November 2024

TEXAS
The University of Texas at Austin

Sigmoid Gating is More Sample Efficient than Softmax Gating in Mixture of Experts

Huy Nguyen, Nhat Ho, Alessandro Rinaldo**

Presenter: Huy Nguyen

PhD Student, Department of Statistics and Data Sciences,

The University of Texas at Austin

1. Introduction

- **Mixture of experts (MoE)** [1] aggregates multiple sub-models called experts, each specializes in a region of the input space, to jointly perform a task.

- For each input, the gating network guarantees that the most relevant experts are assigned more weights.

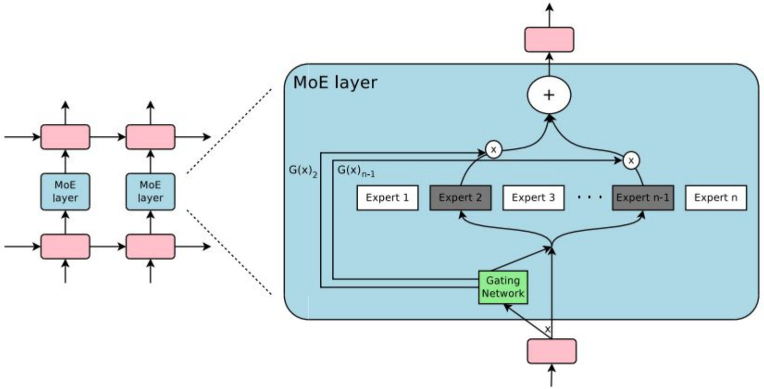

Figure 1: Illustration of the Mixture of Experts.

[1] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. *Adaptive mixtures of local experts*. Neural Computation, 1991.

1. Introduction

- **MoE with softmax gating function:**

$$f_{G_*}(x) := \sum_{i=1}^{k_*} \frac{\exp((\beta_{1i}^*)^\top x + \beta_{0i}^*)}{\sum_{j=1}^{k_*} \exp((\beta_{1j}^*)^\top x + \beta_{0j}^*)} \cdot h(x, \eta_i^*),$$

- **Softmax gating issue:** introduces an unexpected competition among experts, that is, when the weight of one expert increases, those of the others decrease accordingly → expert collapse phenomenon.

1. Introduction

- **Solution to the expert collapse issue:** MoE with sigmoid gating [2]

$$f_{G_*}(x) := \sum_{i=1}^{k_*} \frac{1}{1+\exp(-(\beta_{1i}^*)^\top x - \beta_{0i}^*)} \cdot h(x, \eta_i^*),$$

- **Question:** Under the expert estimation problem, is the sigmoid gating function more sample efficient than the softmax gating function?

- → We consider an MoE-based regression problem.

[2] R. Csordás, K. Irie, and J. Schmidhuber. *Approximating two-layer feedforward networks for efficient transformers*. EMNLP Findings, 2023.

2. Preliminaries

- **The inputs** are sampled from some known probability distribution: $X_1, X_2, \dots, X_n \overset{i.i.d}{\sim} \mu$
- **The outputs** are generated according to the model

$$Y_i = f_{G_*}(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent Gaussian noise variables $\varepsilon_i | X_i \sim \mathcal{N}(0, \nu)$ and

$$f_{G_*}(x) := \sum_{i=1}^{k_*} \frac{1}{1 + \exp(-(\beta_{1i}^*)^\top x - \beta_{0i}^*)} \cdot h(x, \eta_i^*),$$

- Above, k_* is the true number of experts.
- $G_* := \sum_{i=1}^{k_*} \frac{1}{1 + \exp(-\beta_{0i}^*)} \delta_{(\beta_{1i}^*, \eta_i^*)}$ is a mixing measure with unknown parameters $(\beta_{0i}^*, \beta_{1i}^*, \eta_i^*)_{i=1}^{k_*}$.

2. Preliminaries

- **Least square estimation:** we estimate the unknown parameters $(\beta_{0i}^*, \beta_{1i}^*, \eta_i^*)_{i=1}^{k_*}$ through the unknown mixing measure $G_* := \sum_{i=1}^{k_*} \frac{1}{1+\exp(-\beta_{0i}^*)} \delta_{(\beta_{1i}^*, \eta_i^*)}$ as follows:

$$\widehat{G}_n := \arg\min_{G \in \mathcal{G}_k(\Theta)} \sum_{i=1}^{n} \Big(Y_i - f_G(X_i) \Big)^2,$$

where $\mathcal{G}_k(\Theta) := \{G = \sum_{i=1}^{k'} \frac{1}{1+\exp(-\beta_{0i})} \delta_{(\beta_{1i}, \eta_i)} : 1 \leq k' \leq k, \ (\beta_{0i}, \beta_{1i}, \eta_i) \in \Theta\}$

denotes the set of mixing measures with at most k components with $k > k_*$.

2. Preliminaries

- **Challenges:** Since $k > k_*$, there must be some true atom fitted by at least two atoms. Assume $(\hat{\beta}_{1i}^n, \hat{\eta}_i^n) \to (\beta_{11}^*, \eta_1^*)$ for $i \in \{1, 2\}$, then to ensure the convergence of the regression function, the following gating convergence must hold for almost every x:

$$\sum_{i=1}^{2} \frac{1}{1 + \exp(-(\hat{\beta}_{1i}^n)^\top x - \hat{\beta}_{0i}^n)} \to \frac{1}{1 + \exp(-(\beta_{11}^*)^\top x - \beta_{01}^*)},$$

as $n \to \infty$. This limit is attained iff $\beta_{11}^* = 0_d$

❖ **Regime 1:** All the over-specified parameters β_{1i}^* are equal to zero;
❖ **Regime 2:** At least one among the over-specified parameters β_{1i}^* is non-zero.

3. Regression Function Estimation

- **Under the Regime 1**, the regression estimation rate is parametric on the sample size

$$\|f_{\widehat{G}_n} - f_{G_*}\|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}}).$$

- **Under the Regime 2**, since the gating convergence does not hold, the regression estimation cannot converge to the true regression function. Instead, we have

$$\inf_{\overline{G}\in\overline{\mathcal{G}}_k(\Theta)} \|f_{\widehat{G}_n} - f_{\overline{G}}\|_{L^2(\mu)} = \mathcal{O}_P([\log(n)/n]^{\frac{1}{2}})$$

where $\overline{G} \in \overline{\mathcal{G}}_k(\Theta) := \arg\min_{G\in\mathcal{G}_k(\Theta)\setminus\mathcal{G}_{k_*}(\Theta)} \|f_G - f_{G_*}\|_{L^2(\mu)}$.

4. Expert Estimation - Regime 1

- **Summary of expert estimation rates** for

  1. Strongly identifiable experts (ReLU and GELU experts);

  2. Non-strongly identifiable experts (polynomial experts and input-free experts).

	ReLU, GELU Experts	**Polynomial Experts**	**Input-independent Experts**
Sigmoid	$\mathcal{O}_P(n^{-1/4})$	$\mathcal{O}_P(1/\log(n))$	$\mathcal{O}_P(1/\log(n))$
Softmax [3]	$\mathcal{O}_P(n^{-1/4})$	$\mathcal{O}_P(1/\log(n))$	$\mathcal{O}_P(1/\log(n))$

[3] H. Nguyen, N. Ho, A. Rinaldo. *On least square estimation in softmax gating mixture of experts.* In ICML, 2024.

4. Expert Estimation - Regime 2

- **Summary of expert estimation rates** for
  1. Weakly identifiable experts (ReLU, GELU and polynomial experts);
  2. Non-strongly identifiable experts (input-free experts).

	ReLU, GELU Experts	**Polynomial Experts**	**Input-independent Experts**
Sigmoid	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(n^{-1/2})$	$\mathcal{O}_P(1/\log(n))$
Softmax [3]	$\mathcal{O}_P(n^{-1/4})$	$\mathcal{O}_P(1/\log(n))$	$\mathcal{O}_P(1/\log(n))$

[3] H. Nguyen, N. Ho, A. Rinaldo. *On least square estimation in softmax gating mixture of experts.* In ICML, 2024.

5. Conclusion

- From the perspective of the expert estimation problem in the MoE-type regression, we observe that:

  ❖ The sigmoid gating is more sample efficient than the softmax gating;

  ❖ The sigmoid gating is compatible with a broader class of experts than the softmax gating.

THANK YOU!