Breaking the gridlock in Mixture-of-Experts: Consistent and Efficient algorithms

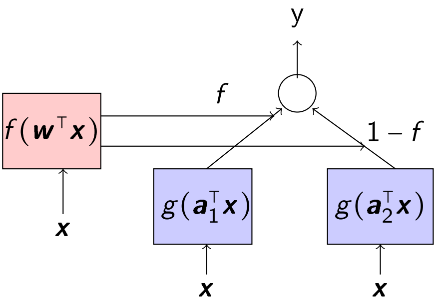
Ashok Vardhan Makkuva

University of Illinois at Urbana-Champaign

Joint work with
Sewoong Oh, Sreeram Kannan, Pramod Viswanath

Mixture-of-Experts (MoE)

- Jacobs, Jordan, Nowlan and Hinton, 1991



$f = \mathrm{sigmoid},\ g = \mathrm{linear}, \tanh, \mathrm{ReLU}, \mathrm{leakyReLU}$

Motivation-I: Modern relevance of MoE

- Outrageously large neural networks

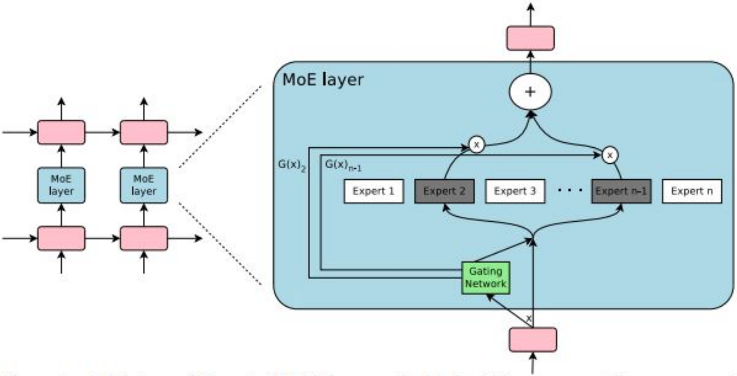


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

Motivation-II: Gated RNNs

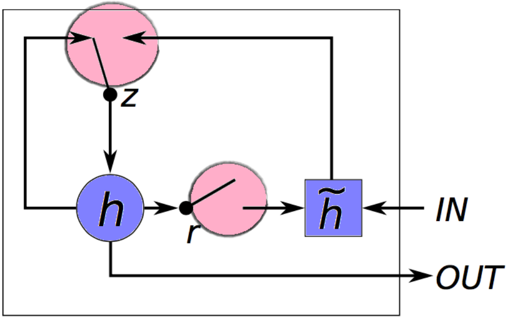
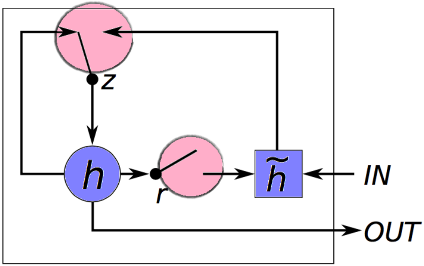


Figure: Gated Recurrent Unit (GRU)

Key features:

- **Gating** mechanism
- Long term memory

Motivation-II: GRU



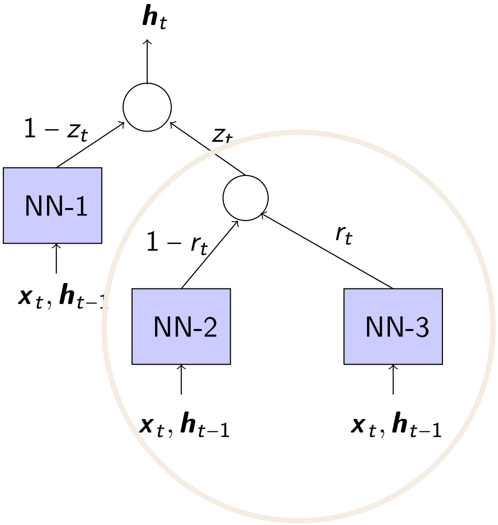
- **Gates:** $z_t, r_t \in [0,1]^d$ depend on the input x_t and the past h_{t-1}
- **States:** $h_t, \tilde{h}_t \in \mathbb{R}^d$

Update equations for each t:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\tilde{h}_t = f(Ax_t + r_t \odot Bh_{t-1})$$

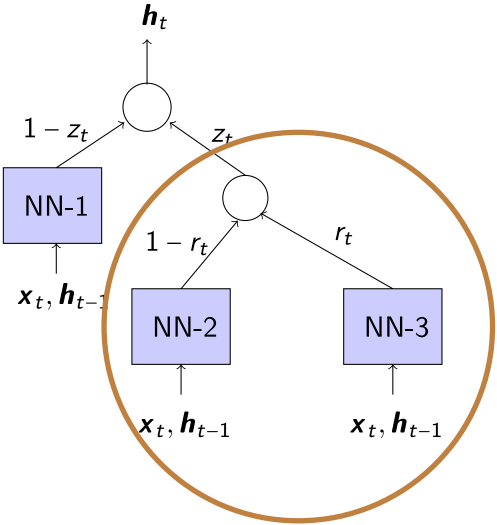
MoE: Building blocks of GRU

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot (1 - r_t) \odot f(Ax_t) + z_t \odot r_t \odot f(Ax_t + Bh_{t-1})$$



MoE: Building blocks of GRU

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot (1 - r_t) \odot f(Ax_t) + z_t \odot r_t \odot f(Ax_t + Bh_{t-1})$$



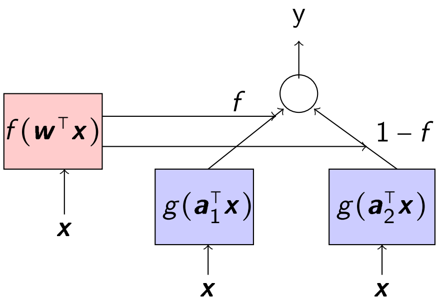
What is known about MoE?

Adaptive mixtures of local experts RA Jacobs, MI Jordan, SJ Nowlan, GE Hinton Neural computation 3 (1), 79-87	3663	1991
Sharing clusters among related groups: Hierarchical Dirichlet processes YW Teh, MI Jordan, MJ Beal, DM Blei Advances in neural information processing systems, 1385-1392	3273	2005
Hierarchical mixtures of experts and the EM algorithm MI Jordan, RA Jacobs Neural computation 6 (2), 181-214	3090	1994

- No provable learning algorithms for parameters[1] ☹

[1] 20 years of MoE, MoE: a literature survey

Open problem for 25+ years



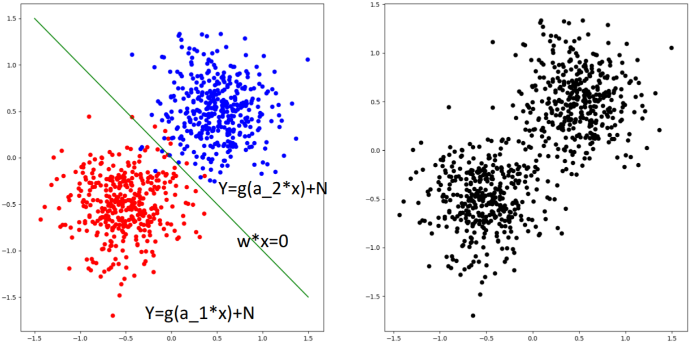
$$\Leftrightarrow P_{y|\boldsymbol{x}} = f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \mathcal{N}(y|g(\boldsymbol{a}_1^\top \boldsymbol{x}), \sigma^2) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x})) \cdot \mathcal{N}(y|g(\boldsymbol{a}_2^\top \boldsymbol{x}), \sigma^2)$$

Open question

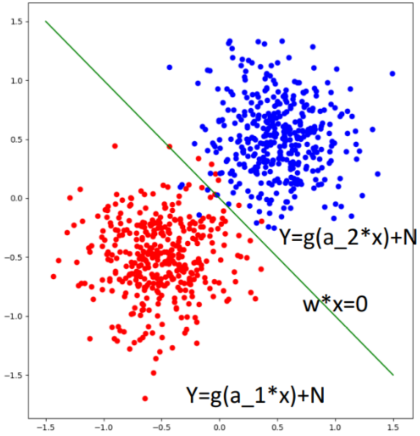
Given n i.i.d. samples $(\boldsymbol{x}^{(i)}, y^{(i)})$, does there exist an efficient learning algorithm with provable theoretical guarantees to learn the regressors $\boldsymbol{a}_1, \boldsymbol{a}_2$ and the gating parameter $\boldsymbol{w}$?

Modular structure

Mixture of classification ($\boldsymbol{w}$) and regression ($\boldsymbol{a}_1, \boldsymbol{a}_2$) problems



Key observation



Key observation

If we know the regressors, learning the gating parameter is easy and vice-versa. How to break the gridlock?

Breaking the gridlock: An overview

Recall the model for MoE:

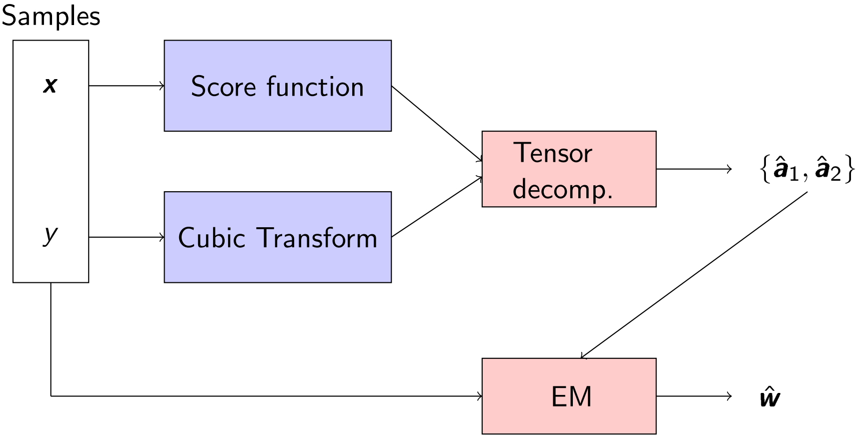
$$P_{y|\boldsymbol{x}} = f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \mathcal{N}(y|g(\boldsymbol{a}_1^\top \boldsymbol{x}), \sigma^2) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x})) \cdot \mathcal{N}(y|g(\boldsymbol{a}_2^\top \boldsymbol{x}), \sigma^2)$$

Main message

We propose a novel algorithm with first recoverable guarantees

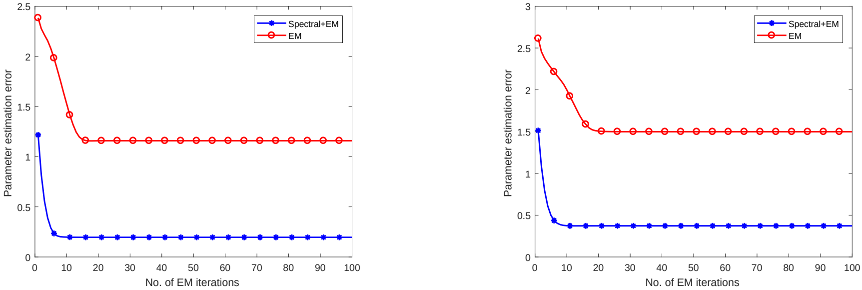
- We learn $(\boldsymbol{a}_1, \boldsymbol{a}_2)$ and $\boldsymbol{w}$ separately
- First recover $(\boldsymbol{a}_1, \boldsymbol{a}_2)$ without knowing $\boldsymbol{w}$ at all
- Later learn $\boldsymbol{w}$ using traditional methods like EM
- Global consistency guarantees (population setting)

Algorithm



Samples
$\boldsymbol{x}$
y
Score function
Cubic Transform
Tensor decomp.
$\{\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{a}}_2\}$
EM
$\hat{\boldsymbol{w}}$

Comparison with EM



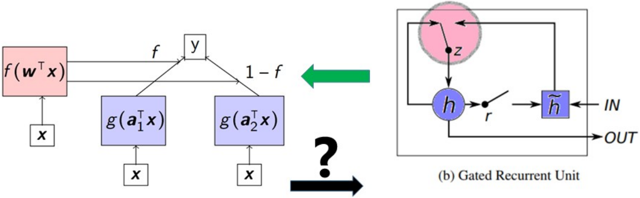
(a) 3 mixtures (b) 4 mixtures

Figure: Plot of parameter estimation error

Summary

- Algorithmic innovation: First provably consistent algorithms for MoE in 25+ years
- Global convergence: Our algorithms work with global initializations

Conclusion



(b) Gated Recurrent Unit

Mixture-of-Experts

1. Theoretical understanding ✓
2. Novel algorithms ✓

Gated recurrent networks

1. Theoretical understanding **?**
2. Algorithms **?**

Poster #210

Thank you!

Poster #210

Thank you!