The University of Texas at Austin
Cockrell School of Engineering

June 2025

LEARNING MIXTURES OF EXPERTS WITH EM

A Mirror Descent Perspective

Quentin Fruytier
PhD Student, The University of Texas at Austin

Table of Contents

Quentin Fruytier
PhD Student, ECE

qdf76@my.utexas.edu

Aryan Mokhtari
Assistant Professor, ECE.

mokhtari@austin.utexas.edu

Sujay Sanghavi
Associate Professor, ECE

sanghavi@mail.utexas.edu

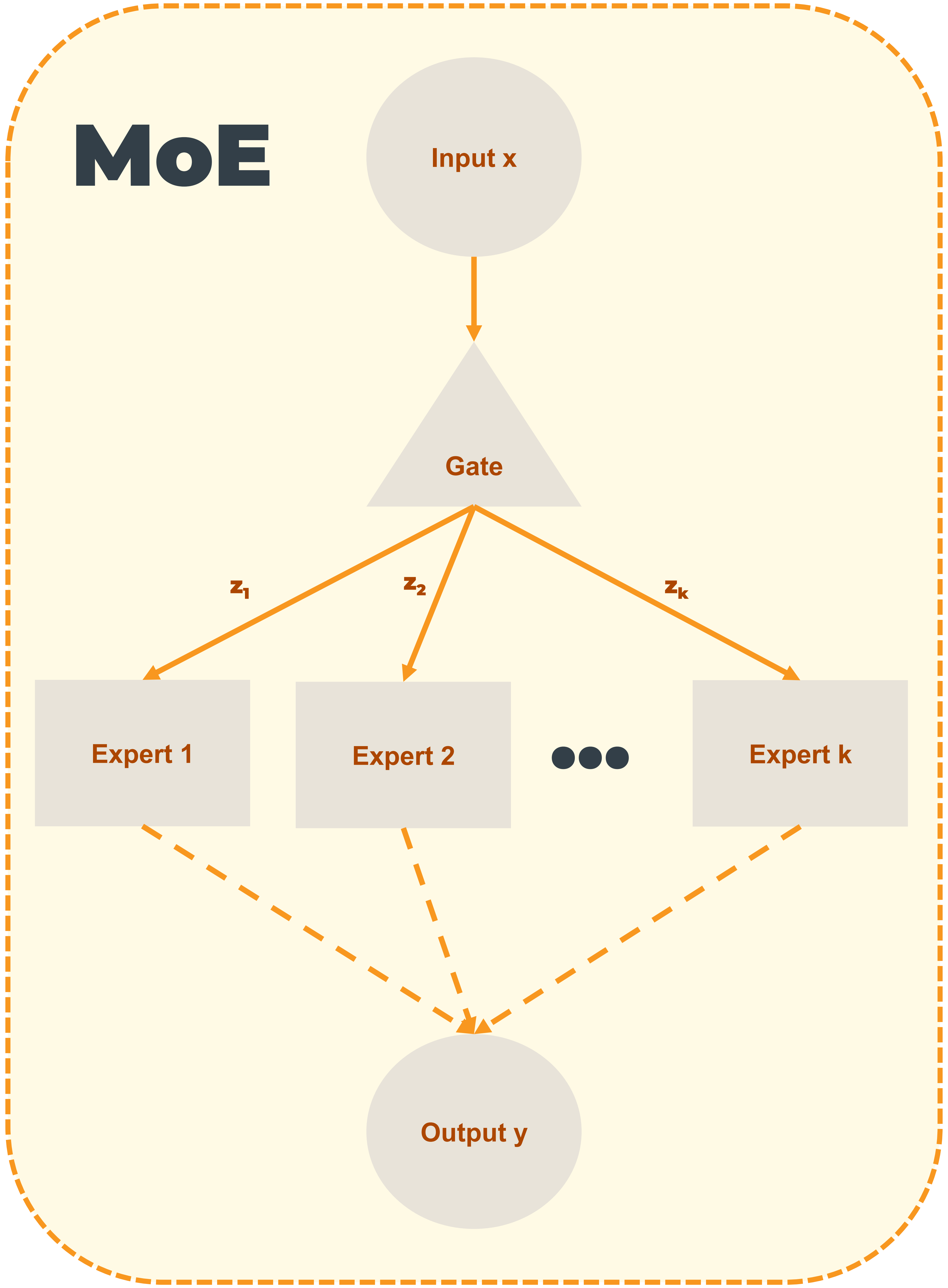
What is a Mixture of Experts (MoE)

Quick Recap:
- MoE **splits the input space** via a "Gate" function and assigns parts to **specialized models** (experts).
- Popular "Gate" function is the softmax given by

$$P(z = i|\boldsymbol{x}; \boldsymbol{w}^*) = \frac{e^{\boldsymbol{x}^\top \boldsymbol{w}_i^*}}{\sum_{j\in[k]} e^{\boldsymbol{x}^\top \boldsymbol{w}_j^*}}, \qquad i \in [k].$$

Training:
- Gradient Descent like methods on log-likelihood.

Our focus:
- Can the classical **Expectation-Maximization (EM)** Algorithm do better?



What is Expectation-Maximization (EM)

EM takes a structured approach to minimizing the negative log-likelihood objective.

- **E-Step:**
  - Compute **expectation of complete data** (**x**,y,z) log-likelihood with respect to the latent variable z conditioned on observable data and current model parameters.
- **M-Step:**
  - Solve the **Maximization (or minimization)** problem for this expectation (or the negative expectation)

Algorithm 1 EM for Mixture of Experts

1: **Input:** Initial $\boldsymbol{\theta}^1 \in \Omega$, data: $(\boldsymbol{X}, Y) \sim p(\boldsymbol{x}, y; \boldsymbol{\theta}^*)$
2: **for** $t = 1$ to T **do**
3: $\quad$ **$\boldsymbol{\theta}$-Update:** Obtain $\boldsymbol{\theta}^{t+1}$ as
4: $\quad\quad$ $\boldsymbol{\theta}^{t+1} \leftarrow \arg\min_{\boldsymbol{\theta} \in \Omega} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t)$
5: **end for**
6: **Output:** $\boldsymbol{\theta}^T = (\boldsymbol{w}^T, \boldsymbol{\beta}^T)$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = -\mathbb{E}_{X,Y}\left[\mathbb{E}_{Z|\boldsymbol{x},y;\boldsymbol{\theta}^t}[\log p(\boldsymbol{x}, y, z; \boldsymbol{\theta})]\right]$$

$$\boldsymbol{w}^{t+1} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} -\mathbb{E}_{\boldsymbol{X},Y}\left[\mathbb{E}_{Z|\boldsymbol{x},y;\boldsymbol{\theta}^t}\left[\log p(z|\boldsymbol{x}; \boldsymbol{w})\right]\right],$$
$$\boldsymbol{\beta}^{t+1} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} -\mathbb{E}_{\boldsymbol{X},Y}\left[\mathbb{E}_{Z|\boldsymbol{x},y;\boldsymbol{\theta}^t}\left[\log p(y|z, \boldsymbol{x}; \boldsymbol{\beta})\right]\right].$$

Theoretical Contributions

Theorem 4.1 [informal]: For a general class of MoE models, the iterations of EM are directly equivalent to projected Mirror Descent with unit step size and Kullback Leibler divergence regularizer.

Theorem 4.2 [informal]: For a general class of MoE models, the iterations of EM are

i. Always at least guaranteed to convergence to a stationary point sub-linearly
ii. Converge sub-linearly (or linearly) to the true parameters under suitable initialization in a region that satisfies specific convexity properties.

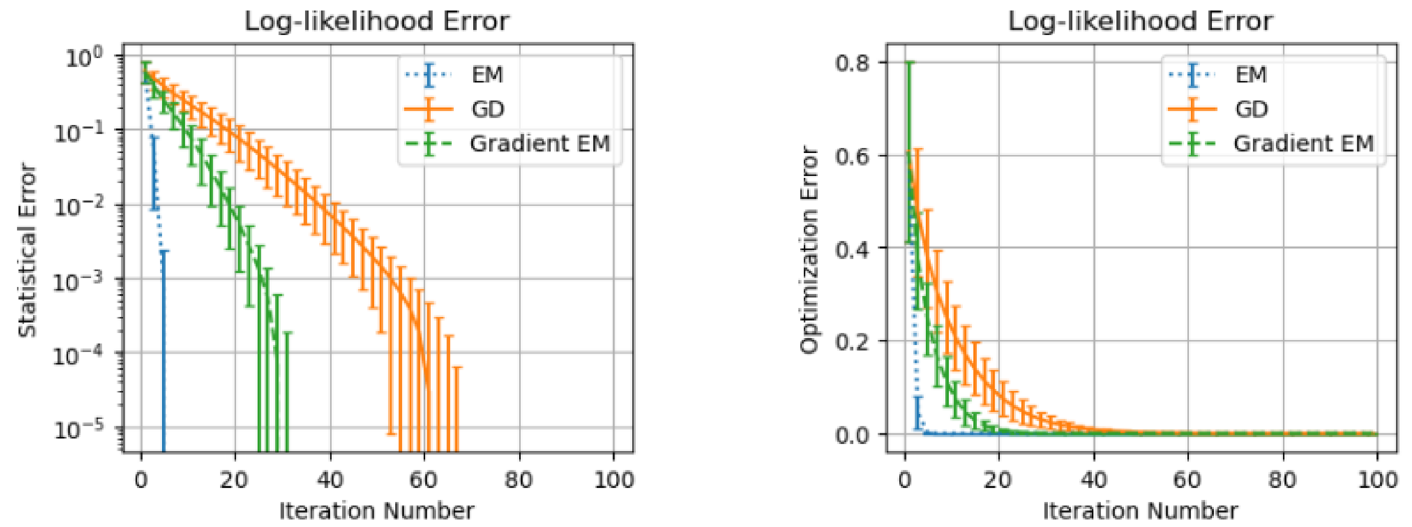
Theorem 5.1 [informal]: For special 2-component mixture of Linear (or Logistic) Experts, the iterations of EM are directly equivalent to Mirror Descent (no projection) with unit step size and Kullback Leibler divergence regularizer.

Following Results for this specific case [informal]:

a. **Corollary B.1:** Recover sufficient conditions for convergence
b. **Theorem B.2:** Link conditions to top eigen values of Missing Information Matrix (MIM)
c. **Theorem B.4:** Link conditions to Signal to Noise Ratio (SNR) of the true model

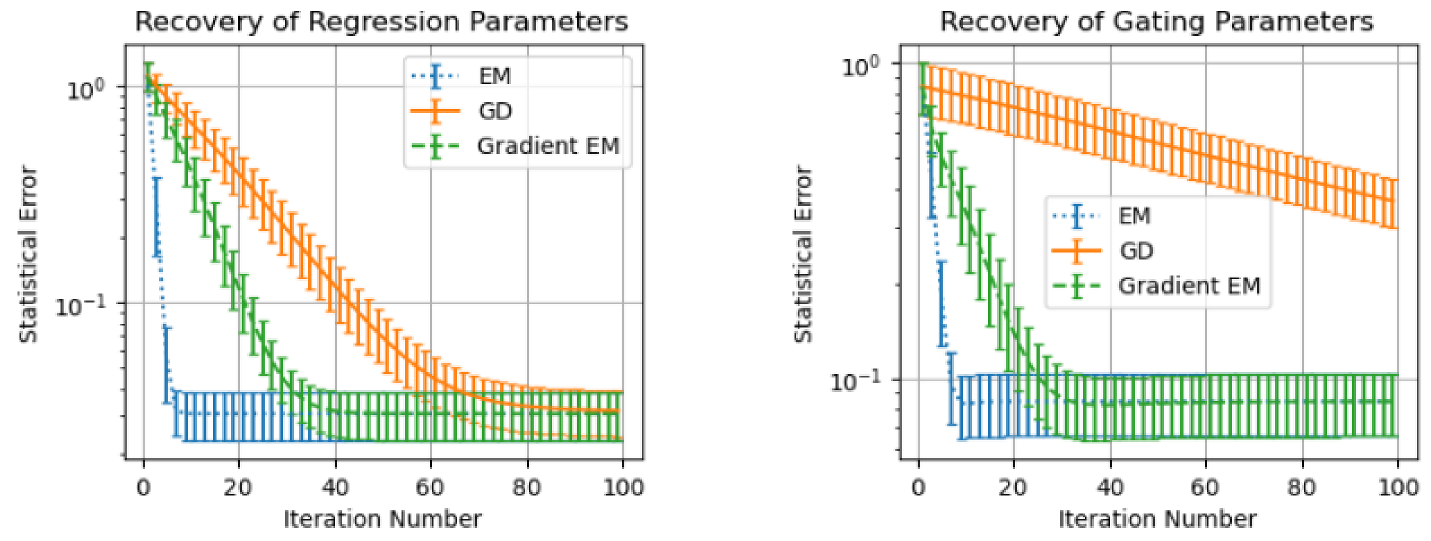
Empirical Validation

Symmetric Mixture of Linear Experts (Synthetic):



(a) (b)

Figure 1: Convergence of objective errors $\mathcal{L}(\boldsymbol{\theta}^t) - \mathcal{L}(\boldsymbol{\theta}^*)$ and $\mathcal{L}(\boldsymbol{\theta}^t) - \mathcal{L}(\boldsymbol{\theta}^T)$ in Fig 1a and Fig 1b, respectively, averaged over 50 instances when fitting a SymMoLinE.



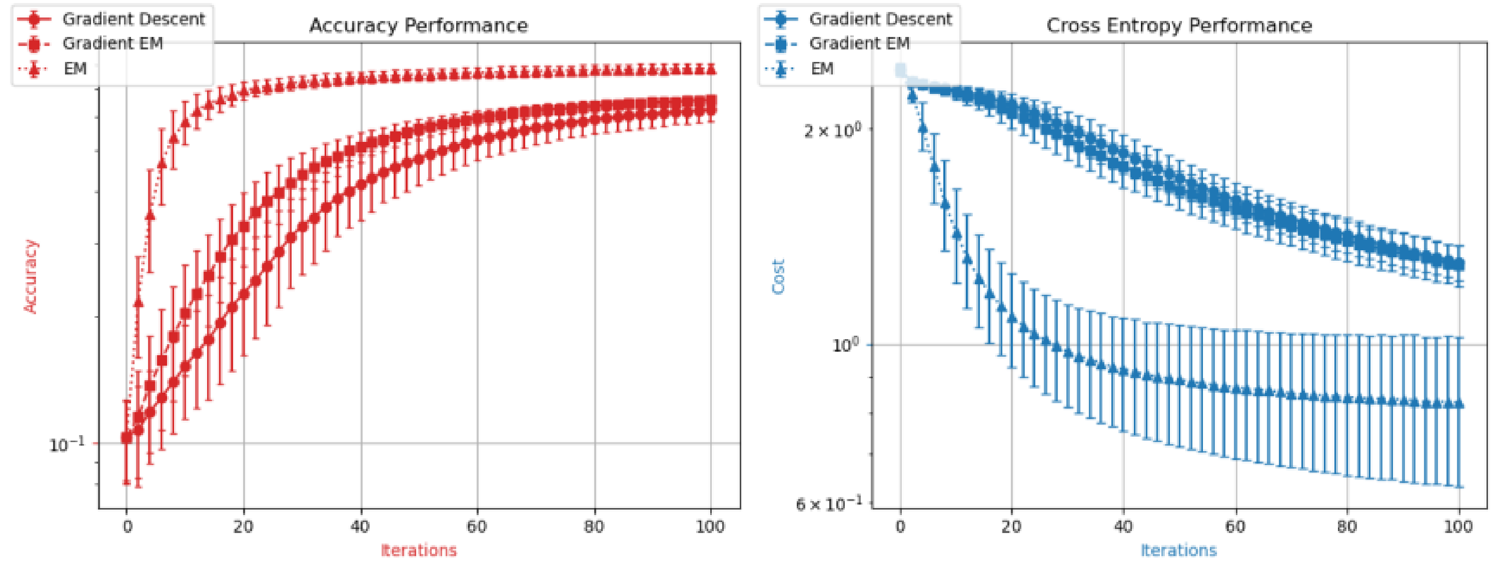
(a) (b)

Figure 2: This figure shows the progress made towards the true parameters, $\frac{\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\|_2}{\|\boldsymbol{\beta}^*\|_2}$ and $\frac{\|\boldsymbol{w}^t - \boldsymbol{w}^*\|_2}{\|\boldsymbol{w}^*\|_2}$ in figures 2a and 2b respectively, averaged over 50 instances when fitting a SymMoLinE

Mixture of 2 Logistic Experts (FMNIST):

Table 2: Performance for 2-Component MoLogE

	Accuracy	Cross Entropy
EM	*78.5%*	*0.827*
Gradient EM	66.0%	1.29
Gradient Descent	62.4%	1.30



(a) (b)

Figure 3: Test accuracy and objective function, $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{y}_i = y_i}$ and $\mathcal{L}(\boldsymbol{\theta}^t)$ in 3a and 3b, respectively, averaged over 25 instances for a 2-component MoLogE train on Random Invert FMNIST.

Conclusion

Takeaway: EM isn't outdated—it's an MD algorithm in disguise with strong convergence properties. This paper aims to:

- ➢ Offer a principled, optimization-theoretic interpretation of EM
- ➢ Unify prior scattered convergence results.
- ➢ Reveal when and why EM converges and at what rate.
- ➢ Validate theoretical guarantees empirically.

Impact: Better understanding of EM and tuning of latent variable models like MoE.

Future Work:

- Scalable EM via mini-batch paradigm.
- Extensions to Deep and Sparse MoE.

THANK YOU.

- Quentin Fruytier, qdf76@my.utexas.edu
- Aryan Mokhtari, mokhtari@austin.utexas.edu
- Sujay Sanghavi, sanghavi@mail.utexas.edu