A Connection between k-MoE and other popular models

Relation to other mixture models. Notice that if let $w_i^* = 0$ in Eq. (1) for all $i \in [k]$, we recover the well-known uniform mixtures of *generalized linear models (GLMs)*. Similarly, allowing for bias parameters in Eq. (1), we can recover the generic mixtures of GLMs. Moreover, if we let g to be the linear function, we get the popular *mixtures of linear regressions* model. These observations highlight that MoE models are a far more stricter generalization of mixtures of GLMs since they allow the mixing probability $p_i^*(x)$ to depend on each input x in a parametric way. This makes the learning of the parameters far more challenging since the gating and expert parameters are inherently coupled.

Relation to feed-forward neural networks. Note that if we let $w_i^* = 0$ and allow for bias parameters in the soft-max probabilities in Eq. (1), taking conditional expectation on both sides yields

$$\hat{y}(x) \triangleq \mathbb{E}[y|x] = \sum_{i\in[k]} w_i^* g(\langle a_i^*, x\rangle), \quad \sum_i w_i^* = 1, w_i^* \in [0,1]. \tag{9}$$

Thus the mapping $x \mapsto \hat{y}(x)$ is exactly the same as that of a 1-hidden -layer neural network with activation function g if we restrict the output layer to positive weights. Thus k-MoE can also be viewed as a probabilistic model for *gated feed-forward networks*.

B Valid class of non-linearities

We slightly modify the class of non-linearities from Makkuva et al. (2019) for our theoretical results. The only key modification is that we use a fourth-order derivative based conditions, as opposed to third-order derivatives used in the above work. Following their notation, let $Z \sim \mathcal{N}(0,1)$ and $Y|Z \sim \mathcal{N}(g(Z), \sigma^2)$, where $g : \mathbb{R} \to \mathbb{R}$. For $(\alpha, \beta, \gamma, \delta) \in \mathbb{R}^4$, define

$$\mathcal{Q}_4(y) \triangleq Y^4 + \alpha Y^3 + \beta Y^2 + \gamma Y,$$

where

$$\mathcal{S}_4(Z) \triangleq \mathbb{E}[\mathcal{Q}_4(y)|Z] = g(Z)^4 + 6g(Z)^2\sigma^2 + \sigma^4 + \alpha(g(Z)^3 + 3g(Z)\sigma^2) + \beta(g(Z)^2 + \sigma^2) + \gamma g(Z).$$

Similarly, define

$$\mathcal{Q}_2(y) \triangleq Y^2 + \delta Y, \quad \mathcal{S}_2(Z) = \mathbb{E}[\mathcal{Q}_2(y)|Z] = g(Z)^2 + \delta g(Z) + \sigma^2.$$

Condition 1. $\mathbb{E}[\mathcal{S}_4'(Z)] = \mathbb{E}[\mathcal{S}_4''(Z)] = \mathbb{E}[\mathcal{S}_4'''(Z)] = 0$ and $\mathbb{E}[\mathcal{S}_4''''(Z)] \neq 0$. Or equivalently, in view of Stein's lemma Stein (1972),

$$\mathbb{E}[\mathcal{S}_4(Z)Z] = \mathbb{E}[\mathcal{S}_4(Z)(Z^2-1)] = \mathbb{E}[\mathcal{S}_4(Z)(Z^3-3Z)] = 0, \text{ and } \mathbb{E}[\mathcal{S}_4(Z)(Z^4-6Z^2+3)] \neq 0.$$

Condition 2. $\mathbb{E}[\mathcal{S}_2'(Z)] = 0$ and $\mathbb{E}[\mathcal{S}_2''(Z)] \neq 0$. Or equivalently,

$$\mathbb{E}[\mathcal{S}_2(Z)Z] = 0 \text{ and } \mathbb{E}[\mathcal{S}_2(Z)(Z^2-1)] \neq 0.$$

Definition 1. We say that the non-linearity g is $(\alpha, \beta, \gamma, \delta)-$valid if there exists a tuple $(\alpha, \beta, \gamma, \delta) \in \mathbb{R}^4$ such that both Condition 1 and Condition 2 are satisfied.

While these conditions might seem restrictive at first, all the widely used non-linearities such as Id, ReLU, leaky-ReLU, sigmoid, etc. belong to this. For some of these non-linear activations, we provide the pre-computed transformations below:

Example 1. If $g = \text{Id}$, then $\mathcal{S}_3(y) = y^4 - 6y^2(1+\sigma^2)$ and $\mathcal{Q}_2(y) = y^2$.

Example 2. If $g = \text{ReLU}$, i.e. $g(z) = \max\{0, z\}$, we have that for any $p, q \in \mathbb{N}$,

$$\mathbb{E}[g(Z)^p Z^q] = \int_0^\infty z^{p+q}\left(\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\right)dz = \frac{1}{2}\mathbb{E}[|Z|^{p+q}] = \frac{(p+q-1)!!}{2}\begin{cases}\sqrt{\frac{2}{\pi}} & \text{if } p+q \text{ is odd}\\ 1 & \text{if } p+q \text{ is even}\end{cases}.$$

Substituting these moments in the linear set of equations $\mathbb{E}[\mathcal{S}_4(Z)Z] = \mathbb{E}[\mathcal{S}_4(Z)(Z^2-1)] = \mathbb{E}[\mathcal{S}_4(Z)(Z^3-3Z)] = 0$, we obtain

$$\begin{bmatrix} 1.5+1.5\sigma^2 & \sqrt{\frac{2}{\pi}}+\sigma^2 & 0.5 \\ 3\sqrt{\frac{2}{\pi}}(1+\sigma^2/2) & 1+\sigma^2 & \frac{1}{2}\sqrt{\frac{2}{\pi}} \\ 3 & \sqrt{\frac{2}{\pi}}+\sigma^2 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = - \begin{bmatrix} \sqrt{\frac{2}{\pi}}(4+6\sigma^2)) \\ 6+6\sigma^2 \\ \sqrt{\frac{2}{\pi}}(12+6\sigma^2) \end{bmatrix}.$$

Solving for (α, β, γ) will yield $\mathcal{S}_4(Z)$. Finally, we have that $\delta = -2\sqrt{\frac{2}{\pi}}$.

C Proofs of Section 3.1

Remark 2. To choose the parameters in Theorem 1, we follow the parameter choices from Ge et al. (2018). Let c be a sufficiently small universal constant (e.g. $c = 0.01$). Assume $\mu \leq c/\kappa^*$, and $\lambda \geq 1/(ca^*_{\min})$. Let $\tau_0 = c\min\{\mu/(\kappa d a^*_{\max}), \lambda\}\sigma_{\min}(M)$. Let $\delta \leq \min\left\{\frac{c\varepsilon_0}{a^*_{\max}\cdot m\sqrt{d}\kappa^{1/2}(M)}, \tau_0/2\right\}$ and $\varepsilon = \min\left\{\lambda\sigma_{\min}(M)^{1/2}, c\delta/\sqrt{\|M\|}, c\varepsilon_0\delta\sigma_{\min}(M)\right\}$.

For any $k \times d$ matrix A, let $A^\dagger$ be its pseudo inverse such that $AA^\dagger = I_{k\times k}$ and $A^\dagger A$ is the projection matrix to the row span of A. Let $\alpha_i^* \triangleq \mathbb{E}[p_i^*(x)], a_i^* = \frac{1}{\alpha_i^*}$ and $\kappa^* = \frac{\alpha^*_{\max}}{\alpha^*_{\min}}$. Let $M = \sum_{i\in[k]} \alpha_i^* a_i^* (a_i^*)^\top$, $\kappa(M) = \frac{\|M\|}{\sigma_{\min}(M)}$.

For the sake of clarity, we now formally state our main assumptions, adapted from Makkuva et al. (2019):

1. x follows a standard Gaussian distribution, i.e. $x \sim \mathcal{N}(0, I_d)$.
2. $\|a_i^*\| = 1$ for all $i \in [k]$ and $\|w_i^*\| \leq R$ for all $i \in [k-1]$.
3. The regressors $a_1^*, \ldots, a_k^*$ are linearly independent and the classifiers $\{w_i^*\}_{i\in[k-1]}$ are orthogonal to the span $\mathcal{S} = \text{span}\{a_1^*, \ldots, a_k^*\}$, and $2k-1 < d$.
4. The non-linearity $g : \mathbb{R} \to \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta) -$ valid, which we define in Appendix B.

Note that while the first three assumptions are same as that of Makkuva et al. (2019), the fourth assumption is slightly different from theirs. Under this assumptions, we first give an alternative characterization of $L_4(\cdot)$ in the following theorem which would be crucial for the proof of Theorem 1.

Theorem 5. *The function $L(\cdot)$ defined in Eq. (6) satisfies that*

$$L_4(A) = \sum_{m\in[k]} \mathbb{E}[p_m^*(x)] \sum_{\substack{i\neq j \\ i,j\in[k]}} \langle a_m^*, a_i\rangle^2 \langle a_m^*, a_j\rangle^2 - \mu \sum_{m,i\in[k]} \mathbb{E}[p_m^*(x)]\langle a_m^*, a_i\rangle^4 + \lambda \sum_{i\in[k]} \Big(\sum_{m\in[k]} \mathbb{E}[p_m^*(x)]\langle a_m^*, a_i\rangle^2 - 1\Big)^2 + \frac{\delta}{2}\|A\|_F^2$$

C.1 Proof of Theorem 5

Proof. For the proof of Theorem 5, we use the notion of score functions defined as Janzamin et al. (2014):

$$\mathcal{S}_m(x) \triangleq (-1)^m \frac{\nabla_x^{(m)} f(x)}{f(x)}, \quad f \text{ is the pdf of } x. \tag{10}$$

In this paper we focus on $m = 2, 4$. When $x \sim \mathcal{N}(0, I_d)$, we know that $\mathcal{S}_2(x) = x \otimes x - I$ and

$$\mathcal{S}_4(x) = x^{\otimes 4} - \sum_{i\in[d]} \text{sym}\left(x \otimes e_i \otimes e_i \otimes x\right) + \sum_{i,j} \text{sym}\left(e_i \otimes e_i \otimes e_j \otimes e_j\right).$$

The score transformations $\mathcal{S}_4(x)$ and $\mathcal{S}_2(x)$ can be viewed as multi-variate polynomials in x of degrees 4 and 2 respectively. For the output y, recall the transforms $\mathcal{Q}_4(y)$ and $\mathcal{Q}_2(y)$ defined in Section 3.1. The following lemma shows that one can construct a fourth-order super symmetric tensor using these special transforms.

Lemma 1 (Super symmetric tensor construction). *Let (x, y) be generated according to Eq. (1) and Assumptions (1)-(4) hold. Then*

$$\mathcal{T}_4 \triangleq \mathbb{E}[\mathcal{Q}_4(y) \cdot \mathcal{S}_4(x)] = c_{g,\sigma} \sum_{i\in[k]} \mathbb{E}[p_i^*(x)] \cdot a_i^* \otimes a_i^* \otimes a_i^* \otimes a_i^*,$$
$$\mathcal{T}_2 \triangleq \mathbb{E}[\mathcal{Q}_2(y) \cdot \mathcal{S}_2(x)] = c'_{g,\sigma} \sum_{i\in[k]} \mathbb{E}[p_i^*(x)] \cdot a_i^* \otimes a_i^*,$$

where $p_i^*(x) = \mathbb{P}[z_i = 1|x]$, $c_{g,\sigma}$ and $c'_{g,\sigma}$ are two non-zero constants depending on g and σ.

Now the proof of the theorem immediately follows from Lemma 1. Recall from Eq. (6) that

$$L_4(A) \triangleq \sum_{\substack{i,j\in[k]\\ i\neq j}} \mathbb{E}[\mathcal{Q}_4(y)t_1(a_i,a_j,x)] - \mu \sum_{i\in[k]} \mathbb{E}[\mathcal{Q}_4(y)t_2(a_i,x)] + \lambda \sum_{i\in[k]} \left(\mathbb{E}[\mathcal{Q}_2(y)t_3(a_i,x)] - 1\right)^2 + \frac{\delta}{2}\|A\|_F^2.$$

Fix $i, j \in [k]$. Notice that we have $t_1(a_i, a_j, x) = \mathcal{S}_4(x)(a_i, a_i, a_j, a_j)/c_{g,\sigma}$. Hence we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{Q}_4(y)t_1(a_i,a_j,x)] &= \frac{1}{c_{g,\sigma}}\mathbb{E}[\mathcal{Q}_4(y)\cdot\mathcal{S}_4(x)](a_i,a_i,a_j,a_j)\\ &= \left(\sum_{m\in[k]} \mathbb{E}[p_m^*(x)](a_m^*)^{\otimes 4}\right)(a_i,a_i,a_j,a_j)\\ &= \sum_{m\in[k]} \mathbb{E}[p_m^*(x)]\langle a_m^*, a_i\rangle^2\langle a_m^*, a_j\rangle^2.\end{aligned}$$

The simplification for the remaining terms is similar and follows directly from definitions of $t_2(\cdot, x)$ and $t_3(\cdot, x)$. □

C.2 Proof of Theorem 1

Proof. The proof is an immediate consequence of Theorem 5 and Theorem C.5 of Ge et al. (2018). □

C.3 Proof of Theorem 2

Proof. Note that our loss function $L_4(A)$ can be written as $\mathbb{E}[\ell(x, y, A)]$ where ℓ is at most a fourth degree polynomial in x, y and A. Hence our finite sample guarantees directly follow from Theorem 1 and Theorem E.1 of Ge et al. (2018). □

C.4 Proof of Lemma 1

Proof. The proof of this lemma essentially follows the same arguments as that of (Makkuva et al., 2019, Theorem 1), where we replace $(\mathcal{S}_3(x), \mathcal{S}_2(x), \mathcal{P}_3(y), \mathcal{P}_2(y))$ with $(\mathcal{S}_4(x), \mathcal{S}_2(x), \mathcal{Q}_4(y), \mathcal{P}_2(y))$ respectively and letting $\mathcal{T}_3$ defined there with our $\mathcal{T}_4$ defined above.

□

D Proofs of Section 3.2

For the convergence analysis of SGD on $L_{\log}$, we use techniques from Balakrishnan et al. (2017) and Makkuva et al. (2019). In particular, we adapt (Makkuva et al., 2019, Lemma 3) and (Makkuva et al., 2019, Lemma 4) to our setting through Lemma 2 and Lemma 3, which are central to the proof of Theorem 3 and Theorem 4. We now sate our lemmas.

Lemma 2. *Under the assumptions of Theorem 3, it holds that*

$$\|G(W, A^*) - W_i^*\| \leq \rho_\sigma \|W - W^*\|.$$

In addition, $W = W^*$ is a fixed point for $G(W, A^*)$.

Lemma 3. *Let the matrix of regressors A be such that $\max_{i\in[k]} \|A_i^\top - (A_i^*)^\top\|_2 = \sigma^2\varepsilon$. Then for any $W \in \Omega$, we have that*

$$\|G(W, A) - G(W, A^*)\| \le \kappa\varepsilon,$$

where κ is a constant depending on g, k and σ. In particular, $\kappa \le (k-1)\frac{\sqrt{6(2+\sigma^2)}}{2}$ for g =linear, sigmoid and ReLU.

Lemma 4 (Deviation of finite sample gradient operator)**.** *For some universal constant c_1, let the number of samples n be such that $n \ge c_1 d\log(1/\delta)$. Then for any fixed set of regressors $A \in \mathbb{R}^{k\times d}$, and a fixed $W \in \Omega$, the bound*

$$\|G_n(W, A) - G(W, A)\| \le \varepsilon_G(n, \delta) \triangleq c_2\sqrt{\frac{d\log(k/\delta)}{n}}$$

holds with probability at least $1 - \delta$.

D.1 Proof of Theorem 3

Proof. The proof directly follows from Lemma 2 and Lemma 3. □

D.2 Proof of Theorem 4

Proof. Let the set of regressors A be such that $\max_{i\in[k]} \|A_i^\top - (A_i^*)^\top\|_2 = \sigma^2\varepsilon_1$. Fix A. For any iteration $t \in [T]$, from Lemma 4 we have the bound

$$\|G_{n/T}(W_t, A) - G(W_t, A)\| \le \varepsilon_G(n/T, \delta/T) \tag{11}$$

with probability at least $1 - \delta/T$. Using an union bound argument, Eq. (11) holds with probability at least $1 - \delta$ for all $t \in [T]$. Now we show that the following bound holds:

$$\|W_{t+1} - W^*\| \le \rho_\sigma \|W_t - W^*\| + \kappa\varepsilon_1 + \varepsilon_G(n/T, \delta/T), \qquad \text{for each } t \in \{0, \dots, T-1\}. \tag{12}$$

Indeed, for any $t \in \{0, \dots, T-1\}$, we have that

$$\begin{aligned}
\|W_{t+1} - W^*\| &= \left\|G_{n/T}(W_t, A) - W^*\right\| \\
&\le \left\|G_{n/T}(W_t, A) - G(W_t, A)\right\| + \|G(W_t, A) - G(W_t, A^*)\| + \|G(W_t, A^*) - W^*\| \\
&\le \varepsilon_G(n/T, \delta/T) + \kappa\varepsilon_1 + \rho_\sigma \|W_t - W^*\|,
\end{aligned}$$

where we used in Lemma 2, Lemma 3 and Lemma 4 in the last inequality to bound each of the terms. From Eq. (11), we obtain that

$$\begin{aligned}
\|W_t - W^*\| &\le \rho_\sigma \|W_{t-1} - W^*\| + \kappa\varepsilon_1 + \varepsilon_G(n/T, \delta/T) \\
&\le \rho_\sigma^2 \|W_{t-2} - W^*\| + (1 + \rho_\sigma)(\kappa\varepsilon_1 + \varepsilon_G(n/T, \delta/T)) \\
&\le \rho_\sigma^t \|W_0 - W^*\| + \left(\sum_{s=0}^{t-1} \rho_\sigma^s\right)(\kappa\varepsilon_1 + \varepsilon_G(n/T, \delta/T)) \\
&\le \rho_\sigma^t \|W_0 - W^*\| + \left(\frac{1}{1-\rho_\sigma}\right)(\kappa\varepsilon_1 + \varepsilon_G(n/T, \delta/T)).
\end{aligned}$$

□

D.3 Proof of Lemma 2

Proof. Recall that the loss function for the population setting, $L_{\log}(W, A)$, is given by

$$L_{\log}(W, A) = -\mathbb{E}\log\left(\sum_{i\in[k]} \frac{e^{\langle w_i, x\rangle}}{\sum_{j\in[k]} e^{\langle w_j, x\rangle}} \cdot \mathcal{N}(y|g(\langle a_i, x\rangle), \sigma^2)\right) = -\mathbb{E}\log\left(\sum_{i\in[k]} p_i(x)N_i\right),$$

where $p_i(x) \triangleq \frac{e^{\langle w_i, x\rangle}}{\sum_{j\in[k]} e^{\langle w_j, x\rangle}}$ and $N_i \triangleq \mathcal{N}(y|g(\langle a_i, x\rangle), \sigma^2)$. Hence for any $i \in [k-1]$, we have

$$\nabla_{w_i} L_{\log}(W, A) = -\mathbb{E}\left(\frac{\nabla_{w_i} p_i(x) N_i + \sum_{j\neq i, j\in[k]} \nabla_{w_i} p_j(x) N_j}{\sum_{i\in[k]} p_i(x) N_i}\right).$$

Moreover,

$$\nabla_{w_i} p_j(x) = \begin{cases} p_i(x)(1-p_i(x))x, & j = i \\ -p_i(x)p_j(x)x, & j \neq i \end{cases}.$$

Hence we obtain that

$$\nabla_{w_i} L_{\log}(W, A) = -\mathbb{E}\left[\frac{p_i(x)N_i}{\sum_{i\in[k]} p_i(x)N_i} - p_i(x)\right]. \tag{13}$$

Notice that if $z \in [k]$ denotes the latent variable corresponding to which expert is chosen, we have that the posterior probability of choosing the ith expert is given by

$$\mathbb{P}\left[z = i|x, y\right] = \frac{p_i(x)N_i}{\sum_{i\in[k]} p_i(x)N_i},$$

whereas,

$$\mathbb{P}\left[z = i|x\right] = p_i(x).$$

Hence, when $A = A^*$ and $W = W^*$, we get that

$$\nabla_{w_i^*} L_{\log}(W^*, A^*) = -\mathbb{E}[\mathbb{P}\left[z = i|x, y\right] - \mathbb{P}\left[z = i|x\right]] = -\mathbb{E}[\mathbb{P}\left[z = i|x\right] + \mathbb{E}[\mathbb{P}\left[z = i|x\right]] = 0.$$

Thus $W = W^*$ is a fixed point for $G(W, A^*)$ since

$$G(W^*, A^*) = \Pi_\Omega(W^* - \alpha\nabla_{W^*} L_{\log}(W^*, A^*)) = W^*.$$

Now we make the observation that the population-gradient updates $W_{t+1} = G(W_t, A)$ are same as the gradient-EM updates. Thus the contraction of the population-gradient operator $G(\cdot, A^*)$ follows from the contraction property of the gradient EM algorithm (Makkuva et al., 2019, Lemma 3). To see this, recall that for k-MoE, the gradient-EM algorithm involves computing the function $Q(W|W_t)$ for the current iterate W_t and defined as:

$$Q(W|W_t) = \mathbb{E}\left[\sum_{i\in[k-1]} p_{W_t}^{(i)}(w_i^\top x) - \log\left(1 + \sum_{i\in[k-1]} e^{w_i^\top x}\right)\right],$$

where $p_{W_t}^{(i)} = \mathbb{P}\left[z = i|x, y, w_t\right]$ corresponds to the posterior probability for the i^{th} expert, given by

$$p_{W_t}^{(i)} = \frac{p_{i,t}(x)\mathcal{N}(y|g(a_i^\top x), \sigma^2)}{\sum_{j\in[k]} p_{j,t}(x)\mathcal{N}(y|g(a_j^\top x), \sigma^2)}, \quad p_{i,t}(x) = \frac{e^{(w_t)_i^\top x}}{1 + \sum_{j\in[k-1]} e^{(w_t)_j^\top x}}.$$

Then the next iterate of the gradient-EM algorithm is given by $W_{t+1} = \Pi_\Omega(W_t + \alpha\nabla_W Q(W|W_t)_{W=W_t})$. We have that

$$\nabla_{w_i} Q(W|W_t)|_{W=W_t} = \mathbb{E}\left[\left(p_{W_t}^{(i)} - \frac{e^{(w_t)_i^\top x}}{1 + \sum_{j\in[k-1]} e^{(w_t)_j^\top x}}\right)x\right] = -\nabla_{w_i} L_{\log}(W_t, A).$$

Hence if we use the same step size α, our population-gradient iterates on the log-likelihood are same as that of the gradient-EM iterates. This finishes the proof. □

D.4 Proof of Lemma 3

Proof. Fix any $W \in \Omega$ and let $A = \begin{bmatrix} a_1^\top \\ \dots \\ a_k^\top \end{bmatrix} \in \mathbb{R}^{k\times d}$ be such that $\max_{i\in[k]} \|a_i - a_i^*\|_2 = \sigma^2\varepsilon_1$ for some $\varepsilon_1 > 0$. Let

$$W' = G(W, A), \quad (W')^* = G(W, A^*).$$

Denoting the i^{th} row of $W' \in \mathbb{R}^{(k-1)\times d}$ by w_i' and that of $(W')^*$ by $(w_i')^*$ for any $i \in [k-1]$, we have that

$$\begin{aligned}\|w_i' - (w_i')^*\|_2 &= \|\Pi_\Omega(w_i - \alpha\nabla_{w_i} L_{\log}(W, A)) - \Pi_\Omega(w_i - \alpha\nabla_{w_i} L_{\log}(W, A^*))\|_2 \\ &\le \alpha \|\nabla_{w_i} L_{\log}(W, A) - \nabla_{w_i} L_{\log}(W, A^*)\|_2 .\end{aligned}$$

Thus it suffices to bound $\|\nabla_{w_i} L_{\log}(W, A) - \nabla_{w_i} L_{\log}(W, A^*)\|_2$. From Eq. (13), we have that

$$\begin{aligned}\nabla_{w_i} L_{\log}(W, A) &= -\mathbb{E}\left[\left(\frac{p_i(x)N_i}{\sum_{i\in[k]} p_i(x)N_i} - p_i(x)\right)x\right], \\ \nabla_{w_i} L_{\log}(W, A^*) &= -\mathbb{E}\left[\left(\frac{p_i(x)N_i^*}{\sum_{i\in[k]} p_i(x)N_i^*} - p_i(x)\right)x\right],\end{aligned}$$

where,

$$p_i(x) = \frac{e^{w_i^\top x}}{1 + \sum_{k\in[k-1]} e^{w_j^\top x}}, \quad N_i \triangleq \mathcal{N}(y|g(a_i^\top x), \sigma^2), \quad N_i^* = \mathcal{N}(y|g((a_i^*)^\top x), \sigma^2).$$

Thus we have

$$\|\nabla_{w_i} L_{\log}(W, A) - \nabla_{w_i} L_{\log}(W, A^*)\|_2 = \left\|\mathbb{E}[(p^{(i)}(A, W) - p^{(i)}(A^*, W))x]\right\|_2, \tag{14}$$

where $p^{(i)}(A, W) \triangleq \frac{p_i(x)N_i}{\sum_{i\in[k]} p_i(x)N_i}$ denotes the posterior probability of choosing the i^{th} expert. Now we observe that Eq. (14) reduces to the setting of (Makkuva et al., 2019, Lemma 4) and hence the conclusion follows.

□

D.5 Proof of Lemma 4

Proof. We first prove the lemma for $k = 2$. For 2-MoE, we have that the posterior probability is given by

$$p_w(x, y) = \frac{f(w^\top x)N_1}{f(w^\top x)N_1 + (1 - f(w^\top x))N_2},$$

where $f(\cdot) = \frac{1}{1+e^{-(\cdot)}}$, $N_1 = \mathcal{N}(y|g(a_1^\top x), \sigma^2)$ and $N_2 = \mathcal{N}(y|g(a_2^\top x), \sigma^2)$ for fixed $a_1, a_2 \in \mathbb{R}^d$. Then we have that

$$\nabla_w L_{\log}(w, A) = -\mathbb{E}[(p_w(x, y) - f(w^\top x)) \cdot x].$$

Hence

$$G(w, A) = \Pi_\Omega(w + \alpha\mathbb{E}[(p_w(x, y) - f(w^\top x)) \cdot x]), \quad G_n(w, A) = \Pi_\Omega(w + \frac{\alpha}{n}\sum_{i\in[n]} (p_w(x_i, y_i) - f(w^\top x_i)) \cdot x_i).$$

Since $0 < \alpha < 1$, we have that

$$\begin{aligned}\|G(w, A) - G_n(w, A)\|_2 &\le \|\mathbb{E}[(p_w(x, y) - f(w^\top x))x] - \frac{1}{n}\sum_{i\in[n]} (p_w(x_i, y_i) - f(w^\top x_i))x_i\|_2 \\ &\le \underbrace{\|\mathbb{E}[p_w(x, y)x] - \sum_{i\in[n]} \frac{p_w(x_i, y_i)x_i}{n}\|_2}_{T_1} + \underbrace{\|\mathbb{E}[f(w^\top x)x] - \sum_{i\in[n]} \frac{f(w^\top x_i)x_i}{n}\|_2}_{T_2}.\end{aligned}$$

We now bound T_1 and T_2.

Bounding T_2: We prove that the random variable $\sum_{i\in[n]} \frac{f(w^\top x_i)x_i}{n} - \mathbb{E}[f(w^\top x)x]$ is sub-gaussian with parameter $L/\sqrt{n}$ for some constant $L > 1$ and thus its squared norm is sub-exponential. We then bound T_2 using standard sub-exponential concentration bounds. Towards the same, we first show that the random variable $f(w^\top x)x - \mathbb{E}[f(w^\top x)x]$ is sub-gaussian with parameter L. Or equivalently, that $f(w^\top x)\langle x, u\rangle - \mathbb{E}[f(w^\top x)\langle x, u\rangle]$ is sub-gaussian for all $u \in \mathbb{S}^d$.

Without loss of generality, assume that $w \neq 0$. First let $u = \vec{w} \triangleq \frac{w}{\|w\|}$. Thus $Z \triangleq \langle \vec{w}, x\rangle \sim \mathcal{N}(0,1)$. We have

$$g(Z) \triangleq f(w^\top x)\langle x, \vec{w}\rangle - \mathbb{E}[f(w^\top x)\langle x, \vec{w}\rangle] = f(\|w\| Z)Z - \mathbb{E}[f(\|w\| Z)Z].$$

It follows that $g(\cdot)$ is Lipschitz since

$$|g'(z)| = |f'(\|w\| z)\|w\| z + f(\|w\| z))| \leq \sup_{t\in\mathbb{R}} |f'(t)t| + 1 = \sup_{t>0} \frac{te^t}{(1+e^t)^2} + 1 \triangleq L.$$

From the Talagaran concentration of Gaussian measure for Lipschitz functions (Ledoux and Talagrand, 1991), it follows that $g(Z)$ is sub-gaussian with parameter L. Now consider any $u \in \mathbb{S}^d$ such that $u \perp w$. Then we have that $Y \triangleq \langle u, x\rangle \sim \mathcal{N}(0,1)$ and $Z \triangleq \langle \vec{w}, x\rangle \sim \mathcal{N}(0,1)$ are independent. Thus,

$$g(Y, Z) \triangleq f(w^\top x)\langle u, x\rangle - \mathbb{E}[f(w^\top x)\langle u, x\rangle] = f(\|w\| Z)Y - \mathbb{E}[f(\|w\| Z)Y]$$

is sub-gaussian with parameter 1 since $f \in [0,1]$ and Y, Z are independent standard gaussians. Since any $u \in \mathbb{S}^d$ can be written as

$$u = P_w(u) + P_{w^\perp}(u),$$

where P_S denotes the projection operator onto the sub-space S, we have that $f(w^\top x)\langle x, u\rangle - \mathbb{E}[f(w^\top x)\langle x, u\rangle]$ is sub-gaussian with parameter L for all $u \in \mathbb{S}^d$. Thus it follows that $\sum_{i\in[n]} \frac{f(w^\top x_i)x_i}{n} - \mathbb{E}[f(w^\top x)x]$ is zero-mean and sub-gaussian with parameter $L/\sqrt{n}$ which further implies that

$$T_2 \leq c_2 L\sqrt{\frac{d\log(1/\delta)}{n}},$$

with probability at least $1 - \delta/2$.

Bounding T_1: Let $Z \triangleq \|\sum_{i\in[n]} \frac{p_w(x_i,y_i)x_i}{n} - \mathbb{E}[p_w(x,y)x]\|_2 = \sup_{u\in\mathbb{S}^d} Z(u)$, where

$$Z(u) \triangleq \sum_{i\in[n]} \frac{p_w(x_i, y_i)\langle x_i, u\rangle}{n} - \mathbb{E}[p_w(x,y)\langle x, u\rangle].$$

Let $\{u_1, \ldots, u_M\}$ be a 1/2-cover of the unit sphere $\mathbb{S}^d$. Hence for any $v \in \mathbb{S}^d$, there exists a $j \in [M]$ such that $\|v - u_j\|_2 \leq 1/2$. Thus,

$$Z(v) \leq Z(u_j) + |Z(v) - Z(u_j)| \leq Z\|v - u_j\|_2 \leq Z(u_j) + Z/2,$$

where we used the fact that $|Z(u) - Z(v)| \leq Z\|u - v\|_2$ for any $u, v \in \mathbb{S}^d$. Now taking supremum over all $v \in \mathbb{S}^d$ yields that $Z \leq 2\max_{j\in[M]} Z(u_j)$. Now we bound $Z(u)$ for a fixed $u \in \mathbb{S}^d$. By symmetrization trick (Vaart and Wellner, 1996), we have

$$\mathbb{P}[Z(u) \geq t] \leq 2\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i p_w(x_i, y_i)\langle x_i, u\rangle \geq t/2\right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher variables. Define the event $E \triangleq \{\frac{1}{n}\sum_{i\in[n]}\langle x_i, u\rangle^2 \leq 2\}$. Since $\langle x_i, u\rangle \sim \mathcal{N}(0,1)$, standard tail bounds imply that $\mathbb{P}[E^c] \leq e^{-n/32}$. Thus we have that

$$\mathbb{P}[Z(u) \geq t] \leq 2\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i p_w(x_i, y_i)\langle x_i, u\rangle \geq t/2 | E\right] + 2e^{-n/32}.$$

Considering the first term, for any $\lambda > 0$, we have

$$\mathbb{E}[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\varepsilon_i p_w(x_i,y_i)\langle x_i,u\rangle\right)|E] \leq \mathbb{E}[\exp\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i\langle x_i,u\rangle\right)|E],$$

where we used the Ledoux-Talagrand contraction for Rademacher process (Ledoux and Talagrand, 1991), since $|p_w(x_i,y_i)| \leq 1$ for all (x_i,y_i). The sub-gaussianity of Rademacher sequence $\{\varepsilon_i\}$ implies that

$$\mathbb{E}[\exp\left(\frac{2\lambda}{n}\sum_{i=1}^{n}\varepsilon_i\langle x_i,u\rangle\right)|E] \leq \mathbb{E}[\exp\left(\frac{2\lambda^2}{n^2}\sum_{i=1}^{n}\langle x_i,u\rangle^2\right)|E] \leq \exp(\frac{4\lambda^2}{n}),$$

using the definition of the event E. Thus the above bound on the moment generating function implies the following tail bound:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i p_w(x_i,y_i)\langle x_i,u\rangle \geq t/2|E\right] \leq \exp\left(-\frac{nt^2}{256}\right).$$

Combining all the bounds together, we obtain that

$$\mathbb{P}\left[Z(u) \geq t\right] \leq 2e^{-nt^2/256} + 2e^{-n/32}.$$

Since $M \leq 2^d$, using the union bound we obtain that

$$\mathbb{P}\left[Z \geq t\right] \leq 2^d(2e^{-nt^2/1024} + 2e^{-n/32}).$$

Since $n \geq c_1 d\log(1/\delta)$, we have that $T_1 = Z \leq c\sqrt{\frac{d\log(1/\delta)}{n}}$ with probability at least $1-\delta/2$. Combining these bounds on T_1 and T_2 yields the final bound on $\varepsilon_G(n,\delta)$.

Now consider any $k \geq 2$. From Eq. (13), defining $N_i \triangleq \mathcal{N}(y|g(a_i^\top x),\sigma^2)$ and $p_i(x) = \frac{e^{w_i^\top x}}{1+\sum_{j\in[k-1]}e^{w_j^\top x}}$, we have that

$$\nabla_{w_i} L_{\log}(W,A) = -\mathbb{E}\left(\frac{p_i(x)N_i}{\sum_{i\in[k]}p_i(x)N_i} - p_i(x)\right)x.$$

Similarly,

$$\nabla_{w_i} L_{\log}^{(n)}(W,A) = -\sum_{j=1}^{n}\frac{1}{n}\left(\frac{p_i(x_j)N_i}{\sum_{i\in[k]}p_i(x_j)N_i} - p_i(x_j)\right)x_j.$$

Since $\|G_n(W,A) - G(W,A)\| = \max_{i\in[k-1]}\|G_n(W,A)_i - G(W,A)_i\|_2$, with out loss of generality, we let $i=1$. The proof for the other cases is similar. Thus we have

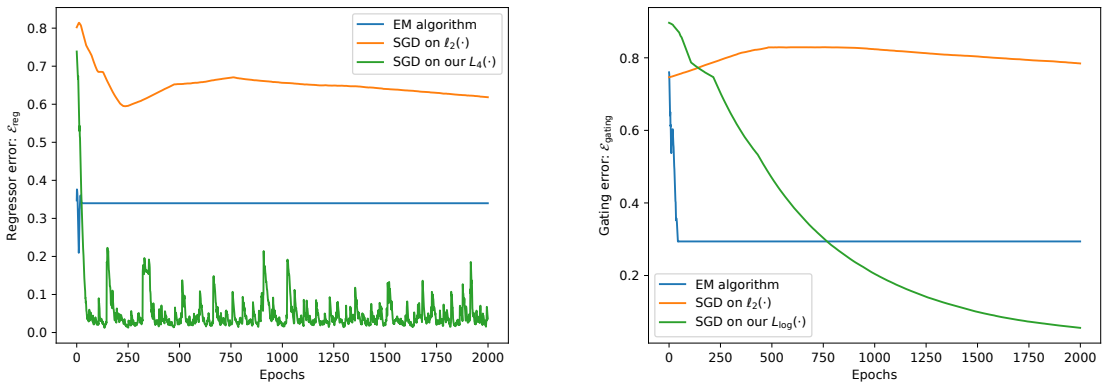
$$\begin{aligned}\|G_n(W,A)_1 - G(W,A)_1\|_2 &\leq \left\|\nabla_{w_1}L_{\log}(W,A) - \nabla_{w_1}L_{\log}^{(n)}(W,A)\right\|_2 \\ &\leq \left\|\sum_{i=1}^{n}\frac{p^{(1)}(x_i,y_i)x_i}{n} - \mathbb{E}[p^{(1)}(x,y)x]\right\|_2 + \left\|\sum_{i=1}^{n}\frac{p_1(x)x}{n} - \mathbb{E}[p_1(x)x]\right\|_2,\end{aligned}$$

where $p^{(1)}(x,y) \triangleq \frac{p_1(x)N_1}{\sum_{i\in[k]}p_i(x)N_i}$. Since $|p^{(1}(x,y)| \leq 1$ and $|p_1(x)| \leq 1$, we can use the same argument as in the bounding of T_1 proof for 2-MoE above to get the parametric bound. This finishes the proof. □

E Additional experiments

E.1 Reduced batch size

In Figure 4 we ran SGD on our loss $L_4(\cdot)$ with 5 different runs with a batch size of 128 and a learning rate of 0.001 for $d = 10$ and $k = 3$. We can see that our algorithm still converges to zero but with a more variance because of noisy gradient estimation and also lesser number of samples than the required sample complexity.



(a) Regressor error

(b) Gating error

Figure 4: Comparison of SGD on our losses $(L_4, L_{\log})$ vs. ℓ_2 and the EM algorithm.