol{\beta}_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{m+2}$	$\gamma_2^{[\eta]}\gamma_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{m+2}$
$\vdots$			$\vdots$			$\vdots$		
$\alpha_k^{[\eta]} + \alpha_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{km+1}$	$\boldsymbol{\beta}_k^{[\eta]} + \boldsymbol{\beta}_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{km+1}$	$\gamma_k^{[\eta]}\gamma_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{km+1}$
$\alpha_{k+1}^{[\eta]} + \alpha_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{km+2}$	$\boldsymbol{\beta}_{k+1}^{[\eta]} + \boldsymbol{\beta}_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{km+2}$	$\gamma_{k+1}^{[\eta]}\gamma_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{km+2}$
$\vdots$			$\vdots$			$\vdots$		
$\alpha_g^{[\eta]} + \alpha_{m-1}^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{gm-1}$	$\boldsymbol{\beta}_g^{[\eta]} + \boldsymbol{\beta}_{m-1}^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{gm-1}$	$\gamma_g^{[\eta]}\gamma_{m-1}^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{gm-1}$
$\alpha_g^{[\eta]} + \alpha_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{gm}$	$\boldsymbol{\beta}_g^{[\eta]} + \boldsymbol{\beta}_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{gm}$	$\gamma_g^{[\eta]}\gamma_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{gm}$
B. Lemma 5								
$\alpha_1^{[\eta]} + \alpha_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_1$	$\boldsymbol{\beta}_1^{[\eta]} + \boldsymbol{\beta}_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_1$	$\gamma_1^{[\eta]} + \gamma_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_1$
$\alpha_1^{[\eta]} + \alpha_2^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_2$	$\boldsymbol{\beta}_1^{[\eta]} + \boldsymbol{\beta}_2^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_2$	$\gamma_1^{[\eta]} + \gamma_2^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_2$
$\vdots$			$\vdots$			$\vdots$		
$\alpha_1^{[\eta]} + \alpha_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{m+1}$	$\boldsymbol{\beta}_1^{[\eta]} + \boldsymbol{\beta}_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{m+1}$	$\gamma_1^{[\eta]} + \gamma_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{m+1}$
$\alpha_2^{[\eta]} + \alpha_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{m+2}$	$\boldsymbol{\beta}_2^{[\eta]} + \boldsymbol{\beta}_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{m+2}$	$\gamma_2^{[\eta]} + \gamma_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{m+2}$
$\vdots$			$\vdots$			$\vdots$		
$\alpha_k^{[\eta]} + \alpha_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{km+1}$	$\boldsymbol{\beta}_k^{[\eta]} + \boldsymbol{\beta}_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{km+1}$	$\gamma_k^{[\eta]} + \gamma_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{km+1}$
$\alpha_{k+1}^{[\eta]} + \alpha_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{km+2}$	$\boldsymbol{\beta}_{k+1}^{[\eta]} + \boldsymbol{\beta}_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{km+2}$	$\gamma_{k+1}^{[\eta]} + \gamma_1^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{km+2}$
$\vdots$			$\vdots$			$\vdots$		
$\alpha_g^{[\eta]} + \alpha_{m-1}^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{gm-1}$	$\boldsymbol{\beta}_g^{[\eta]} + \boldsymbol{\beta}_{m-1}^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{gm-1}$	$\gamma_g^{[\eta]} + \gamma_{m-1}^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{gm-1}$
$\alpha_g^{[\eta]} + \alpha_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\alpha}_{gm}$	$\boldsymbol{\beta}_g^{[\eta]} + \boldsymbol{\beta}_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\boldsymbol{\beta}}_{gm}$	$\gamma_g^{[\eta]} + \gamma_m^{[\lambda]}$	$\longrightarrow$	$\tilde{\gamma}_{gm}$

$$+\frac{\sum_{j=1}^{m}\gamma_j^{[\lambda]}\exp\left(\alpha_j^{[\lambda]}+\boldsymbol{x}^T\boldsymbol{\beta}_j^{[\lambda]}\right)}{\sum_{l=1}^{m}\exp\left(\alpha_l^{[\lambda]}+\boldsymbol{x}^T\boldsymbol{\beta}_l^{[\lambda]}\right)}$$

$$=\frac{\sum_{i=1}^{g}\gamma_i^{[\eta]}\exp\left(\alpha_i^{[\eta]}+\boldsymbol{x}^T\boldsymbol{\beta}_i^{[\eta]}\right)\sum_{l=1}^{m}\exp\left(\alpha_j^{[\lambda]}+\boldsymbol{x}^T\boldsymbol{\beta}_j^{[\lambda]}\right)}{\sum_{k=1}^{g}\exp\left(\alpha_k^{[\eta]}+\boldsymbol{x}^T\boldsymbol{\beta}_k^{[\eta]}\right)\sum_{l=1}^{m}\exp\left(\alpha_l^{[\lambda]}+\boldsymbol{x}^T\boldsymbol{\beta}_l^{[\lambda]}\right)}$$

$$+\frac{\sum_{j=1}^{m}\gamma_j^{[\lambda]}\exp\left(\alpha_j^{[\lambda]}+\boldsymbol{x}^T\boldsymbol{\beta}_j^{[\lambda]}\right)\sum_{k=1}^{g}\exp\left(\alpha_i^{[\eta]}+\boldsymbol{x}^T\boldsymbol{\beta}_i^{[\eta]}\right)}{\sum_{k=1}^{g}\exp\left(\alpha_k^{[\eta]}+\boldsymbol{x}^T\boldsymbol{\beta}_k^{[\eta]}\right)\sum_{l=1}^{m}\exp\left(\alpha_l^{[\lambda]}+\boldsymbol{x}^T\boldsymbol{\beta}_l^{[\lambda]}\right)}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{g}\sum_{l=1}^{m}\gamma_i^{[\eta]}\exp\left([\alpha_i^{[\eta]}+\alpha_j^{[\lambda]}]+\boldsymbol{x}^T\left[\boldsymbol{\beta}_i^{[\eta]}+\boldsymbol{\beta}_j^{[\lambda]}\right]\right)}{\sum_{k=1}^{g}\sum_{l=1}^{m}\exp\left(\left[\alpha_k^{[\eta]}+\alpha_l^{[\lambda]}\right]\boldsymbol{x}^T\left[\boldsymbol{\beta}_k^{[\eta]}+\boldsymbol{\beta}_l^{[\lambda]}\right]\right)} \\
&\quad + \frac{\sum_{i=1}^{g}\sum_{j=1}^{m}\gamma_j^{[\lambda]}\exp\left(\left[\alpha_i^{[\eta]}+\alpha_j^{[\lambda]}\right]+\boldsymbol{x}^T\left[\boldsymbol{\beta}_i^{[\eta]}+\boldsymbol{\beta}_j^{[\lambda]}\right]\right)}{\sum_{k=1}^{g}\sum_{l=1}^{m}\exp\left(\left[\alpha_k^{[\eta]}+\alpha_l^{[\lambda]}\right]+\boldsymbol{x}^T\left[\boldsymbol{\beta}_k^{[\eta]}+\boldsymbol{\beta}_l^{[\lambda]}\right]\right)} \\
&= \frac{\sum_{i=1}^{g}\sum_{l=1}^{m}\left(\gamma_i^{[\eta]}+\gamma_j^{[\lambda]}\right)\exp\left(\left[\alpha_i^{[\eta]}+\alpha_j^{[\lambda]}\right]+\boldsymbol{x}^T\left[\boldsymbol{\beta}_i^{[\eta]}+\boldsymbol{\beta}_j^{[\lambda]}\right]\right)}{\sum_{k=1}^{g}\sum_{l=1}^{m}\exp\left(\left[\alpha_k^{[\eta]}+\alpha_l^{[\lambda]}\right]+\boldsymbol{x}^T\left[\boldsymbol{\beta}_k^{[\eta]}+\boldsymbol{\beta}_l^{[\lambda]}\right]\right)}. \qquad (4.4)
\end{aligned}
$$

On performing the mapping from Table 1B, we can write equation 4.4 as

$$
\begin{aligned}
\eta(\boldsymbol{x})\lambda(\boldsymbol{x}) &= \frac{\sum_{k=1}^{gm}\tilde{\gamma}_k\exp(\tilde{\alpha}_k+\boldsymbol{x}^T\tilde{\boldsymbol{\beta}}_k)}{\sum_{l=1}^{gm}\exp(\tilde{\alpha}_l+\boldsymbol{x}^T\tilde{\boldsymbol{\beta}}_l)} \\
&= \sum_{k=1}^{gm}\tilde{\gamma}_k\pi_k(\boldsymbol{x};\tilde{\boldsymbol{\psi}}_{gm}) \\
&= \eta_{gm}(\boldsymbol{x};\tilde{\boldsymbol{\omega}}_{gm}),
\end{aligned}
$$

where $\tilde{\boldsymbol{\omega}}_{gm}^T = (\tilde{\boldsymbol{\psi}}_1^T, \tilde{\gamma}_1, \ldots, \tilde{\boldsymbol{\psi}}_{gm}^T, \tilde{\gamma}_{gm})$ and $\tilde{\boldsymbol{\psi}}_k^T = (\tilde{\alpha}_k, \tilde{\boldsymbol{\beta}}_k^T)$ for $k = 1, \ldots, gm$. We obtain the result by noting that $\eta_{gm}(\boldsymbol{x}; \tilde{\boldsymbol{\omega}}_{gm}) \in \mathbb{H}$.

Lemmas 1 to 5 imply that the class $\mathbb{H}$ satisfies Assumptions (i)–(v) of Theorem 2; thus Theorem 1 is proved.

5 Conclusion

In this note, we utilized the Stone-Weierstrass theorem to prove that the class of MoE mean functions $\mathbb{M}$ is dense in the class of continuous functions $\mathbb{C}\,(\mathbb{X})$ on the compact domain $\mathbb{X}$.

Unlike in Zeevi et al. (1998), Jiang and Tanner (1999a, 1999b), and Mendes and Jiang (2012), we do not obtain convergence rates. Furthermore, our result does not guarantee statistical estimability of the MoE mean functions. Maximum likelihood (ML) estimation can obtain consistent estimates for mean functions, when g is known (see Zeevi et al., 1998; Jiang & Tanner, 2000; and Nguyen & McLachlan, 2016). Results regarding regularized ML estimation of MoE models were obtained in Khalili (2010). In Grun and Leisch (2007) and Nguyen and McLachlan (2016), the Bayesian information criterion (BIC; Schwarz (1978)) is shown effective for determination of unknown g (see Olteanu & Rynkiewicz, 2011, for theoretical justification of the BIC).

References

Chamroukhi, F., Glotin, H., & Same, A. (2013). Model-based functional mixture discriminant analysis with hidden process regression for curve classification. *Neurocomputing*, *112*, 153–163.

Cotter, N. E. (1990). The Stone-Weierstrass theorem and its application to neural networks. *IEEE Transactions on Neural Networks*, *1*, 290–295.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314.

Dudley, R. M. (2004). *Real analysis and probability*. Cambridge: Cambridge University Press.

Grun, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics and Data Analysis*, *51*, 5247–5252.

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

Jiang, W., & Tanner, M. A. (1999a). Hierachical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, *27*, 987–1011.

Jiang, W., & Tanner, M. A. (1999b). On the approximation rate of hierachical mixtures-of-experts for generalized linear models. *Neural Computation*, *11*, 1183–1198.

Jiang, W., & Tanner, M. A. (2000). On the asymptotic normality of hierachical mixtures-of-experts for generalized linear models. *IEEE Transactions on Information Theory*, *46*, 1005–1013.

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

Khalili, A. (2010). New estimation and feature selection methods in mixture-of-experts models. *Canadian Journal of Statistics*, *38*, 519–539.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mendes, E. F., & Jiang, W. (2012). On convergence rates of mixture of polynomial experts. *Neural Computation*, *24*, 3025–3051.

Nguyen, H. D., & McLachlan, G. J. (2016). Laplace mixture of linear experts. *Computational Statistics and Data Analysis*, *93*, 177–191.

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *Annals of Statistics*, *38*, 1733–1766.

Olteanu, M., & Rynkiewicz, J. (2011). Asymptotic properties of mixture-of-experts models. *Neurocomputing*, *74*, 1444–1449.

Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, *6*, 461–464.

Stone, M. H. (1948). The generalized Weierstrass approximation theorem. *Mathematical Magazine*, *21*, 237–254.

Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, *23*, 1177–1193.

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

Received May 18, 2016; accepted July 12, 2016.