Learning Mixtures of Experts with EM

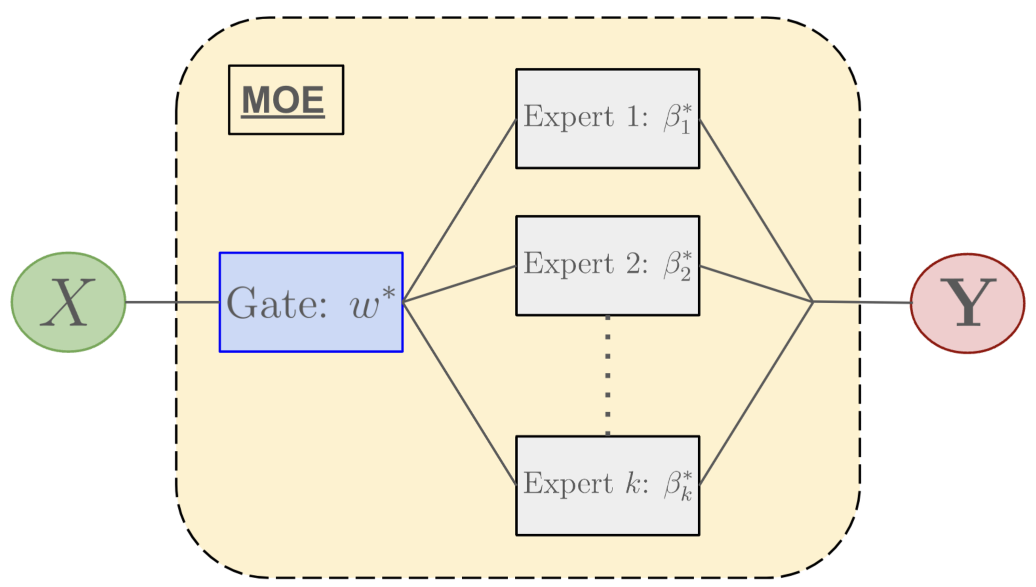
Quentin Fruytier *, Aryan Mokhtari*, and Sujay Sanghavi *

*Department of Electrical and Computer Engineering, The University of Texas at Austin

Mixtures of Experts (MoE)

- Why MoE? Increase model parameters for fixed training and inference costs.
- Many applications:
  - ⇒ LLM: Mixtral (2024), DeepSeek-V3 (2024)
  - ⇒ Transformers: Switch Transformers (2022)
- Assumed Generative Model:

$$p(\boldsymbol{x}, y) = p(\boldsymbol{x}) \sum_{z\in[k]} p(y|\boldsymbol{x}, z) P(z|\boldsymbol{x}).$$



  - ⇒ $(\boldsymbol{x}, y) \in \mathbb{R}^{d\times 1}$: (feature,target) pair.
  - ⇒ $z \in [k]$: Unobserved expert label for $(\boldsymbol{x}, y)$ pair where

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

- **Goal**: Want to find the ground truth parameters $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{\beta}^*)$.
  - ⇒ Find the minimizers of the likelihood, $\mathcal{L}(\boldsymbol{\theta})$:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{X}}\left[\log p(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{X},Y}\left[\log\left(\sum_{z\in[k]} p(y|\boldsymbol{x}, z) P(z|\boldsymbol{x})\right)\right]$$

Expectation Maximization (EM) Algorithm

- **Motivation:**
  - ⇒ We know EM is powerful for learning Mixtures of Gaussians and Mixtures of Regressions, but we lack understanding for MoE.
  - ⇒ We know EM is equivalent to Mirror Descent for exponential family distributions, but this does not include MoE.
- **EM Algorithm for MoE**:
  - ⇒ Iterative global minimization of the EM objective, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$:

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

  - ⇒ EM objective linearly separable in $(\boldsymbol{w}, \boldsymbol{\beta})$:

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

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

EM is Mirror Descent for MoE

- **Mirror Descent (MD):**
  - ⇒ Bregman Divergence:

$$D_h(\boldsymbol{\theta}^t, \boldsymbol{\theta}) := h(\boldsymbol{\theta}) - h(\boldsymbol{\theta}^t) - \langle \nabla h(\boldsymbol{\theta}^t), \boldsymbol{\theta} - \boldsymbol{\theta}^t\rangle.$$

  - ⇒ Iterative global minimization of MD objective:

$$\mathcal{L}(\boldsymbol{\theta}^t) + \langle \nabla\mathcal{L}(\boldsymbol{\theta}^t), \boldsymbol{\theta} - \boldsymbol{\theta}^t\rangle + \frac{1}{\eta} D_h(\boldsymbol{\theta}^t, \boldsymbol{\theta}).$$

- **Symmetric Mixture of 2-Experts:** $\beta^* := \beta_1^* = -\beta_2^*$.
  - ⇒ Symmetric Linear Expert:

$$p(y|\boldsymbol{x}, z = i; \boldsymbol{\beta}_i^*) \propto \exp\left\{\frac{(y - \boldsymbol{x}^\top\boldsymbol{\beta}_i^*)^2}{2}\right\}$$

  - ⇒ Symmetric Logistic Expert:

$$P(y = 1|\boldsymbol{x}, z = i; \boldsymbol{\beta}_i^*) = \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta}_i^*)}{1 + \exp(\boldsymbol{x}^\top\boldsymbol{\beta}_i^*)}$$

Theorem(Simplified). *For $(\boldsymbol{x}, y)$ from a MoE where $y, z|\boldsymbol{x}$ is in an exponential family, the EM Algorithm is equivalent to projected Mirror Descent with unit stepsize and Kullback Leibler Divergence where there is some mirror map $A(\boldsymbol{\theta})$ such that $D_{KL}(\boldsymbol{\theta}_x, \boldsymbol{\phi}_x) = D_A(\boldsymbol{\phi}_x, \boldsymbol{\theta}_x)$.* For symmetric mixture of linear (or logistic) experts, the projection is trivial.

Convergence Analysis From an MD perspective

- **Local Average Convexity**: Convex set Θ containing $\boldsymbol{\theta}^1, \boldsymbol{\theta}^*$ such that for all $\boldsymbol{\phi}, \boldsymbol{\theta} \in \Theta$,

$$\mathcal{L}(\boldsymbol{\phi}) \geq \mathcal{L}(\boldsymbol{\theta}) + \mathbb{E}_X\left[\langle \nabla\mathcal{L}(\boldsymbol{\theta}_x), \boldsymbol{\phi}_x - \boldsymbol{\theta}_x\rangle\right].$$

- **Local Average Strong Relative Convexity**: Convex set Θ containing $\boldsymbol{\theta}^1, \boldsymbol{\theta}^*$ such that for all $\boldsymbol{\phi}, \boldsymbol{\theta} \in \Theta$,

$$\mathcal{L}(\boldsymbol{\phi}) \geq \mathcal{L}(\boldsymbol{\theta}) + \mathbb{E}_X\left[\langle \nabla\mathcal{L}(\boldsymbol{\theta}_x), \boldsymbol{\phi}_x - \boldsymbol{\theta}_x\rangle + \alpha D_h(\boldsymbol{\phi}_x, \boldsymbol{\theta}_x)\right].$$

Corollary(Simplified). *For $(\boldsymbol{x}, y)$ from a General MoE, the EM iterates $\{\boldsymbol{\theta}^t\}_{t\in[T]}$ satisfy:*

1) ***Stationarity.*** *For no additional conditions,*

$$\min_{t\in[T]} \mathbb{E}_X\left[D_{KL}(\boldsymbol{\theta}_{\boldsymbol{x}}^t, \boldsymbol{\theta}_{\boldsymbol{x}}^{t+1})\right] \leq \frac{\mathcal{L}(\boldsymbol{\theta}^1) - \mathcal{L}(\boldsymbol{\theta}^*)}{T}; \qquad (1)$$

2) ***Sub-linear Rate to $\boldsymbol{\theta}^*$.*** *If $\boldsymbol{\theta}^1$ is initialized in Θ, a locally convex region of $\mathcal{L}(\boldsymbol{\theta})$ containing $\boldsymbol{\theta}^*$, then*

$$\mathcal{L}(\boldsymbol{\theta}^T) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \frac{\mathbb{E}_X\left[D_{KL}(\boldsymbol{\theta}_{\boldsymbol{x}}^*, \boldsymbol{\theta}_{\boldsymbol{x}}^1)\right]}{T} \qquad (2)$$

3) ***Linear Rate to $\boldsymbol{\theta}^*$.*** *If $\boldsymbol{\theta}^1$ is initialized in $\Theta \subseteq \Omega$, a locally strongly convex region of $\mathcal{L}(\boldsymbol{\theta})$ relative to $A(\boldsymbol{\theta})$ that contains $\boldsymbol{\theta}^*$, then*

$$\mathcal{L}(\boldsymbol{\theta}^T) - \mathcal{L}(\boldsymbol{\theta}^*) \leq (1-\alpha)^T\left(\mathcal{L}(\boldsymbol{\theta}^1) - \mathcal{L}(\boldsymbol{\theta}^*)\right). \qquad (3)$$

Missing Information Matrix

- **Missing Information Matrix** ($\mathrm{M}(\boldsymbol{\theta})$):

$$\boldsymbol{M}(\boldsymbol{\theta}) = \boldsymbol{I}_{\boldsymbol{x},z,y|\boldsymbol{\theta}}^{-1}\boldsymbol{I}_{z|\boldsymbol{x},y,\boldsymbol{\theta}}$$

  - ⇒ $I_{\boldsymbol{x},z,y|\boldsymbol{\theta}}, I_{z|\boldsymbol{x},y,\boldsymbol{\theta}}$ are the fisher information matrices.
  - ⇒ In our setting,

$$\boldsymbol{I}_{\boldsymbol{x},z,y|\boldsymbol{\theta}} = \nabla^2 A(\boldsymbol{\theta})$$

$$\boldsymbol{I}_{z|\boldsymbol{x},y,\boldsymbol{\theta}} := -\mathbb{E}_{\boldsymbol{X},Y}\mathbb{E}_{Z|\boldsymbol{x},y,\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log P(z|\boldsymbol{x}, y; \boldsymbol{\theta})\right]$$

Theorem(Simplified). *For $(\boldsymbol{x}, y)$ from a symmetric mixture of 2 logistic experts (or 2 linear experts), the objective $\mathcal{L}(\boldsymbol{\theta})$ is α-strongly convex relative to the mirror map $A(\boldsymbol{\theta})$ on the convex set Θ if and only if*

$$\lambda_{\max}(\boldsymbol{M}(\boldsymbol{\theta})) \leq (1-\alpha) \text{ for all } \boldsymbol{\theta} \in \Theta.$$

- Can now obtain sufficient conditions on the Signal to Noise Ratio for the assumptions in part 2) and 3) to be satisfied.

Numerical Experiments

- **Altered FMNIST Experiment**:
  - ⇒ Randomly flip images from a white object on a black background to a black object on a white background.
  - ⇒ Train a Mixture of 2 Logistic Experts.

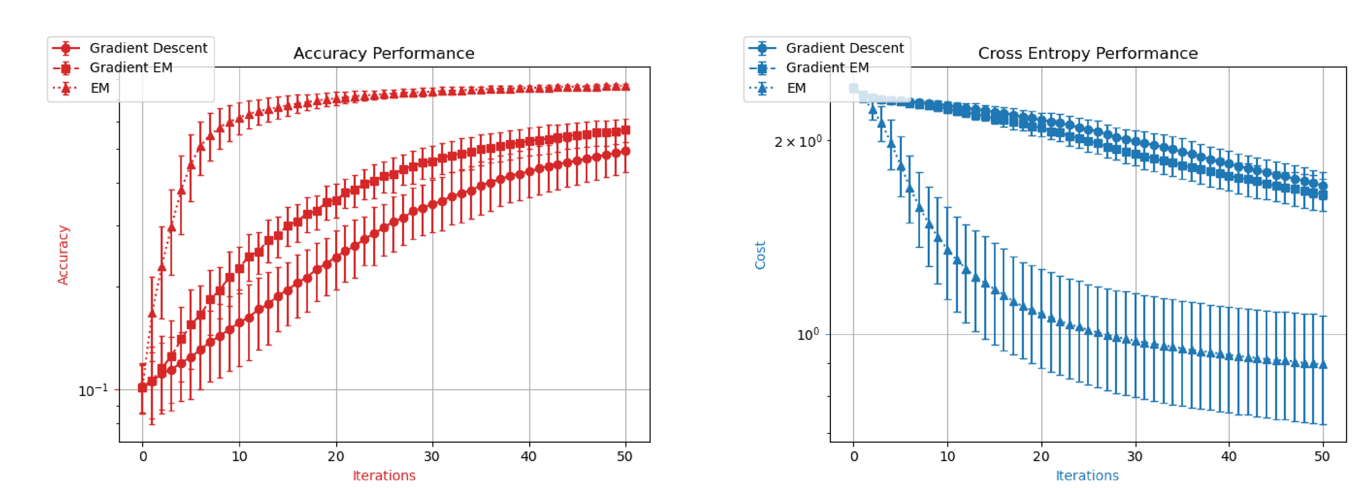


Figure 1: Mixture of 2 Logistic Experts for altered FMNIST dataset

- **Synthetic Experiment** on Symmetric Mixture of 2 Linear Experts.

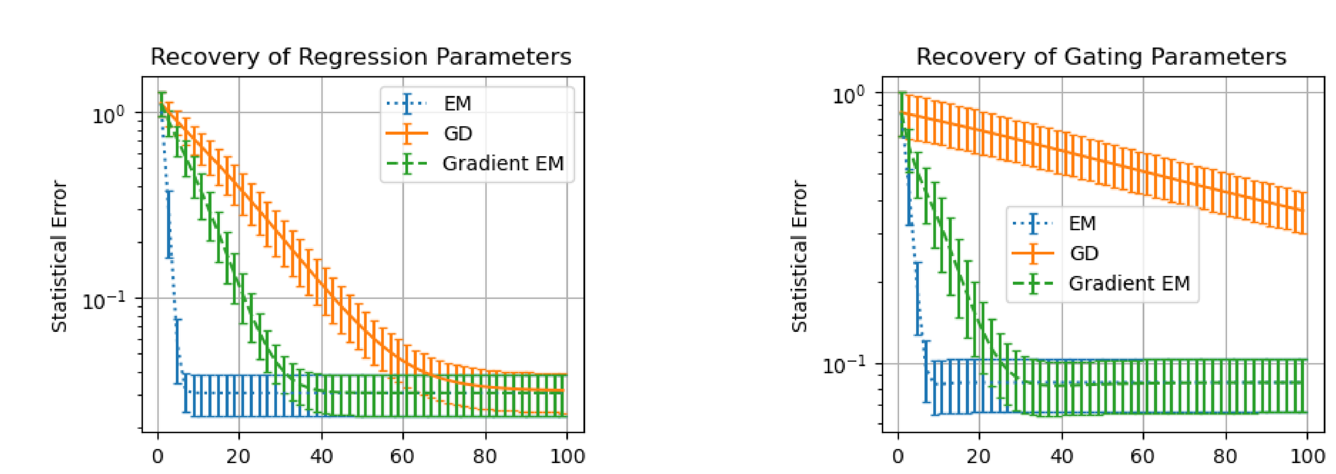


Figure 2: Symmetric Mixture of 2 Linear Experts