Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models

Hien Duy Nguyen[1*] , TrungTin Nguyen[2], Faicel Chamroukhi[2] and Geoffrey John McLachlan[3]

*Correspondence:
h.nguyen5@latrobe.edu.au
[1]Department of Mathematics and
Statistics, La Trobe University,
Bundoora Victoria, Australia
Full list of author information is
available at the end of the article

Abstract

Mixture of experts (MoE) models are widely applied for conditional probability density estimation problems. We demonstrate the richness of the class of MoE models by proving denseness results in Lebesgue spaces, when inputs and outputs variables are both compactly supported. We further prove an almost uniform convergence result when the input is univariate. Auxiliary lemmas are proved regarding the richness of the soft-max gating function class, and their relationships to the class of Gaussian gating functions.

Keywords: Mixture of experts, Conditional probability density functions, Approximation theory, Mixture models, Lebesgue spaces

1 Introduction

Mixture of experts (MoE) models are a widely applicable class of conditional probability density approximations that have been considered as solution methods across the spectrum of statistical and machine learning Yuksel et al. (2012); Masoudnia and Ebrahimpour (2014); Nguyen and Chamroukhi (2018).

Let $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X} \subseteq \mathbb{R}^d$ and $\mathbb{Y} \subseteq \mathbb{R}^q$, for $d, q \in \mathbb{N}$. Suppose that the input and output random variables, $X \in \mathbb{X}$ and $Y \in \mathbb{Y}$, are related via the conditional probability density function (PDF) $f(y|x)$ in the functional class:

$$\mathcal{F} = \left\{ f : \mathbb{Z} \to [0, \infty) \mid \int_{\mathbb{Y}} f(y|x) \, d\lambda(y) = 1, \forall x \in \mathbb{X} \right\},$$

where λ denotes the Lebesgue measure. The MoE approach seeks to approximate the unknown target conditional PDF f by a function of the MoE form:

$$m(y|x) = \sum_{k=1}^{K} \text{Gate}_k(x) \, \text{Expert}_k(y),$$

where **Gate** $= (\text{Gate}_k)_{k \in [K]} \in \mathcal{G}^K$ ($[K] = \{1, \dots, K\}$), $\text{Expert}_1, \dots, \text{Expert}_K \in \mathcal{E}$, and $K \in \mathbb{N}$. Here, we say that m is a K-component MoE model with gates arising from the class $\mathcal{G}^K$ and experts arising from $\mathcal{E}$, where $\mathcal{E}$ is a class of PDFs with support $\mathbb{Y}$.

The most popular choices for $\mathcal{G}^K$ are the parametric soft-max and Gaussian gating classes:

$$\mathcal{G}_S^K = \left\{ \mathbf{Gate} = (\text{Gate}_k\,(\cdot;\boldsymbol{\gamma}))_{k\in[K]} \,|\forall k \in [K]\,, \text{Gate}_k\,(\cdot;\boldsymbol{\gamma}) = \frac{\exp\left(a_k + \boldsymbol{b}_k^\top \cdot\right)}{\sum_{l=1}^K \exp\left(a_l + \boldsymbol{b}_l^\top \cdot\right)}, \boldsymbol{\gamma} \in \mathbb{G}_S^K \right\}$$

and

$$\mathcal{G}_G^K = \left\{ \mathbf{Gate} = (\text{Gate}_k\,(\cdot;\boldsymbol{\gamma}))_{k\in[K]} \,|\forall k \in [K]\,, \text{Gate}_k\,(\cdot;\boldsymbol{\gamma}) = \frac{\pi_k \phi\,(\cdot;\boldsymbol{v}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \phi\,(\cdot;\boldsymbol{v}_l, \boldsymbol{\Sigma}_l)}, \boldsymbol{\gamma} \in \mathbb{G}_G^K \right\},$$

respectively, where

$$\mathbb{G}_S^K = \left\{ \boldsymbol{\gamma} = (a_1, \dots, a_K, \boldsymbol{b}_1, \dots, \boldsymbol{b}_K) \in \mathbb{R}^K \times \left(\mathbb{R}^d\right)^K \right\}$$

and

$$\mathbb{G}_G^K = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K) \in \Pi_{K-1} \times \left(\mathbb{R}^d\right)^K \times \mathbb{S}_d^K \right\}.$$

Here,

$$\phi\,(\cdot;\boldsymbol{v}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\cdot - \boldsymbol{v})^\top \boldsymbol{\Sigma}^{-1}(\cdot - \boldsymbol{v})\right]$$

is the multivariate normal density function with mean vector $\boldsymbol{v}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}^\top = (\pi_1, \dots, \pi_K)$ is a vector of weights in the simplex:

$$\Pi_{K-1} = \left\{ \boldsymbol{\pi} = (\pi_k)_{k\in[K]} \,|\forall k \in [K]\,, \pi_k > 0, \sum_{k=1}^K \pi_k = 1 \right\},$$

and $\mathbb{S}_d$ is the class of $d \times d$ symmetric positive definite matrices. The soft-max and Gaussian gating classes were first introduced by Jacobs et al. (1991); Jordan and Xu (1995), respectively. Typically, one chooses experts that arise from some location-scale class:

$$\mathcal{E}_\psi = \left\{ g_\psi\,(\cdot;\boldsymbol{\mu}, \sigma) : \mathbb{Y} \to [0, \infty) \,|g_\psi\,(\cdot;\boldsymbol{\mu}, \sigma) = \frac{1}{\sigma^q} \psi\left(\frac{\cdot - \boldsymbol{\mu}}{\sigma}\right), \boldsymbol{\mu} \in \mathbb{R}^q, \sigma \in (0, \infty) \right\},$$

where ψ is a PDF, with respect to $\mathbb{R}^q$ in the sense that $\psi : \mathbb{R}^q \to [0, \infty)$ and $\int_{\mathbb{R}^q} \psi\,(\boldsymbol{y})\,d\lambda\,(\boldsymbol{y}) = 1$.

We shall say that $f \in \mathcal{L}_p\,(\mathbb{Z})$ for any $p \in [1, \infty)$ if

$$\|f\|_{p,\mathbb{Z}} = \left(\int_{\mathbb{Z}} |\mathbf{1}_{\mathbb{Z}} f|^p\,d\lambda\,(\boldsymbol{z})\right)^{1/p} < \infty,$$

where $\mathbf{1}_{\mathbb{Z}}$ is the indicator function that takes value 1 when $\boldsymbol{z} \in \mathbb{Z}$, and 0 otherwise. Further, we say that $f \in \mathcal{L}_\infty\,(\mathbb{Z})$ if

$$\|f\|_{\infty,\mathbb{Z}} = \inf\left\{ a \geq 0 |\lambda\,\left(\left\{\boldsymbol{z} \in \mathbb{Z}| \,|f\,(\boldsymbol{z})| > a\right\}\right) = 0 \right\} < \infty.$$

We shall refer to $\|\cdot\|_{p,\mathbb{Z}}$ as the $\mathcal{L}_p$ norm on $\mathbb{Z}$, for $p \in [0, \infty]$, and where the context is obvious, we shall drop the reference to $\mathbb{Z}$.

Suppose that the target conditional PDF f is in the class $\mathcal{F}_p = \mathcal{F} \cap \mathcal{L}_p$. We address the problem of approximating f, with respect to the $\mathcal{L}_p$ norm, using MoE models in the soft-max and Gaussian gated classes,

$$\mathcal{M}_S^\psi = \left\{ m_K^\psi : \mathbb{Z} \to [0,\infty) \,|\, m_K^\psi \left(\boldsymbol{y}|\boldsymbol{x}\right) = \sum_{k=1}^K \text{Gate}_k \left(\boldsymbol{x}\right) g_\psi \left(\boldsymbol{y}; \boldsymbol{\mu}_k, \sigma_k\right), \right.$$
$$\left. g_\psi \in \mathcal{E}_\psi \cap \mathcal{L}_\infty, \textbf{Gate} \in \mathcal{G}_S^K, \boldsymbol{\mu}_k \in \mathbb{Y}, \sigma_k \in (0,\infty), k \in [K], K \in \mathbb{N} \right\}$$

and

$$\mathcal{M}_G^\psi = \left\{ m_K^\psi : \mathbb{Z} \to [0,\infty) \,|\, m_K^\psi \left(\boldsymbol{y}|\boldsymbol{x}\right) = \sum_{k=1}^K \text{Gate}_k \left(\boldsymbol{x}\right) g_\psi \left(\boldsymbol{y}; \boldsymbol{\mu}_k, \sigma_k\right), \right.$$
$$\left. g_\psi \in \mathcal{E}_\psi \cap \mathcal{L}_\infty, \textbf{Gate} \in \mathcal{G}_G^K, \boldsymbol{\mu}_k \in \mathbb{Y}, \sigma_k \in (0,\infty), k \in [K], K \in \mathbb{N} \right\},$$

by showing that both $\mathcal{M}_S^\psi$ and $\mathcal{M}_G^\psi$ are dense in the class $\mathcal{F}_p$, when $\mathbb{X} = [0,1]^d$ and $\mathbb{Y}$ is a compact subset of $\mathbb{R}^q$. Our denseness results are enabled by the indicator function approximation result of Jiang and Tanner (1999a), and the finite mixture model denseness theorems of Nguyen et al. (2020); Nguyen et al. (2020a).

Our theorems contribute to an enduring continuity of sustained interest in the approximation capabilities of MoE models. Related to our results are contributions regarding the approximation capabilities of the conditional expectation function of the classes $\mathcal{M}_S^\psi$ and $\mathcal{M}_G^\psi$ (Wang and Mendel 1992; Zeevi et al. 1998; Jiang and Tanner 1999a; Krzyzak and Schafer 2005; Mendes and Jiang 2012; Nguyen et al. 2016; Nguyen et al. 2019) and the approximation capabilities of subclasses of $\mathcal{M}_S^\psi$ and $\mathcal{M}_G^\psi$, with respect to the Kullback–Leibler divergence (Jiang and Tanner 1999b; Norets 2010; Norets and Pelenis 2014). Our results can be seen as complements to the Kullback–Leibler approximation theorems of Norets (2010); Norets and Pelenis (2014), by the relationship between the Kullback–Leibler divergence and the $\mathcal{L}_2$ norm (Zeevi and Meir 1997). That is, when $f > 1/\kappa$, for all $\left(\boldsymbol{x}, \boldsymbol{y}\right) \in \mathbb{Z}$ and some constant $\kappa > 0$, we have that the integrated conditional Kullback–Leibler divergence considered by Norets and Pelenis (2014):

$$\int_{\mathbb{X}} D\left(f\left(\cdot|\boldsymbol{x}\right) \| m_K^\psi \left(\cdot|\boldsymbol{x}\right)\right) \mathrm{d}\lambda\left(\boldsymbol{x}\right) = \int_{\mathbb{X}} \int_{\mathbb{Y}} f\left(\boldsymbol{y}|\boldsymbol{x}\right) \log \frac{f\left(\boldsymbol{y}|\boldsymbol{x}\right)}{m_K^\psi \left(\boldsymbol{y}|\boldsymbol{x}\right)} \mathrm{d}\lambda\left(\boldsymbol{y}\right) \mathrm{d}\lambda\left(\boldsymbol{x}\right)$$

satisfies

$$\int_{\mathbb{X}} D\left(f\left(\cdot|\boldsymbol{x}\right) \| m_K^\psi \left(\cdot|\boldsymbol{x}\right)\right) \mathrm{d}\lambda\left(\boldsymbol{x}\right) \leq \kappa^2 \left\| f - m_K^\psi \right\|_{2,\mathbb{Z}}^2,$$

and thus a good approximation in the integrated Kullback–Leibler divergence is guaranteed if one can find a good approximation in the $\mathcal{L}_2$ norm, which is guaranteed by our main result.

The remainder of the manuscript proceeds as follows. The main result is presented in Section 2. Technical lemmas are provided in Section 3. The proofs of our results are then presented in Section 4. Proofs of required lemmas that do not appear elsewhere are provided in Section 5. A summary of our work and some conclusions are drawn in Section 6.

2 Main results

Denote the class of bounded functions on $\mathbb{Z}$ by

$$\mathcal{B}\left(\mathbb{Z}\right) = \left\{ f \in \mathcal{L}_\infty\left(\mathbb{Z}\right) \,|\, \exists a \in [0,\infty), \text{ such that } \left|f\left(\boldsymbol{z}\right)\right| \leq a, \forall \boldsymbol{z} \in \mathbb{Z} \right\},$$

and write its norm as $\|f\|_{\mathcal{B}(\mathbb{Z})} = \sup_{z \in \mathbb{Z}} |f(z)|$. Further, let $\mathcal{C}$ denote the class of continuous functions. Note that if $\mathbb{Z}$ is compact and $f \in \mathcal{C}$, then $f \in \mathcal{B}$.

Theorem 1 *Assume that* $\mathbb{X} = [0,1]^d$ *for* $d \in \mathbb{N}$. *There exists a sequence* $\left\{ m_K^{\psi} \right\}_{K \in \mathbb{N}} \subset \mathcal{M}_S^{\psi}$, *such that if* $\mathbb{Y} \subset \mathbb{R}^q$ *is compact,* $f \in \mathcal{F} \cap \mathcal{C}$, *and* $\psi \in \mathcal{C}(\mathbb{R}^q)$ *is a PDF on support* $\mathbb{R}^q$, *then* $\lim_{K \to \infty} \left\| f - m_K^{\psi} \right\|_p = 0$, *for* $p \in [1, \infty)$.

Since convergence in Lebesgue spaces does not imply point-wise modes of convergence, the following result is also useful and interesting in some restricted scenarios. Here, we note that the mode of convergence is almost uniform, which implies almost everywhere convergence and convergence in measure (cf. Bartle 1995, Lem 7.10 and Thm. 7.11). The almost uniform convergence of $\left\{ m_K^{\psi} \right\}_{K \in \mathbb{N}}$ to f in the following result is to be understood in the sense of Bartle (1995), Def. 7.9. That is, for every $\delta > 0$, there exists a set $\mathbb{E}_\delta \subset \mathbb{Z}$ with $\lambda(\mathbb{Z}) < \delta$, such that $\left\{ m_K^{\psi} \right\}_{K \in \mathbb{N}}$ converges to f, uniformly on $\mathbb{Z} \backslash \mathbb{E}_\delta$.

Theorem 2 *Assume that* $\mathbb{X} = [0,1]$. *There exists a sequence* $\left\{ m_K^{\psi} \right\}_{K \in \mathbb{N}} \subset \mathcal{M}_S^{\psi}$, *such that if* $\mathbb{Y} \subset \mathbb{R}^q$ *is compact,* $f \in \mathcal{F} \cap \mathcal{C}$, *and* $\psi \in \mathcal{C}(\mathbb{R}^q)$ *is a PDF on support* $\mathbb{R}^q$, *then* $\lim_{K \to \infty} m_K^{\psi} = f$, *almost uniformly.*

The following result establishes the connection between the gating classes $\mathcal{G}_S^K$ and $\mathcal{G}_G^K$.

Lemma 1 *For each* $K \in \mathbb{N}$, $\mathcal{G}_S^K \subset \mathcal{G}_G^K$. *Further, if we define the class of Gaussian gating vectors with equal covariance matrices:*

$$\mathcal{G}_E^K = \left\{ \mathbf{Gate} = (Gate_k(\cdot; \boldsymbol{\gamma}))_{k \in [K]} \, | \forall k \in [K], Gate_k(\cdot; \boldsymbol{\gamma}) = \frac{\pi_k \phi(\cdot; \boldsymbol{v}_k, \boldsymbol{\Sigma})}{\sum_{l=1}^K \pi_l \phi(\cdot; \boldsymbol{v}_l, \boldsymbol{\Sigma})}, \boldsymbol{\gamma} \in \mathbb{G}_E^K \right\},$$

where

$$\mathbb{G}_E^K = \left\{ \boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_K, \boldsymbol{\Sigma}) \in \Pi_{K-1} \times \left(\mathbb{R}^d \right)^K \times \mathbb{S}_d \right\},$$

then $\mathcal{G}_E^K \subset \mathcal{G}_S^K$.

We can directly apply Lemma 1 to establish the following corollary to Theorems 1 and 2, regarding the approximation capability of the class $\mathcal{M}_G^{\psi}$.

Corollary 1 *Theorems 1 and 2 hold when* $\mathcal{M}_S^{\psi}$ *is replaced by* $\mathcal{M}_G^{\psi}$ *in their statements.*

3 Technical lemmas

Let $\mathbb{K}^n = \left\{ (k_1, \ldots, k_d) \in [n]^d \right\}$ and $\kappa : \mathbb{K}^n \to [n^d]$ be a bijection for each $n \in \mathbb{N}$. For each $(k_1, \ldots, k_d) \in \mathbb{K}^n$ and $k \in [n^d]$, we define $\mathbb{X}_k^n = \mathbb{X}_{\kappa(k_1, \ldots, k_d)}^n = \prod_{i=1}^d \mathbb{I}_{k_i}^n$, where $\mathbb{I}_{k_i}^n = [(k_i - 1)/n, k_i/n)$ for $k_i \in [n-1]$, and $\mathbb{I}_n^n = [(n-1)/n, 1]$.

We call $\left\{ \mathbb{X}_k^n \right\}_{k \in [n^d]}$ a fine partition of $\mathbb{X}$, in the sense that $\mathbb{X} = [0,1]^d = \bigcup_{k=1}^{n^d} \mathbb{X}_k^n$, for each n, and that $\lambda \left(\mathbb{X}_k^n \right) = n^{-d}$ gets smaller, as n increases. The following result from Jiang and Tanner (1999a) establishes the approximation capability of soft-max gates.

Lemma 2 (Jiang and Tanner, 1999, p. 1189) *For each $n \in \mathbb{N}$, $p \in [1, \infty)$ and $\epsilon > 0$, there exists a gating functions*

$$\mathbf{Gate} = (Gate_k \left(\cdot; \boldsymbol{\gamma}\right))_{k \in [n^d]} \in \mathcal{G}_S^{n^d}$$

for some $\boldsymbol{\gamma} \in \mathbb{G}_S^{n^d}$, such that

$$\sup_{k \in [n^d]} \left\| \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} - Gate_k \left(\cdot; \boldsymbol{\gamma}\right) \right\|_{p, \mathbb{X}} \leq \epsilon.$$

When, $d = 1$, we have also the following almost uniform convergence alternative to Lemma 2.

Lemma 3 *Let $\mathbb{X} = [0, 1]$. Then, for each $n \in \mathbb{N}$, there exists a sequence of gating functions:*

$$\left\{ \mathbf{Gate}_l = \left(Gate_k \left(\cdot; \boldsymbol{\gamma}_l\right)\right)_{k \in [n^d]} \right\}_{l \in \mathbb{N}} \subset \mathcal{G}_S^n,$$

defined by $\left\{ \boldsymbol{\gamma}_l \right\}_{l \in \mathbb{N}} \subset \mathbb{G}_S^n$, such that

$$Gate_k \left(\cdot; \boldsymbol{\gamma}_l\right) \to \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}},$$

almost uniformly, simultaneously for all $k \in \left[n^d\right]$.

For PDF ψ on support $\mathbb{R}^q$, define the class of finite mixture models by

$$\mathcal{H}^\psi = \left\{ h_K^\psi : \mathbb{R}^q \to [0, \infty) \, | h_K^\psi \left(\boldsymbol{y}\right) = \sum_{k=1}^{K} c_k g_\psi \left(\boldsymbol{y}; \boldsymbol{\mu}_k, \sigma_k\right), \right.$$

$$\left. g_\psi \in \mathcal{E}_\psi \cap \mathcal{L}_\infty, (c_k)_{k \in [K]} \in \Pi_{K-1}, \boldsymbol{\mu}_k \in \mathbb{Y}, \sigma_k \in (0, \infty), k \in [K], K \in \mathbb{N} \right\}.$$

We require the following result, from Nguyen et al. (2020), regarding the approximation capabilities of $\mathcal{H}^\psi$.

Lemma 4 (Nguyen et al., 2020a, Thm. 2(b)) *If $f \in \mathcal{C}\left(\mathbb{Y}\right)$ is a PDF on $\mathbb{Y}$, $\psi \in \mathcal{C}\left(\mathbb{R}^q\right)$ is a PDF on $\mathbb{R}^q$, and $\mathbb{Y} \subset \mathbb{R}^q$ is compact, then there exists a sequence $\left\{ h_K^\psi \right\}_{K \in \mathbb{N}} \subset \mathcal{H}^\psi$, such that $\lim_{K \to \infty} \left\| f - h_K^\psi \right\|_{\mathcal{B}(\mathbb{Y})} = 0.$*

4 Proofs of main results

4.1 Proof of Theorem 1

To prove the result, it suffices to show that for each $\epsilon > 0$, there exists a $m_K^\psi \in \mathcal{M}_S^\psi$, such that

$$\left\| f - m_K^\psi \right\|_p < \epsilon.$$

The main steps of the proof are as follows. We firstly approximate $f\left(\boldsymbol{y}|\boldsymbol{x}\right)$ by

$$\upsilon_n \left(\boldsymbol{y}|\boldsymbol{x}\right) = \sum_{k=1}^{n^d} \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} f \left(\boldsymbol{y}|\boldsymbol{x}_k^n\right), \tag{1}$$

where $\boldsymbol{x}_k^n \in \mathbb{X}_k^n$, for each $k \in \left[n^d\right]$, such that

$$\left\| f - \upsilon_n \right\|_p < \frac{\epsilon}{3}, \tag{2}$$

for all $n \geq N_1(\epsilon)$, for some sufficiently large $N_1(\epsilon) \in \mathbb{N}$. Then we approximate $\upsilon_n(\boldsymbol{y}|\boldsymbol{x})$ by

$$\eta_n(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{n^d} \text{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) f(\boldsymbol{y}|\boldsymbol{x}_k^n), \tag{3}$$

where $\boldsymbol{\gamma}_n \in \mathbb{G}_S^{n^d}$ and $\textbf{Gate} = (\text{Gate}_k(\cdot; \boldsymbol{\gamma}_n))_{k \in [n^d]} \in \mathcal{G}_S^{n^d}$, so that

$$\|\upsilon_n - \eta_n\|_p \leq \sup_{k \in [n^d]} \left\| \text{Gate}_k(\cdot; \boldsymbol{\gamma}) - \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} \right\|_{p, \mathbb{X}} \sum_{k=1}^{n^d} \left\| f(\cdot|\boldsymbol{x}_k^n) \right\|_{p, \mathbb{Y}} < \frac{\epsilon}{3}, \tag{4}$$

using Lemma 2.

Finally, we approximate $\eta_n(\boldsymbol{y}|\boldsymbol{x})$ by $m_{K_n}^{\psi}(\boldsymbol{y}|\boldsymbol{x})$, where

$$m_{K_n}^{\psi}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{n^d} \text{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}) h_{n_k}^k(\boldsymbol{y}|\boldsymbol{x}_k^n) \tag{5}$$

and

$$h_{n_k}^k(\boldsymbol{y}|\boldsymbol{x}_k^n) = \sum_{i=1}^{n_k} c_i^k g_{\psi}(\boldsymbol{y}; \boldsymbol{\mu}_i^k, \sigma_i^k) \in \mathcal{H}^{\psi} \tag{6}$$

for $n_k \in \mathbb{N}$ ($k \in [n^d]$), such that $K_n = \sum_{k=1}^{n^d} n_k$. Here, we establish that there exists $N_2(\epsilon, n, \boldsymbol{\gamma}_n) \in \mathbb{N}$, so that when $n_k \geq N_2(\epsilon, n, \boldsymbol{\gamma}_n)$,

$$\left\| \eta_n - m_{K_n}^{\psi} \right\|_p \leq \sup_{k \in [n^d]} \| \text{Gate}_k(\cdot; \boldsymbol{\gamma}) \|_{p, \mathbb{X}} \sum_{k=1}^{n^d} \left\| f(\cdot|\boldsymbol{x}_k^n) - h_{n_k}^k(\cdot|\boldsymbol{x}_k^n) \right\|_{p, \mathbb{Y}} < \frac{\epsilon}{3}. \tag{7}$$

Results (2)–(7) then imply that for each $\epsilon > 0$, there exists $N_1(\epsilon)$, $\boldsymbol{\gamma}_n$, and $N_2(\epsilon, n, \boldsymbol{\gamma}_n)$, such that for all $K_n = \sum_{k=1}^{n^d} n_k$, where $n_k \geq N_2(\epsilon, n, \boldsymbol{\gamma}_n)$ (for each $k \in [n^d]$) and $n \geq N_1(\epsilon)$. The following inequality results from an application of the triangle inequality:

$$\left\| f - m_{K_n}^{\psi} \right\|_p \leq \left\| f - \upsilon_n \right\|_p + \| \upsilon_n - \eta_n \|_p + \left\| \eta_n - m_{K_n}^{\psi} \right\|_p < 3 \times \frac{\epsilon}{3} = \epsilon.$$

We now focus our attention to proving each of the results: (2)–(7). To prove (2), we note that since f is uniformly continuous (because $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$ is compact, and $f \in \mathcal{C}$), there exists a function (1) such that for all $\varepsilon > 0$,

$$\sup_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{Z}} \left| f(\boldsymbol{y}|\boldsymbol{x}) - \upsilon(\boldsymbol{y}|\boldsymbol{x}) \right| < \varepsilon. \tag{8}$$

We can construct such an approximation by considering the fact that as n increases, the diameter $\delta_n = \sup_{k \in n^d} \text{diam}(\mathbb{X}_k^n)$ of the fine partition goes to zero. By the uniform continuity of f, for every $\varepsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that if $\left\| (\boldsymbol{x}_1, \boldsymbol{y}_1) - (\boldsymbol{x}_2, \boldsymbol{y}_2) \right\| < \delta(\epsilon)$, then $\left| f(\boldsymbol{y}_1|\boldsymbol{x}_1) - f(\boldsymbol{y}_2|\boldsymbol{x}_2) \right| < \varepsilon$, for all pairs $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in \mathbb{Z}$. Here, $\|\cdot\|$ denotes the Euclidean norm. Furthermore, for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{Z}$, we have

$$\left| f(\boldsymbol{y}|\boldsymbol{x}) - \upsilon_n(\boldsymbol{y}|\boldsymbol{x}) \right| = \left| \sum_{k=1}^{n^d} \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} \left[f(\boldsymbol{y}|\boldsymbol{x}) - f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right] \right|$$

$$\leq \sum_{k=1}^{n^d} \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} \left| f(\boldsymbol{y}|\boldsymbol{x}) - f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right|, \tag{9}$$

by the triangle inequality.

Since $\boldsymbol{x}_k^n \in \mathbb{X}_k^n$, for each k and n, we have the fact that $\left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}_k^n, \boldsymbol{y}) \right\| < \delta_n$ for $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{X}_k^n \times \mathbb{Y}$. By uniform continuity, for each ε, we can find a sufficiently small $\delta(\epsilon)$, such that $\left| f(\boldsymbol{y}|\boldsymbol{x}) - f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right| < \varepsilon$, if $\left\| (\boldsymbol{x}, \boldsymbol{y}) - (\boldsymbol{x}_k^n, \boldsymbol{y}) \right\| < \delta(\epsilon)$, for all k. The desired result (8) can be obtained by noting that the right hand side of (9) consists of only one non-zero summand for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{Z}$, and by choosing $n \in \mathbb{N}$ sufficiently large, so that $\delta_n < \delta(\epsilon)$.

By (8), we have the fact that $\upsilon_n \to f$, point-wise. We can bound υ_n as follows:

$$\upsilon_n(\boldsymbol{y}|\boldsymbol{x}) \le \sum_{i=1}^{n^p} \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} \sup_{\boldsymbol{\zeta} \in \mathbb{Y}, \boldsymbol{\xi} \in \mathbb{X}} f(\boldsymbol{\zeta}|\boldsymbol{\xi}) = \sup_{\boldsymbol{\zeta} \in \mathbb{Y}, \boldsymbol{\xi} \in \mathbb{X}} f(\boldsymbol{\zeta}|\boldsymbol{\xi}), \tag{10}$$

where the right-hand side is a constant and is therefore in $\mathcal{L}_p$, since $\mathbb{Z}$ is compact. An application of the Lebesgue dominated convergence theorem in $\mathcal{L}_p$ then yields (2).

Next we write

$$\begin{aligned} \|\upsilon_n - \eta_n\|_p &= \left\| \sum_{k=1}^{n^d} \mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} f(\boldsymbol{y}|\boldsymbol{x}_k^n) - \sum_{k=1}^{n^d} \mathrm{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right\|_p \\ &\le \sum_{k=1}^{n^d} \left\| \left[\mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) \right] f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right\|_p. \end{aligned}$$

Since the norm arguments are separable in $\boldsymbol{x}$ and $\boldsymbol{y}$, we apply Fubini's theorem to get

$$\begin{aligned} \|\upsilon_n - \eta_n\|_p &= \sum_{k=1}^{n^d} \left\| \left[\mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) \right] \right\|_{p,\mathbb{X}} \left\| f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right\|_{p,\mathbb{Y}} \\ &\le \sup_{k \in [n^d]} \left\| \left[\mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) \right] \right\|_{p,\mathbb{X}} \sum_{k=1}^{n^d} \left\| f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right\|_{p,\mathbb{Y}} \end{aligned}$$

Because $f \in \mathcal{B}$ and n^d is finite, for any fixed $n \in \mathbb{N}$, we have $C_1(n) = \sum_{k=1}^{n^d} \left\| f(\boldsymbol{y}|\boldsymbol{x}_k^n) \right\|_{p,\mathbb{Y}} < \infty$. For each $\epsilon > 0$, we need to choose a $\boldsymbol{\gamma}_n \in \mathbb{G}_S^{n^d}$, such that

$$\sup_{k \in [n^d]} \left\| \left[\mathbf{1}_{\{\boldsymbol{x} \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k(\boldsymbol{x}; \boldsymbol{\gamma}_n) \right] \right\|_{p,\mathbb{X}} < \frac{\epsilon}{3 C_1(n)},$$

which can be achieved via a direct application of Lemma 2. We have thus shown (4).

Lastly, we are required to approximate $f(\boldsymbol{y}|\boldsymbol{x}_k^n)$ for each $k \in [n^d]$, by a function of form (6). Since $\mathbb{Y}$ is compact and f and ψ are continuous, we can apply of Lemma 4, directly. Note that over a set of finite measure, convergence in $\|\cdot\|_{\mathcal{B}}$ implies convergence in $\mathcal{L}_p$ norm, for all $p \in [1, \infty]$ (cf. Oden and Demkowicz 2010, Prop. 3.9.3).

We can then write (5) as

$$\begin{aligned} m_{K_n}^{\psi}(\boldsymbol{y}|\boldsymbol{x}) &= \sum_{k=1}^{n^d} \frac{\exp\left(a_{n,k} + \boldsymbol{b}_{n,k}^{\top} \boldsymbol{x}\right)}{\sum_{l=1}^{n^d} \exp\left(a_{n,l} + \boldsymbol{b}_{n,l}^{\top} \boldsymbol{x}\right)} h_{n_k}^k(\boldsymbol{y}|\boldsymbol{x}_k^n) \\ &= \sum_{k=1}^{n^d} \sum_{i=1}^{n_k} \frac{\exp\left(a_{n,k} + \boldsymbol{b}_{n,k}^{\top} \boldsymbol{x}\right)}{\sum_{l=1}^{n^d} \exp\left(a_{n,l} + \boldsymbol{b}_{n,l}^{\top} \boldsymbol{x}\right)} \frac{c_i^k}{\sum_{l=1}^{n_k} c_l^k} g_{\psi}\left(\boldsymbol{y}; \boldsymbol{\mu}_i^k, \sigma_i^k\right) \\ &= \sum_{k=1}^{n^d} \sum_{i=1}^{n_k} \frac{\exp\left(\log c_i^k + a_{n,k} + \boldsymbol{b}_{n,k}^{\top} \boldsymbol{x}\right)}{\sum_{l=1}^{n^d} \sum_{j=1}^{n_k} \exp\left(\log c_j^k + a_{n,l} + \boldsymbol{b}_{n,l}^{\top} \boldsymbol{x}\right)} g_{\psi}\left(\boldsymbol{y}; \boldsymbol{\mu}_i^k, \sigma_i^k\right), \tag{11} \end{aligned}$$

where $\boldsymbol{\gamma}_n = \left(a_{n,1}, \ldots, a_{n,n^d}, \boldsymbol{b}_{n,1}, \ldots, \boldsymbol{b}_{n,n^d}\right)$. From (11), we observe that $m_{K_n}^{\psi} \in \mathcal{M}_S^{\psi}$, with $K_n = \sum_{k=1}^{n^d} n_k$.

To obtain (7), we write

$$
\left\| \eta_n - m_{K_n}^{\psi} \right\|_p = \left\| \sum_{k=1}^{n^d} \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_n\right) f\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) - \sum_{k=1}^{n^d} \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}\right) h_{n_k}^k\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) \right\|_p
$$

$$
\leq \sum_{k=1}^{n^d} \left\| \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_n\right) \left[f\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) - h_{n_k}^k\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) \right] \right\|_p.
$$

By separability and Fubini's theorem, we then have

$$
\left\| \eta_n - m_{K_n}^{\psi} \right\| \leq \sum_{k=1}^{n^d} \left\| \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_n\right) \right\|_{p,\mathbb{X}} \left\| f\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) - h_{n_k}^k\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) \right\|_{p,\mathbb{Y}}
$$

$$
\leq \sup_{k \in \left[n^d\right]} \left\| \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_n\right) \right\|_{p,\mathbb{X}} \sum_{k=1}^{n^d} \left\| f\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) - h_{n_k}^k\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) \right\|_{p,\mathbb{Y}}.
$$

Let $C_2\left(n, \boldsymbol{\gamma}_n\right) = \sup_{k \in \left[n^d\right]} \left\| \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_n\right) \right\|_{p,\mathbb{X}}$. Then, we apply Lemma 4 n^d times to establish the existence of a constant $N_2\left(\epsilon, n, \boldsymbol{\gamma}_n\right) \in \mathbb{N}$, such that for all $k \in \left[n^d\right]$ and $n_k \geq N_2\left(\epsilon, n, \boldsymbol{\gamma}_n\right)$,

$$
\left\| f\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) - h_{n_k}^k\left(\boldsymbol{y}|\boldsymbol{x}_k^n\right) \right\|_{p,\mathbb{Y}} \leq \frac{\epsilon}{3C_2\left(n, \boldsymbol{\gamma}_n\right) n^d}.
$$

Thus, we have

$$
\left\| \eta_n - m_{K_n}^{\psi} \right\| \leq C_2\left(n, \boldsymbol{\gamma}_n\right) \times n^d \times \frac{\epsilon}{3C_2\left(n, \boldsymbol{\gamma}_n\right) n^d} = \frac{\epsilon}{3},
$$

which completes our proof.

4.2 Proof of Theorem 2

The proof is procedurally similar to that of Theorem 1 and thus we only seek to highlight the important differences. Firstly, for any $\epsilon > 0$, we approximate $f\left(\boldsymbol{y}|\boldsymbol{x}\right)$ by $\upsilon_n\left(\boldsymbol{x}|\boldsymbol{y}\right)$ of form (1), with $d = 1$. Result (2) implies uniform convergence, in the sense that there exists an $N_1(\epsilon) \in \mathbb{N}$, such that for all $n \geq N_1(\epsilon)$,

$$
\left\| f - \upsilon_n \right\|_{\mathcal{B}} < \frac{\epsilon}{3}. \tag{12}
$$

We now seek to approximate υ_n by η_n of form (3), with $\boldsymbol{\gamma}_n = \boldsymbol{\gamma}_l$ for some $l \in \mathbb{N}$. Upon application of Lemma 3, it follows that for each $k \in \left[n^d\right]$ and $\varepsilon > 0$, there exists a measurable set $\mathbb{B}_k(\varepsilon) \subseteq \mathbb{X}$, such that

$$
\lambda\left(\mathbb{B}_k(\varepsilon)\right) < \frac{\varepsilon}{n^d \lambda\left(\mathbb{Y}\right)}
$$

and

$$
\left\| \mathrm{Gate}_k\left(\cdot; \boldsymbol{\gamma}_l\right) - \mathbf{1}_{\left\{\boldsymbol{x} \in \mathbb{X}_k^n\right\}} \right\|_{\mathcal{B}\left(\mathbb{B}_k^{\complement}(\varepsilon)\right)} < \frac{\epsilon}{3}
$$

for all $l \geq M_k\left(\epsilon, n\right)$, for some $M_k\left(\epsilon, n\right) \in \mathbb{N}$. Here, $(\cdot)^{\complement}$ is the set complement operator.

Since $f \in \mathcal{B}$, we have the bound $C(n) = \sum_{k=1}^{n^d} \left\| f\left(y | x_k^n\right) \right\|_{\mathcal{B}(\mathbb{Y})} < \infty$. Write $\mathbb{B}(\varepsilon) = \bigcup_{k=1}^{n^d} \mathbb{B}_k(\varepsilon)$. Then, $\mathbb{B}^{\complement}(\varepsilon) = \bigcap_{k=1}^{n^d} \mathbb{B}_k^{\complement}(\varepsilon)$,

$$\lambda\left(\mathbb{B}(\varepsilon)\right) \leq \sum_{k=1}^{n^d} \lambda\left(\mathbb{B}_k^{\complement}(\varepsilon)\right) < \frac{\varepsilon}{\lambda\left(\mathbb{Y}\right)},$$

and

$$\left\| \mathrm{Gate}_k\left(\cdot; \boldsymbol{\gamma}_l\right) - \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} \leq \min_{k \in [n^d]} \left\| \mathrm{Gate}_k\left(\cdot; \boldsymbol{\gamma}_l\right) - \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} \right\|_{\mathcal{B}\left(\mathbb{B}_k^{\complement}(\varepsilon)\right)} < \frac{\epsilon}{3C(n)},$$

for all $l \geq M(\epsilon, n) = \max_{k \in [n^d]} M_k(\epsilon, n)$. Here we use the fact that the supremum over some intersect of sets is less than or equal to the minimum of the supremum over each individual set.

Upon defining $\mathbb{C}(\varepsilon) = \mathbb{B}(\varepsilon) \times \mathbb{Y} \subset \mathbb{Z}$, we observe that

$$\lambda\left(\mathbb{C}(\varepsilon)\right) = \lambda\left(\mathbb{B}(\varepsilon)\right) \lambda\left(\mathbb{Y}\right) \leq \frac{\varepsilon}{\lambda\left(\mathbb{Y}\right)} \times \lambda\left(\mathbb{Y}\right) = \varepsilon,$$

and $\mathbb{C}(\varepsilon) \subset \mathbb{B}(\varepsilon) \times \mathbb{Y}$. Note also that

$$\left(\mathbb{B}(\varepsilon) \times \mathbb{Y}\right)^{\complement} = \mathbb{Z} \setminus \left(\mathbb{B}(\varepsilon) \times \mathbb{Y}\right) = \mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}$$

and

$$\mathbb{C}^{\complement}(\varepsilon) = \left(\mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}\right) \cup \left(\mathbb{B}(\varepsilon) \times \mathbb{Y}^{\complement}\right) \cup \left(\mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}^{\complement}\right).$$

It follows that

$$\left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} \leq \max\left\{ \left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}\right)}, \left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{B}(\varepsilon) \times \mathbb{Y}^{\complement}\right)}, \left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}^{\complement}\right)} \right\}.$$

Since $\mathbb{B}(\varepsilon) \times \mathbb{Y}^{\complement}$ and $\mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}^{\complement}$ are empty, via separability, we have

$$\left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} = \left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}\right)}$$

$$= \sup_{z \in \mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}} \left| \sum_{k=1}^{n^d} \left[\mathbf{1}_{\{x \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k\left(x; \boldsymbol{\gamma}_l\right) \right] f\left(y | x_k^n\right) \right|$$

$$\leq \sup_{z \in \mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}} \sum_{k=1}^{n^d} \left| \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k\left(x; \boldsymbol{\gamma}_l\right) \right| f\left(y | x_k^n\right)$$

$$\leq \sum_{k=1}^{n^d} \sup_{z \in \mathbb{B}^{\complement}(\varepsilon) \times \mathbb{Y}} \left| \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k\left(x; \boldsymbol{\gamma}_l\right) \right| f\left(y | x_k^n\right)$$

$$= \sum_{k=1}^{n^d} \left\| \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k\left(x; \boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} \left\| f\left(y | x_k^n\right) \right\|_{\mathcal{B}(\mathbb{Y})}$$

$$\leq \sup_{k \in [n]} \left\| \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k\left(x; \boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} \sum_{k=1}^{n^d} \left\| f\left(y | x_k^n\right) \right\|_{\mathcal{B}(\mathbb{Y})}.$$

Recall that the $\sum_{k=1}^{n^d} \left\| f\left(y | x_k^n\right) \right\|_{\mathcal{B}(\mathbb{Y})} = C(n) < \infty$ and that we can choose $l \geq M(\epsilon, n)$ so that

$$\sup_{k \in [n]} \left\| \mathbf{1}_{\{x \in \mathbb{X}_k^n\}} - \mathrm{Gate}_k\left(x; \boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}\left(\mathbb{B}^{\complement}(\varepsilon)\right)} < \frac{\epsilon}{3C(n)},$$

and thus

$$\|\upsilon_n - \eta_n\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} < \frac{\epsilon}{3C(n)} \times C(n) = \frac{\epsilon}{3},\tag{13}$$

as required.

Finally, by noting that for each $k \in [n^d]$, both (6) and $f\left(\cdot|\boldsymbol{x}_k^n\right)$ are continuous over $\mathbb{Y}$, we apply Lemma 4 to obtain an $N_2\left(\epsilon, n, l\right) \in \mathbb{N}$, such that for any $\epsilon > 0$ and $n_k \geq N_2\left(\epsilon, n, l\right)$, we have

$$\left\| f\left(\cdot|\boldsymbol{x}_k^n\right) - h_{n_k}^k\left(\cdot|\boldsymbol{x}_k^n\right) \right\|_{\mathcal{B}(\mathbb{Y})} < \frac{\epsilon}{3M_1 n}.$$

Here $M_1 = \sup_{k \in [n^d]} \left\| \mathrm{Gate}_k\left(\cdot; \boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}(\mathbb{X})} < \infty$, since $\mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_l\right)$ is continuous in $\boldsymbol{x}$, and $\mathbb{X}$ is compact. Therefore, for all $K_n = \sum_{k=1}^{n^d} n_k$, $n_k \geq N_2\left(\epsilon, n, l\right)$,

$$\left\| \eta_n - m_{K_n}^{\psi} \right\|_{\mathcal{B}} \leq \sup_{k \in [n^d]} \left\| \mathrm{Gate}_k\left(\boldsymbol{x}; \boldsymbol{\gamma}_l\right) \right\|_{\mathcal{B}(\mathbb{X})} \sum_{k=1}^{n^d} \left\| f\left(\cdot|\boldsymbol{x}_k^n\right) - h_{n_k}^k\left(\cdot|\boldsymbol{x}_k^n\right) \right\|_{\mathcal{B}(\mathbb{Y})}$$

$$= M_1 \times n^d \times \frac{\epsilon}{3M_1 n^d} = \frac{\epsilon}{3}.\tag{14}$$

In summary, via (12), (13), and (14), for each $\epsilon > 0$, for any $\varepsilon > 0$, there exists a $\mathbb{C}(\varepsilon) \subset \mathbb{Z}$ and constants $N_1(\epsilon), M\left(\epsilon, n\right), N_2\left(\epsilon, n, l\right) \in \mathbb{N}$, such that for all $K_n = \sum_{k=1}^{n^d} n_k$, with $n_k \geq N_2\left(\epsilon, n, l\right)$, $l \geq M\left(\epsilon, n\right)$, and $n \geq N_1(\epsilon)$, it follows that $\lambda\left(\mathbb{C}(\varepsilon)\right) < \varepsilon$, and

$$\left\| f - m_{K_n}^{\psi} \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} \leq \left\| f - \upsilon_n \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} + \left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} + \left\| \eta_n - m_{K_n}^{\psi} \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)}$$

$$\leq \left\| f - \upsilon_n \right\|_{\mathcal{B}} + \left\| \upsilon_n - \eta_n \right\|_{\mathcal{B}\left(\mathbb{C}^{\complement}(\varepsilon)\right)} + \left\| \eta_n - m_{K_n}^{\psi} \right\|_{\mathcal{B}}$$

$$< 3 \times \frac{\epsilon}{3} = \epsilon.$$

This completes the proof.

5   Proofs of lemmas

5.1   Proof of Lemma 1

We firstly prove that any gating vector from $\mathcal{G}_S^K$ can be equivalently represented as an element of $\mathcal{G}_G^K$. For any $\boldsymbol{x} \in \mathbb{R}^d$, $d \in \mathbb{N}$, $k \in [K]$, $a_k \in \mathbb{R}$, $\boldsymbol{b}_k \in \mathbb{R}^d$, and $K \in \mathbb{N}$, choose $\boldsymbol{v}_k = \boldsymbol{b}_k$, $\tau_k = a_k + \boldsymbol{b}_k^{\top} \boldsymbol{b}_k / 2$ and

$$\pi_k = \exp\left(\tau_k\right) / \sum_{l=1}^{K} \exp\left(\tau_l\right).$$

This implies that $\sum_{l=1}^{K} \pi_l = 1$, $\pi_l > 0$, for all $l \in [K]$, and

$$\frac{\exp\left(a_k + \boldsymbol{b}_k^{\top} \boldsymbol{x}\right)}{\sum_{l=1}^{K} \exp\left(a_k + \boldsymbol{b}_k^{\top} \boldsymbol{x}\right)} = \frac{\exp\left(\tau_k - \boldsymbol{v}_k^{\top} \boldsymbol{v}_k / 2 + \boldsymbol{v}_k^{\top} \boldsymbol{x}\right)}{\sum_{l=1}^{K} \exp\left(\tau_l - \boldsymbol{v}_l^{\top} \boldsymbol{v}_l / 2 + \boldsymbol{v}_l^{\top} \boldsymbol{x}\right)}$$

$$= \frac{\exp\left(\tau_k\right) \exp\left(-\left(\boldsymbol{x} - \boldsymbol{v}_k\right)^{\top} \left(\boldsymbol{x} - \boldsymbol{v}_k\right) / 2\right)}{\sum_{l=1}^{K} \exp\left(\tau_l\right) \exp\left(-\left(\boldsymbol{x} - \boldsymbol{v}_l\right)^{\top} \left(\boldsymbol{x} - \boldsymbol{v}_l\right) / 2\right)}$$

$$= \frac{\pi_k \left(2\pi\right)^{-d/2} \exp\left(-\left(\boldsymbol{x} - \boldsymbol{v}_k\right)^{\top} \left(\boldsymbol{x} - \boldsymbol{v}_k\right) / 2\right)}{\sum_{l=1}^{K} \pi_l \left(2\pi\right)^{-d/2} \exp\left(-\left(\boldsymbol{x} - \boldsymbol{v}_l\right)^{\top} \left(\boldsymbol{x} - \boldsymbol{v}_l\right) / 2\right)}$$

$$= \frac{\pi_k \phi\left(\boldsymbol{x}; \boldsymbol{v}_k, \mathbf{I}\right)}{\sum_{l=1}^{K} \pi_l \phi\left(\boldsymbol{x}; \boldsymbol{v}_l, \mathbf{I}\right)},$$

where $\mathbf{I}$ is the identity matrix of appropriate size. This proves that $\mathcal{G}_S^K \subset \mathcal{G}_G^K$.

Next, to show that $\mathcal{G}_E^K \subset \mathcal{G}_S^K$, we write

$$\frac{\pi_k \phi\left(\mathbf{x}; \mathbf{v}_k, \mathbf{\Sigma}\right)}{\sum_{l=1}^K \pi_l \phi\left(\mathbf{x}; \mathbf{v}_l, \mathbf{\Sigma}\right)}$$

$$= \frac{\pi_k \left|2\pi\mathbf{\Sigma}\right|^{-1/2} \exp\left(-\left(\mathbf{x}-\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{v}_k\right)/2\right)}{\sum_{l=1}^K \pi_l \left|2\pi\mathbf{\Sigma}\right|^{-1/2} \exp\left(-\left(\mathbf{x}-\mathbf{v}_l\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{v}_l\right)/2\right)}$$

$$= \frac{1}{\sum_{l=1}^K \exp\left(-\log\left(\pi_l^{-2}/\pi_k^{-2}\right)/2 - \left(\mathbf{x}-\mathbf{v}_l\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{v}_l\right)/2 - \left(\mathbf{x}-\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{v}_k\right)/2\right)},$$

and note that

$$\left(\mathbf{x}-\mathbf{v}_l\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{v}_l\right) - \left(\mathbf{x}-\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mathbf{v}_k\right)$$

$$= -2\left(\mathbf{v}_l-\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\mathbf{x} + \left(\mathbf{v}_l+\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{v}_l-\mathbf{v}_k\right).$$

Thus, we have

$$\frac{\pi_k \phi\left(\mathbf{x}; \mathbf{v}_k, \mathbf{\Sigma}\right)}{\sum_{l=1}^K \pi_l \phi\left(\mathbf{x}; \mathbf{v}_l, \mathbf{\Sigma}\right)}$$

$$= \frac{1}{\sum_{l=1}^K \exp\left(-\log\left(\pi_l^{-2}/\pi_k^{-2}\right)/2 - \left(\mathbf{v}_l+\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\mathbf{v}_l-\mathbf{v}_k\right)/2 - \left(\mathbf{v}_l-\mathbf{v}_k\right)^\top \mathbf{\Sigma}^{-1}\mathbf{x}\right)}.$$

Next, notice that we can write

$$\frac{\exp\left(a_k + \mathbf{b}_k^\top \mathbf{x}\right)}{\sum_{l=1}^K \exp\left(a_l + \mathbf{b}_l^\top \mathbf{x}\right)} = \frac{1}{\sum_{l=1}^K \exp\left(\alpha_l + \boldsymbol{\beta}_l^\top \mathbf{x}\right)},$$

where $\alpha_l = a_l - a_k$ and $\boldsymbol{\beta}_l = \boldsymbol{\beta}_l - \boldsymbol{\beta}_k$. We now choose a_k and $\mathbf{b}_k$, such that for every $l \in [K]$,

$$\alpha_l = a_l - a_k = -\frac{1}{2}\log\left(\frac{\pi_l^{-2}}{\pi_k^{-2}}\right) - \frac{1}{2}\left(\mathbf{v}_l^\top \mathbf{\Sigma}^{-1}\mathbf{v}_l - \mathbf{v}_k^\top \mathbf{\Sigma}^{-1}\mathbf{v}_k\right),$$

and

$$\boldsymbol{\beta}_l = \boldsymbol{\beta}_l - \boldsymbol{\beta}_k = \mathbf{v}_l^\top \mathbf{\Sigma}^{-1} - \mathbf{v}_k^\top \mathbf{\Sigma}^{-1}.$$

To complete the proof, we choose

$$a_k = \log\left(\pi_k\right) - \frac{1}{2}\mathbf{v}_k^\top \mathbf{\Sigma}^{-1}\mathbf{v}_k$$

and $\mathbf{b}_k = \mathbf{v}_k^\top \mathbf{\Sigma}^{-1}$, for each $k \in [K]$.

5.2 Proof of Lemma 3

For $l \in [0, \infty)$, write

$$\text{Gate}_k\left(x, l\right) = \frac{\exp\left(\left[x-c_k\right]lk\right)}{\sum_{i=1}^n \exp\left(\left[x-c_i\right]li\right)},$$

where $x \in \mathbb{X} = [0, 1]$, and $c_k = (k-1)/(2k)$. We identify that $\mathbf{Gate} = \left(\text{Gate}_k\left(x, l\right)\right)_{k \in [n]}$ belongs to the class $\mathcal{G}_S^n$. The proof of the Section 4 Proposition from Jiang and Tanner (1999a) reveals that for all $k \in [n]$,

$$\text{Gate}_k\left(x, l\right) \to \mathbf{1}_{\left\{x \in \mathbb{I}_k^n\right\}}$$

almost everywhere in λ, as $l \to \infty$. The result then follows via an application of Egorov's theorem (cf. Folland 1999, Thm. 2.33).

6 Summary and conclusions

Using recent results mixture model approximation results Nguyen et al. (2020) and Nguyen et al. (2020a), and the indicator approximation theorem of Jiang and Tanner (1999a) (cf. Section 3), we have proved two approximation theorems (Theorems 1 and 2) regarding the class of soft-max gated MoE models with experts arising from arbitrary location-scale families of conditional density functions. Via an equivalence result (Lemma 1), the results of Theorems 1 and 2 also extend to the setting of Gaussian gated MoE models (Corollary 1), which can be seen as a generalization of the soft-max gated MoE models.

Although we explicitly make the assumption that $\mathbb{X} = [0, 1]^d$, for the sake of mathematical argument (so that we can make direct use of Lemma 2), a simple shift-and-scale argument can be used to generalize our result to cases where $\mathbb{X}$ is any generic compact domain. The compactness assumption regarding the input domain is common in the MoE and mixture of regression models literature, as per the works of Jiang and Tanner (1999b); Jiang and Tanner (1999a); Norets (2010); Montuelle and Le Pennec (2014); Pelenis (2014); Devijver (2015a); Devijver (2015b).

The assumption permits the application of the result to the settings where the inputs X is assumed to be non-random design vectors that take value on some compact set $\mathbb{X}$. This is often the case when there is only a finite number of possible design vector elements for which X can take. Otherwise, the assumption also permits the scenario where X is some random element with compactly supported distribution, such as uniformly distributed, or beta distributed inputs. Unfortunately, the case of random X over an unbounded domain (e.g., if X has multivariate Gaussian distribution) is not covered under our framework. An extension to such cases would require a more general version of Lemma 2, which we believe is a nontrivial direction for future work.

Like the input, we also assume that the output domain is restricted to a compact set $\mathbb{Y}$. However, the output domain of the approximating class of MoE models is unrestricted to $\mathbb{Y}$ and thus the functions (i.e., we allow ψ to be a PDF over $\mathbb{R}^q$). The restrictions placed on $\mathbb{Y}$ is also common in the mixture approximation literature, as per the works of Zeevi and Meir (1997); Li and Barron (1999); Rakhlin et al. (2005), and is also often made in the context of nonparametric regression (see, e.g., Gyorfi et al. 2002; Cucker and Zhou 2007). Here, our use of the compactness of $\mathbb{Y}$ is to bound the integral of ν_n, in (10). A more nuanced approach, such as via the use of a generalized Lebesgue spaces (see e.g., Castillo and Rafeiro 2010; Cruze-Uribe and Fiorenza 2013), may lead to result for unbounded $\mathbb{Y}$. This is another exciting future direction of our research program.

A trivial modification to the proof of Lemma 4 allows us to replace the assumption that f is a PDF with a sub-PDF assumption (i.e., $\int_{\mathbb{Y}} f \, \mathrm{d}\lambda \leq 1$), instead. This in turn permits us to replace the assumption that $f(\cdot|\boldsymbol{x})$ is a conditional PDF in Theorems 1 and 2 with sub-PDF assumptions as well (i.e., for each $\boldsymbol{x} \in \mathbb{X}$, $\int_{\mathbb{Y}} f(\boldsymbol{y}|\boldsymbol{x}) \, \mathrm{d}\lambda(\boldsymbol{y}) \leq 1$). Thus, in this modified form, we have a useful interpretation for situations when the input Y is unbounded. That is, when Y is unbounded, we can say that the conditional PDF f can be arbitrarily well approximated in $\mathcal{L}_p$ norm by a sequence $\left\{ m_K^\psi \right\}_{K \in \mathbb{N}}$ of either soft-max or Gaussian gated MoEs over any compact subdomain $\mathbb{Y}$ of the unbounded domain of Y. Thus, although we cannot provide guarantees of the entire domain of Y, we are able to guarantee arbitrary approximate fidelity over any arbitrarily large compact subdomain. This is a useful result in practice since one is often not interested in the entire domain of Y, but only on some

subdomain where the probability of Y is concentrated. This version of the result resembles traditional denseness results in approximation theory, such as those of Cheney and Light (2000), Ch. 20.

Finally, our results can be directly applied to provide approximation guarantees for a large number of currently used models in applied statistics and machine learning research. Particularly, our approximation guarantees are applicable to the recent MoE models of Ingrassia et al. (2012); Chamroukhi et al. (2013); Ingrassia et al. (2014); Chamroukhi (2016); Nguyen and McLachlan (2016); Deleforge et al. (2015a); Deleforge et al. (2015b); Kalliovirta et al. (2016); Perthame et al. (2018), among many others. Here, we may guarantee that the underlying data generating processes, if satisfying our assumptions, can be adequately well approximated by sufficiently complex forms of the models considered in each of the aforementioned work.

The rate and manner of which good approximation can be achieved as a function of the number of experts K and the sample size is a currently active research area, with pioneering work conducted in Cohen and Le Pennec (2012); Montuelle and Le Pennec (2014). More recent results in this direction appear in Nguyen et al. (2020b); Nguyen et al. (2021); Nguyen et al. (2021).

Availability of data and materials
Not applicable.

Code availability
Not applicable.

Declarations

Competing interests
None.

Author details
[1]Department of Mathematics and Statistics, La Trobe University, Bundoora Victoria, Australia. [2]Normandie Univ, UNICAEN, CNRS, LMNO, 14000 Caen, France. [3]School of Mathematics and Physics, University of Queensland, St. Lucia Brisbane, Australia.

References
Bartle, R.: The Elements of Integration and Lebesgue Measure. Wiley, New York (1995)
Castillo, R. E., Rafeiro, H.: An Introductory Course in Lebesgue Spaces. Springer, Switzerland (2010)
Chamroukhi, F.: Robust mixture of experts modeling using the t distribution. Neural Netw. **79**, 20–36 (2016)
Chamroukhi, F., Mohammed, S., Trabelsi, D., Oukhellou, L., Amirat, Y.: Joint segmentation of multivariate time series with hidden process regression for human activity recognition. Neurocomputing. **120**, 633–644 (2013)
Cheney, W., Light, W.: A Course in Approximation Theory. Brooks/Cole, Pacific Grove (2000)

Cohen, S., Le Pennec, E.: Conditional density estimation by penalized likelihood model selection and application. ArXiv (arXiv:1103.2021) (2012)

Cruze-Uribe, D. V., Fiorenza, A.: Variable Lebesgue Spaces: Foundations and Harmonic Analysis. Birkhauser, Basel (2013)

Cucker, F., Zhou, D.-X.: Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge (2007)

Deleforge, A., Forbes, F., Horaud, R.: High-dimensional regression with Gaussian mixtures and partially-latent response variables. Stat. Comput. **25**, 893–911 (2015)

Deleforge, A., Forbes, F., Horaud, R.: Acoustic space learning for sound-source separation and localization on binaural manifolds. Int. J. Neural Syst. **25**, 1440003 (2015)

Devijver, E.: An ℓ_1-oracle inequality for the Lasso in multivariate finite mixture of multivariate Gaussian regression models. ESAIM: Probab. Stat. **19**, 649–670 (2015)

Devijver, E.: Finite mixture regression: a sparse variable selection by model selection for clustering. Electron. J. Stat. **9**, 2642–2674 (2015)

Folland, G. B.: Real Analysis: Modern Techniques and Their Applications. Wiley, New York (1999)

Gyorfi, L., Kohler, M., Krzyzak, A., Walk, H.: A Distribution-free Theory Of Nonparametric Regression. Springer, New York (2002)

Ingrassia, S., Minotti, S. C., Punzo, A.: Model-based clustering via linear cluster-weighted models. Comput. Stat. Data Anal. **71**, 159–182 (2014)

Ingrassia, S., Minotti, S. C., Vittadini, G.: Local statistical modeling via a cluster-weighted approach with elliptical distributions. J. Classif. **29**, 363–401 (2012)

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

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

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

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

Kalliovirta, L., Meitz, M., Saikkonen, P.: Gaussian mixture vector autoregression. J. Econ. **192**, 485–498 (2016)

Krzyzak, A., Schafer, D.: Nonparametric regression estimation by normalized radial basis function networks. IEEE Trans. Inf. Theory. **51**, 1003–1010 (2005)

Li, J. Q., Barron, A. R.: Mixture density estimation. In: Solla, S. A., Leen, T. K., Mueller, K. R. (eds.) Advances in Neural Information Processing Systems. MIT Press, Cambridge, (1999)

Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artif. Intell. Rev. **42**, 275–293 (2014)

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

Montuelle, L., Le Pennec, E.: Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach. Electron. J. Stat. **8**, 1661–1695 (2014)

Nguyen, H. D., Chamroukhi, F.: Practical and theoretical aspects of mixture-of-experts modeling: an overview. WIREs Data Min. Knowl. Disc. **8**(4), 1246 (2018)

Nguyen, H. D., Chamroukhi, F., Forbes, F.: Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. Neurocomputing. **366**, 208–214 (2019)

Nguyen, T. T., Chamroukhi, F., Nguyen, H. D., Forbes, F.: Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. arXiv preprint arXiv:2104.08959 (2021). http://arxiv.org/abs/2104.08959

Nguyen, T. T., Chamroukhi, F., Nguyen, H. D., McLachlan, G. J.: Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. arXiv:2008.09787 (2020)

Nguyen, H. D., Lloyd-Jones, L. R., McLachlan, G. J.: A universal approximation theorem for mixture-of-experts models. Neural Comput. **28**, 2585–2593 (2016)

Nguyen, H. D., McLachlan, G. J.: Laplace mixture of linear experts. Comput. Stat. Data Anal. **93**, 177–191 (2016)

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., Forbes, F.: A non-asymptotic penalization criterion for model selection in mixture of experts models. ArXiv (arXiv:2104.02640) (2021)

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., McLachlan, G. J.: Approximation by finite mixtures of continuous density functions that vanish at infinity. Cogent Math. Stat. **7**, 1750861 (2020)

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., McLachlan, G. J.: An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. ArXiv (arXiv:2009.10622) (2020)

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

Norets, A., Pelenis, J.: Posterior consistency in conditional density estimation by covariate dependent mixtures. Econ. Theory. **30**, 606–646 (2014)

Oden, J. T., Demkowicz, L. F.: Applied Functional Analysis. CRC Press, Boca Raton (2010)

Pelenis, J.: Bayesian regression with heteroscedastic error density and parametric mean function. J. Econ. **178**, 624–638 (2014)

Perthame, E., Forbes, F., Deleforge, A.: Inverse regression approach to robust nonlinear high-to-low dimensional mapping. J. Multivar. Anal. **163**, 1–14 (2018)

Rakhlin, A., Panchenko, D., Mukherjee, S.: Risk bounds for mixture density estimation. ESAIM: Probab. Stat. **9**, 220–229 (2005)

Wang, L.-X., Mendel, J. M.: Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. IEEE Trans. Neural Netw. **3**, 807–814 (1992)

Yuksel, S. E., Wilson, J. N., Gader, P. D.: Twenty years of mixture of experts. IEEE Trans. Neural Netw. Learn. Syst. **23**, 1177–1193 (2012)

Zeevi, A. J., Meir, R.: Density estimation through convex combinations of densities: approximation and estimation bounds. Neural Comput. **10**, 99–109 (1997)

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.