Demystifying Softmax Gating Function in Gaussian Mixture of Experts

Huy Nguyen[◇], TrungTin Nguyen[†], Nhat Ho[◇]

The University of Texas at Austin[◇]
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK[†]

November 5, 2023

Introduction

- **Problem.** Establish the convergence rates of maximum likelihood estimation under the softmax gating Gaussian mixture of experts.
- **Goals.** Understand the effects of softmax gating function on Gaussian mixture of experts via the parameter estimation problem.

Preliminaries

Setup. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R} \overset{\text{i.i.d}}{\sim} g_{G_*}(Y|X)$:

$$g_{G_*}(Y|X) := \sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*), \tag{1}$$

where

- k_* is the true number of experts of the form $(a_i^*)^\top X + b_i^*$;
- $f(\cdot|\mu, \sigma)$ is a Gaussian density function with mean μ and variance σ;
- $G_* := \sum_{i=1}^{k_*} \exp(\beta_{0i}^*)\delta_{(\beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*)}$ is a true but unknown mixing measure;
- True parameters $(\beta_{0i}^*, \beta_{1i}^*, a_i^*, b_i^*, \sigma_i^*) \in \Theta \subset \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}_+$.

Preliminaries

Setup. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R} \overset{\text{i.i.d}}{\sim} g_{G_*}(Y|X)$:

$$g_{G_*}(Y|X) := \sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top X + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top X + \beta_{0j}^*)} \cdot f(Y|(a_i^*)^\top X + b_i^*, \sigma_i^*).$$

Assumptions:

- The covariate $X \in \mathcal{X}$ follows a continuous distribution, where $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$, while the parameter space Θ is compact;
- Expert parameters $(a_1^*, b_1^*, \sigma_1^*), \ldots, (a_{k_*}^*, b_{k_*}^*, \sigma_{k_*}^*)$ are pairwise distinct;
- At least one among parameters $\beta_{11}^*, \ldots, \beta_{1k_*}^*$ is different from zero.

Preliminaries

Maximum likelihood estimation (MLE). Since the true number of experts k_* is unknown in practice, we use MLE within a class of all mixing measures with at most k atoms, where $k \geq k_*$:

$$\widehat{G}_n \in \arg\max_{G \in \mathcal{O}_k(\Theta)} \frac{1}{n} \sum_{i=1}^{n} \log(g_G(Y_i|X_i)), \quad (2)$$

where we define $\mathcal{O}_k(\Theta) := \{G = \sum_{i=1}^{k'} \exp(\beta_{0i}) \delta_{(\beta_{1i}, a_i, b_i, \sigma_i)} : 1 \leq k' \leq k$ and $(\beta_{0i}, \beta_{1i}, a_i, b_i, \sigma_i) \in \Theta\}$.

In the paper, we study the convergence rate of the MLE under the following two settings:

- **Exact-fitted settings**: when k_* is known, we set $k = k_*$;
- **Over-fitted settings**: when k_* becomes unknown, we set $k > k_*$.

Density Estimation Rate

Proposition 1.

Under the Hellinger distance $h(\cdot,\cdot)$, the density estimation $g_{\widehat{G}_n}(Y|X)$ converges to the true density $g_{G_*}(Y|X)$ at the following rate:

$$\mathbb{P}\left(\mathbb{E}_X[h(g_{\widehat{G}_n}(\cdot|X), g_{G_*}(\cdot|X))] > C\sqrt{\log(n)/n}\right) \lesssim n^{-c}, \tag{3}$$

where c and C are universal constants.

- Under either the exact-fitted or over-fitted settings, the density estimation rate is of order $\mathcal{O}(n^{-1/2})$ (up to some logarithmic factor), which is parametric on the sample size.

Voronoi cells. A Voronoi cell of G generated by generated by the true component $\omega_j^* := (\beta_{1j}^*, a_j^*, b_j^*, \sigma_j^*)$ of G_*, for $1 \leq j \leq k_*$, is defined as

$$\mathcal{A}_j \equiv \mathcal{A}_j(G) := \{i \in \{1, 2, \ldots, k\} : \|\omega_i - \omega_j^*\| \leq \|\omega_i - \omega_\ell^*\|, \ \forall \ell \neq j\}, \tag{4}$$

where $\omega_i := (\beta_{1i}, a_i, b_i, \sigma_i)$.

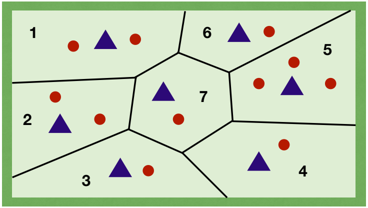

Figure: Illustration of Voronoi cells. Blue triangles represent for the components ω_j^* of G_* (true components), while red rounds stand for the components ω_i of G (fitted components).

Voronoi loss. Then, the loss function of interest is

$$\mathcal{D}_1(G, G_*) := \inf_{t_1, t_2} \sum_{j=1}^{k_*} \Bigg[\sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) \|(\Delta_{t_2}\beta_{1ij}, \Delta a_{ij}, \Delta b_{ij}, \Delta \sigma_{ij})\| + \Big| \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta_{0j}^* + t_1) \Big| \Bigg], \tag{5}$$

where $\Delta_{t_2}\beta_{1ij} := \beta_{1i} - \beta_{1j}^* - t_2$, $\Delta a_{ij} := a_i - a_j^*$, $\Delta b_{ij} := b_i - b_j^*$ and $\Delta \sigma_{ij} := \sigma_i - \sigma_j^*$.

Exact-fitted Settings

Theorem 1.

Given the exact-fitted settings, i.e., $k = k_*$, we find that

$$\mathbb{E}_X[h(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_1(G, G_*), \quad (6)$$

for any $G \in \mathcal{E}_{k_*}(\Theta) := \mathcal{O}_{k_*}(\Theta) \setminus \mathcal{O}_{k_*-1}(\Theta)$. As a result, there exist universal constants $C_1 > 0$ and $c_1 > 0$ such that:

$$\mathbb{P}\Big(\mathcal{D}_1(\widehat{G}_n, G_*) > C_1\sqrt{\log(n)/n}\Big) \lesssim n^{-c_1}. \quad (7)$$

Setting	**Loss Function**	$g_{G_*}(Y\|X)$	$\exp(\beta^*_{0j})$	β^*_{1j}, b^*_j	a^*_j, σ^*_j
Exact-fitted	$\mathcal{D}_1$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/2})$

Figure: Summary of the convergence rates of density estimation and parameter estimation under the exact-fitted settings.

Over-fitted Settings

Main Challenges. To establish the Hellinger lower bound, we use the Taylor expansion to decompose the term $g_{\widehat{G}_n}(Y|X) - g_{G_*}(Y|X)$ into a combination of linearly independent elements.

However, there are two interactions among softmax gating and expert parameters via the following partial differential equations (PDEs):

$$\frac{\partial^2 u}{\partial \beta_1 \, \partial b} = \frac{\partial u}{\partial a}; \qquad \frac{\partial^2 u}{\partial b^2} = 2 \, \frac{\partial u}{\partial \sigma}, \tag{8}$$

where $u(Y|X; \beta_1, a, b, \sigma) := \exp(\beta_1^\top X) \cdot f(Y|a^\top X + b, \sigma)$. The above PDEs lead to a number of linearly dependent derivative terms.

Over-fitted Settings

Voronoi loss. The loss function of interest is given by

$$\mathcal{D}_2(G, G_*) := \inf_{t_1, t_2} \Bigg\{ \sum_{\substack{j:|\mathcal{A}_j|=1, \\ i \in \mathcal{A}_j}} \exp(\beta_{0i}) \|(\Delta_{t_2}\beta_{1ij}, \Delta a_{ij}, \Delta b_{ij}, \Delta\sigma_{ij})\|$$

$$+ \sum_{\substack{j:|\mathcal{A}_j|>1, \\ i \in \mathcal{A}_j}} \exp(\beta_{0i}) \Big(\|(\Delta_{t_2}\beta_{1ij}, \Delta b_{ij})\|^{\bar{r}(|\mathcal{A}_j|)} + \|(\Delta a_{ij}, \Delta\sigma_{ij})\|^{\bar{r}(|\mathcal{A}_j|)/2} \Big)$$

$$+ \sum_{j=1}^{k_*} \Big| \sum_{i \in \mathcal{A}_j} \exp(\beta_{0i}) - \exp(\beta^*_{0j} + t_1) \Big| \Bigg\}. \tag{9}$$

Lemma 1.

For any $d \geq 1$, we have $\bar{r}(2) = 4$ and $\bar{r}(3) = 6$. We conjecture that $\bar{r}(m) = 2m$.

Over-fitted Settings

Theorem 2.

Under the over-fitted settings, i.e. $k > k_*$, we obtain that

$$\mathbb{E}_X[h(g_G(\cdot|X), g_{G_*}(\cdot|X))] \gtrsim \mathcal{D}_2(G, G_*), \tag{10}$$

for any $G \in \mathcal{O}_k(\Theta)$. Consequently, there exist universal constants $C_2 > 0$ and $c_2 > 0$ such that

$$\mathbb{P}\Big(\mathcal{D}_2(\widehat{G}_n, G_*) > C_2\sqrt{\log(n)/n}\Big) \lesssim n^{-c_2}. \tag{11}$$

Setting	**Loss Function**	$g_{G_*}(Y\|X)$	$\exp(\beta_{0j}^*)$	β_{1j}^*, b_j^*	a_j^*, σ_j^*
Over-fitted	$\mathcal{D}_2$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/2})$	$\mathcal{O}(n^{-1/2\bar{r}(\|\mathcal{A}_j\|)})$	$\mathcal{O}(n^{-1/\bar{r}(\|\mathcal{A}_j\|)})$

Figure: Summary of the convergence rates of density estimation and parameter estimation under the over-fitted settings.