Organization. The appendix is organized as follows:

- Appendix A and Appendix B contain the requisite material for method of moments and the convergence analysis of EM respectively.
- Appendix C details the class of non-linearities for which our results hold.
- Appendix D contains all the proofs of Section 3. Two technical lemmas needed to prove Theorem 2 are relegated to Appendix E and Appendix F.
- Appendix G provides convergence guarantees for Gradient EM.
- Appendix H contains additional experiments for the comparison of joint-EM and our algorithm for the synthetic data.

A. Toolbox for method of moments

In this section, we introduce the key techniques that are useful in parameter estimation of mixture models via the method of moments.

Stein's identity (Stein's lemma) is a well-known result in probability and statistics and is widely used in estimation and inference taks. A refined version of the Stein's lemma (Stein, 1972) for higher-order moments is the key to parameter estimation in mixture of generalized linear models. We utilize this machinery in proving Theorem 1. We first recall the Stein's lemma.

Lemma 1 (Stein's lemma (Stein, 1972))**.** *Let $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$ and $g : \mathbb{R}^d \to \mathbb{R}$ be a function such that both $\mathbb{E}[\nabla_{\boldsymbol{x}} g(\boldsymbol{x})]$ and $\mathbb{E}[g(\boldsymbol{x}) \cdot \boldsymbol{x}]$ exist and are finite. Then*

$$\mathbb{E}[g(\boldsymbol{x}) \cdot \boldsymbol{x}] = \mathbb{E}[\nabla_{\boldsymbol{x}} g(\boldsymbol{x})].$$

The following lemma, which can be viewed as an extension of Stein's lemma for higher-order moments, is the central technique behind parameter estimation in M-GLMs.

Lemma 2 ((Sedghi et al., 2014))**.** *Let $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$ and $\mathcal{S}_3(\boldsymbol{x})$ be as defined in (6) and let $\mathcal{S}_2(\boldsymbol{x}) \triangleq \boldsymbol{x} \otimes \boldsymbol{x} - I_d$. Then for any $g : \mathbb{R}^d \to \mathbb{R}$ satisfying some regularity conditions, we have*

$$\mathbb{E}[g(\boldsymbol{x}) \cdot \mathcal{S}_2(\boldsymbol{x})] = \mathbb{E}[\nabla^{(2)}_{\boldsymbol{x}} g(\boldsymbol{x})], \quad \mathbb{E}[g(\boldsymbol{x}) \cdot \mathcal{S}_3(\boldsymbol{x})] = \mathbb{E}[\nabla^{(3)}_{\boldsymbol{x}} g(\boldsymbol{x})].$$

B. Toolbox for EM convergence analysis

Recall that the domain of our gating parameters is $\Omega = \{\boldsymbol{w} : \|\boldsymbol{w}\| \leq 1\}$. Then the population EM for the mixture of experts consists of the following two steps:

- **E-step:** Using the current estimate $\boldsymbol{w}_t$ to compute the function $Q(\cdot|\boldsymbol{w}_t)$.
- **M-step:** $\boldsymbol{w}_{t+1} = \mathrm{argmax}_{\|\boldsymbol{w}\| \leq 1} Q(\boldsymbol{w}|\boldsymbol{w}_t)$.

Thus the EM can be viewed as a deterministic procedure which maps $\boldsymbol{w}_t \mapsto M(\boldsymbol{w}_t)$ where

$$M(\boldsymbol{w}) = \mathrm{argmax}_{\boldsymbol{w}' \in \Omega} Q(\boldsymbol{w}'|\boldsymbol{w}).$$

Our convergence analysis relies on tools from (Balakrishnan et al., 2017) where they provided local convergence results on both the EM and gradient EM algorithms. In particular, they showed that if we initialize EM in a sufficiently small neighborhood around the true parameters, the EM iterates converge geometrically to the true parameters under some strong-concavity and gradient stability conditions. We now formally state the assumptions in (Balakrishnan et al., 2017) under which the convergence guarantees hold. We will show in the next section that these conditions hold *globally* in our setting.

Assumption 1 (Convexity of the domain)**.** Ω is convex.

Assumption 2 (Strong-concavity)**.** $Q(\cdot|\boldsymbol{w}^*)$ is a λ-strongly concave function over a r-neighborhood of $\boldsymbol{w}^*$, i.e. $\mathcal{B}(\boldsymbol{w}^*, r) \triangleq \{\boldsymbol{w} \in \Omega : \|\boldsymbol{w} - \boldsymbol{w}^*\| \leq r\}$.

Remark 1. An important point to note is that the true parameter $\boldsymbol{w}^*$ is a fixed point for the EM algorithm, i.e. $M(\boldsymbol{w}^*) = \boldsymbol{w}^*$. This is also known as *self-consistency* of the EM algorithm. Hence it is reasonable to expect that in a sufficiently small neighborhood around $\boldsymbol{w}^*$ there exists a unique maximizer for $Q(\cdot|\boldsymbol{w}^*)$.

Assumption 3 (First-order stability condition)**.** Assume that

$$\|\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}^*) - \nabla Q(M(\boldsymbol{w})|\boldsymbol{w})\| \leq \gamma \|\boldsymbol{w} - \boldsymbol{w}^*\|, \quad \forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}^*, r).$$

Remark 2. Intuitively, the gradient stability condition enforces the gradient maps $\nabla Q(\cdot|\boldsymbol{w})$ and $\nabla Q(\cdot|\boldsymbol{w}^*)$ to be close whenever $\boldsymbol{w}$ lies in a neighborhood of $\boldsymbol{w}^*$. This will ensure that the mapped output $M(\boldsymbol{w})$ stays closer to $\boldsymbol{w}^*$.

Theorem 4 (Theorem 1, (Balakrishnan et al., 2017))**.** *If the above assumptions are met for some radius $r > 0$ and $0 \leq \gamma < \lambda$, then the map $\boldsymbol{w} \mapsto M(\boldsymbol{w})$ is contractive over $\mathcal{B}(\boldsymbol{w}^*, r)$, i.e.*

$$\|M(\boldsymbol{w}) - \boldsymbol{w}^*\| \leq \left(\frac{\gamma}{\lambda}\right) \|\boldsymbol{w} - \boldsymbol{w}^*\|, \quad \forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}^*, r),$$

and consequently, the EM iterates $\{\boldsymbol{w}_t\}_{t \geq 0}$ converge geometrically to $\boldsymbol{w}^*$, i.e.

$$\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq \left(\frac{\gamma}{\lambda}\right)^t \|\boldsymbol{w}_0 - \boldsymbol{w}^*\|,$$

whenever the initialization $\boldsymbol{w}_0 \in \mathcal{B}(\boldsymbol{w}^*, r)$.

C. Class of non-linearities

In this section, we characterize the class of non-linearities for which our theoretical results for the recovery of regressors hold. 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 $(\alpha, \beta, \gamma) \in \mathbb{R}^3$, define

$$\mathcal{P}_3(y) \triangleq Y^3 + \alpha Y^2 + \beta Y, \quad \mathcal{S}_3(Z) = \mathbb{E}[\mathcal{P}_3(y)|Z] = g(Z)^3 + \alpha g(Z)^2 + g(Z)(\beta + 3\sigma^2) + \alpha\sigma^2,$$

and

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

Condition 1. $\mathbb{E}[\mathcal{S}_3'(Z)] = \mathbb{E}[\mathcal{S}_3''(Z)] = 0$ and $\mathbb{E}[\mathcal{S}_3'''(Z)] \neq 0$.

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

We are now ready to define the (α, β, γ)-valid class of non-linearities.

Definition 1. *We say that the non-linearity g is (α, β, γ)-valid if there exists $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ such that both Condition 1 and Condition 2 are satisfied.*

We have that

$$\begin{aligned}
\mathcal{S}_3'(Z) &= 3g(Z)^2 g'(Z) + 2\alpha g(Z) g'(Z) + g'(Z)(\beta + 3\sigma^2) \\
&= 2\alpha g(Z) g'(Z) + \beta g'(Z) + 3g(Z)^2 g'(Z) + 3g'(Z)\sigma^2, \\
\mathcal{S}_3''(Z) &= 2\alpha \left(g'(Z)^2 + g(Z) g''(Z)\right) + \beta g''(Z) + 3g''(Z)(g(Z)^2 + \sigma^2) + 6g(Z) g'(Z)^2.
\end{aligned}$$

Thus $\mathbb{E}[\mathcal{S}_3'(Z)] = \mathbb{E}[\mathcal{S}_3''(Z)] = 0$ implies that

$$\begin{bmatrix} 2\mathbb{E}(g(Z)g'(Z)) & \mathbb{E}(g'(Z)) \\ 2\mathbb{E}\left(g'(Z)^2 + g(Z)g''(Z)\right) & \mathbb{E}(g''(Z)) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} -3\mathbb{E}(g(Z)^2 g'(Z) + g'(Z)\sigma^2) \\ -3\mathbb{E}(g''(Z)(g(Z)^2 + \sigma^2) + 2g(Z)g'(Z)^2) \end{bmatrix}$$

To ensure Condition 1, we need the pair (α, β) obtained by solving the above linear equation to satisfy $\mathbb{E}[\mathcal{S}_3'''(Z)] \neq 0$. Similarly, $\mathbb{E}[\mathcal{S}_2'(Z)] = 0$ implies that

$$\gamma = \frac{-2\mathbb{E}[g(Z)g'(Z)]}{\mathbb{E}[g'(Z)]}.$$

Thus Condition 2 stipulates that $\mathbb{E}[\mathcal{S}_2''(Z)] \neq 0$ with this choice of γ. It turns out that these conditions hold for a wide class of non-linearities and in particular, when g is either the identity function, or the sigmoid function, or the ReLU. For these three choices of popular non-linearities, the values of the tuple (α, β, γ) are provided below (which are obtained by solving the linear equations mentioned above).

Example 1. If g is the identity mapping, then $\mathcal{P}_3(y) = y^3 - 3y(1+\sigma^2)$ and $\mathcal{S}_2(y) = y^2$.

Example 2. If g is the sigmoid function, i.e. $g(z) = \frac{1}{1+e^{-z}}$, then α and β can be obtained by solving the following linear equation:

$$\begin{bmatrix} 0.2066 & 0.2066 \\ 0.0624 & -0.0001 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} -0.1755 - 0.6199\sigma^2 \\ -0.0936 \end{bmatrix}$$

The second-order transformation is given by $\mathcal{S}_2(y) = y^2 - y$ (since $\gamma = -1$ when g is sigmoid).

Example 3. If g is the ReLU function, i.e. $g(z) = \max\{0, z\}$, then $\alpha = -3\sqrt{\frac{2}{\pi}}, \beta = 3\left(\frac{4}{\pi} - \sigma^2 - 1\right)$ and $\gamma = -2\sqrt{\frac{2}{\pi}}$.

D. Proofs of Section 3

In this section, for the simplicity of the notation we denote the true parameters as $\boldsymbol{w}_i$'s and $\boldsymbol{a}_i$'s dropping the $*$ sign.

D.1. Proof of Theorem 1 for $k = 2$

Proof. Suppose that g is the linear activation function. For $k = 2$, (1) implies that

$$P_{y|\boldsymbol{x}} = f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \mathcal{N}(y|\boldsymbol{a}_1^\top \boldsymbol{x}, \sigma^2) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x})) \cdot \mathcal{N}(y|\boldsymbol{a}_2^\top \boldsymbol{x}, \sigma^2), \quad \boldsymbol{x} \sim \mathcal{N}(0, I_d), \tag{13}$$

where $f(\cdot)$ is the sigmoid function. Using the fact $\mathbb{E}[Z^3] = \mu^3 + 3\mu\sigma^2$ for any Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, we get

$$\mathbb{E}[y^3|\boldsymbol{x}] = f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_1^\top \boldsymbol{x})^3 + 3(\boldsymbol{a}_1^\top \boldsymbol{x})\sigma^2) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x}))((\boldsymbol{a}_1^\top \boldsymbol{x})^3 + 3(\boldsymbol{a}_1^\top \boldsymbol{x})\sigma^2).$$

Moreover,

$$\mathbb{E}[y|\boldsymbol{x}] = f(\boldsymbol{w}^\top \boldsymbol{x})(\boldsymbol{a}_1^\top \boldsymbol{x}) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x}))(\boldsymbol{a}_2^\top \boldsymbol{x}).$$

Thus,

$$\mathbb{E}[y^3 - 3y(1+\sigma^2)|\boldsymbol{x}] = f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x})) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x}))((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x})).$$

If we define $\mathcal{P}_3(y) \triangleq y^3 - 3y(1+\sigma^2)$, in view of Lemma 2 we get that

$$\begin{aligned} \mathcal{T}_3 = \mathbb{E}[\mathcal{P}_3(y) \cdot \mathcal{S}_3(\boldsymbol{x})] &= \mathbb{E}[(y^3 - 3y(1+\sigma^2)) \cdot \mathcal{S}_3(\boldsymbol{x})] \\ &= \mathbb{E}\left[\left(f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x}))\right) \cdot \mathcal{S}_3(\boldsymbol{x})\right] + \mathbb{E}\left[\left(1 - f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_2^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_2^\top \boldsymbol{x}))\right) \cdot \mathcal{S}_3(\boldsymbol{x})\right] \\ &= \mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)} \left(f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x}))\right)\right] + \mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)} \left(1 - f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_2^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_2^\top \boldsymbol{x}))\right)\right]. \end{aligned} \tag{14}$$

Using the chain rule for multi-derivatives, the first term simplifies to

$$\begin{aligned} \mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)} \left(f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x}))\right)\right] = \mathbb{E}[f'''((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x}))] \cdot \boldsymbol{w} \otimes \boldsymbol{w} \otimes \boldsymbol{w} + \mathbb{E}[f''(3(\boldsymbol{a}_1^\top \boldsymbol{x})^2 - 3)] \cdot \\ (\boldsymbol{w} \otimes \boldsymbol{w} \otimes \boldsymbol{a}_1 + \boldsymbol{w} \otimes \boldsymbol{a}_1 \otimes \boldsymbol{w} + \boldsymbol{a}_1 \otimes \boldsymbol{w} \otimes \boldsymbol{w}) + \\ \mathbb{E}[f'(6(\boldsymbol{a}_1^\top \boldsymbol{x}))] \cdot (\boldsymbol{a}_1 \otimes \boldsymbol{a}_1 \otimes \boldsymbol{w} + \boldsymbol{a}_1 \otimes \boldsymbol{w} \otimes \boldsymbol{a}_1 + \boldsymbol{w} \otimes \boldsymbol{a}_1 \otimes \boldsymbol{a}_1) + 6\mathbb{E}[f] \cdot \boldsymbol{a}_1 \otimes \boldsymbol{a}_1 \otimes \boldsymbol{a}_1. \end{aligned} \tag{15}$$

Since $f(z) = \frac{1}{1+e^{-z}}$, $f'(\cdot), f'''(\cdot)$ are even functions whereas $f''(\cdot)$ is an odd function. Furthermore, both $\boldsymbol{x} \mapsto (\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x})$ and $\boldsymbol{x} \mapsto \boldsymbol{a}_1^\top \boldsymbol{x}$ are odd functions whereas $\boldsymbol{x} \mapsto 3(\boldsymbol{a}_1^\top \boldsymbol{x})^2 - 3$ is an even function. Since $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$, $-\boldsymbol{x} \overset{(d)}{=} \boldsymbol{x}$. Thus all the expectation terms in (15) equal zero except for the last term since $\mathbb{E}[f(\boldsymbol{w}^\top \boldsymbol{x})] = \frac{1}{2} > 0$. We have,

$$\mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)} \left(f(\boldsymbol{w}^\top \boldsymbol{x})((\boldsymbol{a}_1^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_1^\top \boldsymbol{x}))\right)\right] = 3 \cdot \boldsymbol{a}_1 \otimes \boldsymbol{a}_1 \otimes \boldsymbol{a}_1.$$

Similarly,

$$\mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)}\left(1-f(\boldsymbol{w}^\top\boldsymbol{x})((\boldsymbol{a}_2^\top\boldsymbol{x})^3-3(\boldsymbol{a}_2^\top\boldsymbol{x}))\right)\right]=3\cdot\boldsymbol{a}_2\otimes\boldsymbol{a}_2\otimes\boldsymbol{a}_2.$$

Together, we have that

$$\mathcal{T}_3=3\cdot\boldsymbol{a}_1\otimes\boldsymbol{a}_1\otimes\boldsymbol{a}_1+3\cdot\boldsymbol{a}_2\otimes\boldsymbol{a}_2\otimes\boldsymbol{a}_2.$$

Now consider an arbitrary link function g belonging to the class of non-linearities described in Appendix C. Then

$$P_{y|\boldsymbol{x}}=f(\boldsymbol{w}^\top\boldsymbol{x})\cdot\mathcal{N}(y|g(\boldsymbol{a}_1^\top\boldsymbol{x}),\sigma^2)+(1-f(\boldsymbol{w}^\top\boldsymbol{x}))\cdot\mathcal{N}(y|g(\boldsymbol{a}_2^\top\boldsymbol{x}),\sigma^2),\quad \boldsymbol{x}\sim\mathcal{N}(0,I_d),$$

implies that

$$\mathbb{E}[y^3|\boldsymbol{x}]=f(\boldsymbol{w}^\top\boldsymbol{x})(g(\boldsymbol{a}_1^\top\boldsymbol{x})^3+3g(\boldsymbol{a}_1^\top\boldsymbol{x})\sigma^2)+(1-f(\boldsymbol{w}^\top\boldsymbol{x}))(g(\boldsymbol{a}_2^\top\boldsymbol{x})^3+3g(\boldsymbol{a}_2^\top\boldsymbol{x})\sigma^2),$$

and

$$\begin{aligned}\mathbb{E}[y^2|\boldsymbol{x}]&=f(\boldsymbol{w}^\top\boldsymbol{x})(g(\boldsymbol{a}_1^\top\boldsymbol{x})^2+\sigma^2)+(1-f(\boldsymbol{w}^\top\boldsymbol{x}))(g(\boldsymbol{a}_2^\top\boldsymbol{x})^2+\sigma^2),\\ \mathbb{E}[y|\boldsymbol{x}]&=f(\boldsymbol{w}^\top\boldsymbol{x})g(\boldsymbol{a}_1^\top\boldsymbol{x})+(1-f(\boldsymbol{w}^\top\boldsymbol{x}))g(\boldsymbol{a}_2^\top\boldsymbol{x}).\end{aligned}$$

If we define $\mathcal{P}_3(y)\triangleq y^3+\alpha y^2+\beta y$, we have that

$$\begin{aligned}\mathcal{T}_3=\mathbb{E}[\mathcal{P}_3(y)\cdot\mathcal{S}_3(\boldsymbol{x})]=\mathbb{E}[\mathbb{E}[y^3+\alpha y^2+\beta y|\boldsymbol{x}]\cdot\mathcal{S}_3(\boldsymbol{x})]&\\ =\mathbb{E}\left[f(\boldsymbol{w}^\top\boldsymbol{x})\left(g(\boldsymbol{a}_1^\top\boldsymbol{x})^3+\alpha g(\boldsymbol{a}_1^\top\boldsymbol{x})^2+g(\boldsymbol{a}_1^\top\boldsymbol{x})(\beta+3\sigma^2)\right)\cdot\mathcal{S}_3(\boldsymbol{x})\right]+&\\ \mathbb{E}\left[(1-f(\boldsymbol{w}^\top\boldsymbol{x}))\left(g(\boldsymbol{a}_2^\top\boldsymbol{x})^3+\alpha g(\boldsymbol{a}_2^\top\boldsymbol{x})^2+g(\boldsymbol{a}_2^\top\boldsymbol{x})(\beta+3\sigma^2)\right)\cdot\mathcal{S}_3(\boldsymbol{x})\right]&\\ =\mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)}\left(f(\boldsymbol{w}^\top\boldsymbol{x})\left(g(\boldsymbol{a}_1^\top\boldsymbol{x})^3+\alpha g(\boldsymbol{a}_1^\top\boldsymbol{x})^2+g(\boldsymbol{a}_1^\top\boldsymbol{x})(\beta+3\sigma^2)\right)\right)\right]+&\\ \mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)}\left(f(\boldsymbol{w}^\top\boldsymbol{x})\left(g(\boldsymbol{a}_2^\top\boldsymbol{x})^3+\alpha g(\boldsymbol{a}_2^\top\boldsymbol{x})^2+g(\boldsymbol{a}_2^\top\boldsymbol{x})(\beta+3\sigma^2)\right)\right)\right]&\\ \overset{(a)}{=}\mathbb{E}[f]\mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)}\left(g(\boldsymbol{a}_1^\top\boldsymbol{x})^3+\alpha g(\boldsymbol{a}_1^\top\boldsymbol{x})^2+g(\boldsymbol{a}_1^\top\boldsymbol{x})(\beta+3\sigma^2)\right)\right]\cdot\boldsymbol{a}_1\otimes\boldsymbol{a}_1\otimes\boldsymbol{a}_1+&\\ \mathbb{E}[1-f]\mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)}\left(g(\boldsymbol{a}_2^\top\boldsymbol{x})^3+\alpha g(\boldsymbol{a}_2^\top\boldsymbol{x})^2+g(\boldsymbol{a}_2^\top\boldsymbol{x})(\beta+3\sigma^2)\right)\right]\cdot\boldsymbol{a}_2\otimes\boldsymbol{a}_2\otimes\boldsymbol{a}_2&\\ =c_{g,\sigma}\left(\mathbb{E}[f]\cdot\boldsymbol{a}_1\otimes\boldsymbol{a}_1\otimes\boldsymbol{a}_1+\mathbb{E}[1-f]\cdot\boldsymbol{a}_2\otimes\boldsymbol{a}_2\otimes\boldsymbol{a}_2\right),&\end{aligned}$$

where (a) follows from the choice of α and β and the fact that $\boldsymbol{w}\perp\{\boldsymbol{a}_1,\boldsymbol{a}_2\}$, and $c_{g,\sigma}\triangleq\mathbb{E}\left[\left(g(Z)^3+\alpha g(Z)^2+g(Z)(\beta+3\sigma^2)\right)'''\right]$ where $Z\sim\mathcal{N}(0,1)$. The proof for $\mathcal{T}_2$ is similar.

□

D.2. Proof of Theorem 1 for general k

Proof. The proof for general k closely follows that of $k=2$, described in Appendix D.1. For the general k, we first prove the theorem when g is the identity function, i.e.

$$P_{y|\boldsymbol{x}}=\sum_{i\in[k]}P_{i|\boldsymbol{x}}P_{y|\boldsymbol{x},i}=\sum_{i\in[k]}\frac{e^{\boldsymbol{w}_i^\top\boldsymbol{x}}}{\sum_{i\in[k]}e^{\boldsymbol{w}_i^\top\boldsymbol{x}}}\cdot\mathcal{N}(y|\boldsymbol{a}_i^\top\boldsymbol{x},\sigma^2),\quad \boldsymbol{x}\sim\mathcal{N}(0,I_d).$$

Denoting $P_{i|\boldsymbol{x}}$ by $p_i(\boldsymbol{x})$, we have that

$$\begin{aligned}\mathbb{E}[y^3|\boldsymbol{x}]&=\sum_{i\in[k]}p_i(\boldsymbol{x})\left((\boldsymbol{a}_i^\top\boldsymbol{x})^3+3(\boldsymbol{a}_i^\top\boldsymbol{x})\sigma^2\right),\\ \mathbb{E}[y|\boldsymbol{x}]&=\sum_{i\in[k]}p_i(\boldsymbol{x})(\boldsymbol{a}_i^\top\boldsymbol{x}).\end{aligned}$$

Hence

$$\mathbb{E}[y^3 - 3y(1+\sigma^2)|\boldsymbol{x}] = \sum_{i\in[k]} p_i(\boldsymbol{x}) \left((\boldsymbol{a}_i^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_i^\top \boldsymbol{x}) \right)$$

If we let $\mathcal{P}_3(y) \triangleq y^3 - 3y(1+\sigma^2)$, we get

$$\mathbb{E}[\mathcal{P}_3(y) \cdot \mathcal{S}_3(\boldsymbol{x})] = \sum_{i\in[k]} \mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)} \left(p_i(\boldsymbol{x}) \left((\boldsymbol{a}_i^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_i^\top \boldsymbol{x}) \right) \right) \right]$$

Since $\boldsymbol{x} \sim \mathcal{N}(0, I_d)$ and $\boldsymbol{a}_i \perp \mathrm{span}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{k-1}\}$, we have that $\boldsymbol{a}_i^\top \boldsymbol{x} \perp (\boldsymbol{w}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{w}_{k-1}^\top \boldsymbol{x})$. Moreover, $\mathbb{E}[(\boldsymbol{a}_i^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_i^\top \boldsymbol{x})] = \mathbb{E}[(\boldsymbol{a}_i^\top \boldsymbol{x})^2 - 1] = \mathbb{E}[\boldsymbol{a}_i^\top \boldsymbol{x}] = 0$ for each i. Using the chain-rule for multi-derivatives, the above equation thus simplifies to

$$\mathbb{E}[\mathcal{P}_3(y) \cdot \mathcal{S}_3(\boldsymbol{x})] = \sum_{i\in[k]} \mathbb{E}[p_i(\boldsymbol{x})] \cdot \mathbb{E}\left[\nabla_{\boldsymbol{x}}^{(3)} \left((\boldsymbol{a}_i^\top \boldsymbol{x})^3 - 3(\boldsymbol{a}_i^\top \boldsymbol{x}) \right) \right] = \sum_{i\in[k]} 6\mathbb{E}[p_i(\boldsymbol{x})] \cdot \boldsymbol{a}_i \otimes \boldsymbol{a}_i \times \boldsymbol{a}_i.$$

For a generic $g : \mathbb{R} \to \mathbb{R}$ which is $(\alpha, \beta, \gamma)-$valid, let $\mathcal{P}_3(y) = y^3 + \alpha y^2 + \beta y$. Then it is easy to see that the same proof goes through except for a change in the coefficients of rank-1 terms, i.e.

$$\mathbb{E}[\mathcal{P}_3(y) \cdot \mathcal{S}_3(\boldsymbol{x})] = \sum_{i\in[k]} \alpha_i \mathbb{E}[p_i(\boldsymbol{x})] \cdot \boldsymbol{a}_i \otimes \boldsymbol{a}_i \otimes \boldsymbol{a}_i,$$

where $\alpha_i \triangleq \mathbb{E}\left[\left(g(Z)^3 + \alpha g(Z)^2 + g(Z)(\beta + 3\sigma^2) \right)''' \right]$ where $Z \sim \mathcal{N}(0,1)$ and $'''$ denotes the third-derivative with respect to Z. Note that Condition 2 together with the fact that $\mathbb{E}[p_i(\boldsymbol{x})] > 0$ ensures that $\alpha_i \neq 0$ and thus the coefficients of the rank-1 terms are non-zero. The proof for $\mathcal{T}_2$ is similar. □

D.3. Proof of Theorem 2

The following two lemmas are central to the proof of Theorem 2. Let $\boldsymbol{A}^\top = [\boldsymbol{a}_1 | \ldots | \boldsymbol{a}_k] \in \mathbb{R}^{d \times k}$ denote the matrix of regressor parameters whereas $\boldsymbol{W}^\top = [\boldsymbol{w}_1 | \ldots | \boldsymbol{w}_{k-1}] \in \mathbb{R}^{d \times (k-1)}$ denote the matrix of gating parameters. With a slight change of notation, when $\boldsymbol{A} = \boldsymbol{A}^*$, we denote the EM operator $M(\boldsymbol{W})$ as either $M(\boldsymbol{W}, \boldsymbol{A}^*)$ or $M(\boldsymbol{w})$, introduced in Section 3. For the general case, we simply denote it by $M(\boldsymbol{W}, \boldsymbol{A})$. In the following lemmas, we use the norm $\|\boldsymbol{A}\| = \max_{i\in[k]} \|\boldsymbol{A}_i^\top\|_2$ where $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ is a matrix of regressors, similarly for any matrix of classifiers $\boldsymbol{W} \in \mathbb{R}^{(k-1)\times d}$.

Lemma 3 (Contraction of the EM operator)**.** *Under the assumptions of Theorem 2, we have that*

$$\|M(\boldsymbol{W}, \boldsymbol{A}^*) - \boldsymbol{W}^*\| \leq \kappa_\sigma \|\boldsymbol{W} - \boldsymbol{W}^*\|.$$

Moreover, $\boldsymbol{W} = \boldsymbol{W}^*$ is a fixed point for $M(\boldsymbol{W}, \boldsymbol{A}^*)$.

Lemma 4 (Robustness of the EM operator)**.** *Let the matrix of regressors $\boldsymbol{A}$ be such that $\max_{i\in[k]} \|\boldsymbol{A}_i^\top - (\boldsymbol{A}_i^*)^\top\|_2 = \sigma^2 \varepsilon_1$. Then for any $\boldsymbol{W} \in \Omega$, we have that*

$$\|M(\boldsymbol{W}, \boldsymbol{A}) - M(\boldsymbol{W}, \boldsymbol{A}^*)\| \leq \kappa \varepsilon_1,$$

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

We are now ready to prove Theorem 2.

Proof. We first note that the EM iterates $\{\boldsymbol{W}_t\}_{t\geq 1}$ evolve according to

$$\boldsymbol{W}_t = M(\boldsymbol{W}_{t-1}, \boldsymbol{A}), \quad t \geq 1$$

Thus

$$\begin{aligned} \|\boldsymbol{W}_t - \boldsymbol{W}^*\| = \|M(\boldsymbol{W}_{t-1}, \boldsymbol{A}) - \boldsymbol{W}^*\| &= \|M(\boldsymbol{W}_{t-1}, \boldsymbol{A}) - M(\boldsymbol{W}^*, \boldsymbol{A}^*)\| \\ &\leq \|M(\boldsymbol{W}_{t-1}, \boldsymbol{A}) - M(\boldsymbol{W}_{t-1}, \boldsymbol{A}^*)\| + \|M(\boldsymbol{W}_{t-1}, \boldsymbol{A}^*) - \boldsymbol{W}^*\| \\ &\leq k\varepsilon_1 + \kappa_\sigma \|\boldsymbol{W}_{t-1} - \boldsymbol{W}^*\|, \end{aligned}$$

where the last inequality follows from Lemma 3 and Lemma 4. Recursively using the above inequality, we obtain that

$$\|\boldsymbol{W}_t - \boldsymbol{W}^*\| \leq (\kappa_\sigma)^t \|\boldsymbol{W}_0 - \boldsymbol{W}^*\| + \kappa\varepsilon_1(1 + \kappa_\sigma + \ldots + \kappa_\sigma^{t-1}) \leq (\kappa_\sigma)^t \|\boldsymbol{W}_0 - \boldsymbol{W}^*\| + \frac{\kappa\varepsilon_1}{1-\kappa_\sigma}.$$

□

D.4. Proof of Theorem 3

Proof. We are given that $(\boldsymbol{a}_1, \boldsymbol{a}_2) = (\boldsymbol{a}_1^*, \boldsymbol{a}_2^*)$. Denoting $\boldsymbol{w}^*$ with $\boldsymbol{w}$, from (13), we have that

$$\begin{aligned}\mathbb{E}[y|\boldsymbol{x}] &= f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{a}_1^\top \boldsymbol{x} + (1 - f(\boldsymbol{w}^\top \boldsymbol{x})) \cdot \boldsymbol{a}_2^\top \boldsymbol{x}, &(16)\\ &= \boldsymbol{a}_2^\top \boldsymbol{x} + f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot (\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}. &(17)\end{aligned}$$

Thus,

$$\frac{\mathbb{E}[y|\boldsymbol{x}] - \boldsymbol{a}_2^\top \boldsymbol{x}}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}} = f(\boldsymbol{w}^\top \boldsymbol{x}).$$

Notice that in the above equation we have $(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}$ in the denominator. But this equals zero with zero probability whenever $\boldsymbol{x}$ is generated from a continuous distribution; in our case $\boldsymbol{x}$ is Gaussian. Thus we may write

$$\begin{aligned}\mathbb{E}\left[\left(\frac{y - \boldsymbol{a}_2^\top \boldsymbol{x}}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}}\right) \cdot \boldsymbol{x}\right] \overset{\text{✗}}{=} \mathbb{E}\left[\left(\frac{\mathbb{E}[y|\boldsymbol{x}] - \boldsymbol{a}_2^\top \boldsymbol{x}}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}}\right) \cdot \boldsymbol{x}\right] &= \mathbb{E}\left[f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{x}\right]\\ &= \mathbb{E}\left[f'(\boldsymbol{w}^\top \boldsymbol{x})\right] \cdot \boldsymbol{w}\\ &= \mathbb{E}_{Z \sim \mathcal{N}(0,1)} f'(\|\boldsymbol{w}\| Z) \cdot \boldsymbol{w}\\ &\propto \boldsymbol{w}.\end{aligned}$$

However, it turns out that the above chain of equalities does not hold. Surprisingly, the first equality, which essentially is the law of iterated expectations, is not valid in this case as $\frac{y - \boldsymbol{a}_2^\top \boldsymbol{x}}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}}$ is not integrable. To see this, notice that the model in (13) can also be written as

$$y \overset{(d)}{=} Z(\boldsymbol{a}_1^\top \boldsymbol{x}) + (1 - Z)(\boldsymbol{a}_2^\top \boldsymbol{x}) + \sigma N, \quad Z \sim \mathrm{Bern}(f(\boldsymbol{w}^\top \boldsymbol{x})), N \sim \mathcal{N}(0,1).$$

Thus,

$$\mathrm{Ratio} \triangleq \frac{y - \boldsymbol{a}_2^\top \boldsymbol{x}}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}} \overset{(d)}{=} Z + \frac{\sigma N}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}}.$$

Since Z is independent of N and $\frac{N}{(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}}$ is a Cauchy random variable, it follows that the random variable Ratio is not integrable. To deal with the non-integrability of Ratio, we look at its conditional cdf, given by

$$\mathbb{P}\left[\mathrm{Ratio} \leq z|\boldsymbol{x}\right] = f(\boldsymbol{w}^\top \boldsymbol{x})\Phi\left((z-1)\frac{|\Delta_x|}{\sigma}\right) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x}))\Phi\left(z\frac{|\Delta_x|}{\sigma}\right), \quad \Delta_x = (\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x},$$

where $\Phi(\cdot)$ is the standard Gaussian cdf. Substituting $z = 0.5$ and using the fact that $\Phi(z) + \Phi(-z) = 1$, we obtain

$$\begin{aligned}\mathbb{P}\left[\mathrm{Ratio} \leq 0.5|\boldsymbol{x}\right] &= f(\boldsymbol{w}^\top \boldsymbol{x})\Phi\left(-\frac{|\Delta_x|}{2\sigma}\right) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x}))\Phi\left(\frac{|\Delta_x|}{2\sigma}\right)\\ &= \Phi\left(\frac{|(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right) + f(\boldsymbol{w}^\top \boldsymbol{x})\left(1 - 2\Phi\left(\frac{|(\boldsymbol{a}_1 - \boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right)\right).\end{aligned}$$

Since $\Phi\left(\frac{|(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right)$ is a symmetric function in $\boldsymbol{x}$ its first moment with $\boldsymbol{x}$ equals zero. Furthermore, if we assume that $\boldsymbol{w}$ is orthogonal to $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\left\{\text{Ratio} \leq 0.5\right\} \cdot \boldsymbol{x}\right] &= \mathbb{E}\left[\mathbb{P}\left[\text{Ratio} \leq 0.5|\boldsymbol{x}\right] \cdot \boldsymbol{x}\right] \\
&= \mathbb{E}\left[f(\boldsymbol{w}^\top \boldsymbol{x})\left(1-2\Phi\left(\frac{|(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right)\right)\cdot \boldsymbol{x}\right] \\
&= \mathbb{E}[f'(\boldsymbol{w}^\top \boldsymbol{x})]\cdot \mathbb{E}\left(1-2\Phi\left(\frac{|(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right)\right)\cdot \boldsymbol{w}+ \\
&\qquad \mathbb{E}[f(\boldsymbol{w}^\top \boldsymbol{x})]\cdot \underbrace{\mathbb{E}\left[\nabla_{\boldsymbol{x}}\left(1-2\Phi\left(\frac{|(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right)\right)\right]}_{=0,\text{ since derivative of a even function is odd}} \\
&= \mathbb{E}[f'(\boldsymbol{w}^\top \boldsymbol{x})]\cdot \mathbb{E}\left(1-2\Phi\left(\frac{|(\boldsymbol{a}_1-\boldsymbol{a}_2)^\top \boldsymbol{x}|}{2\sigma}\right)\right)\cdot \boldsymbol{w} \\
&\propto \boldsymbol{w}.
\end{aligned}$$

Thus, if $\|\boldsymbol{w}\| = 1$, we have that

$$\frac{\mathbb{E}\left[\mathbb{1}\left\{\text{Ratio} \leq 0.5\right\} \cdot \boldsymbol{x}\right]}{\|\mathbb{E}\left[\mathbb{1}\left\{\text{Ratio} \leq 0.5\right\} \cdot \boldsymbol{x}\right]\|} = \boldsymbol{w}.$$

In the finite sample regime, $\mathbb{E}\left[\mathbb{1}\left\{\text{Ratio} \leq 0.5\right\} \cdot \boldsymbol{x}\right]$ can be estimated from samples using the empirical moments and its normalized version will be an estimate of $\boldsymbol{w}$. □

E. Proof of Lemma 4

We need the following lemma which establishes the stability of the minimizers for strongly convex functions under Lipschitz perturbations.

Lemma 5. *Suppose $\Omega \subseteq \mathbb{R}^d$ is a closed convex subset, $f : \Omega \to \mathbb{R}$ is a λ-strongly convex function for some $\lambda > 0$ and B is an L-Lipschitz continuous function on Ω. Let $\boldsymbol{w}_f = \operatorname{argmin}_{\boldsymbol{w}\in\Omega} f(\boldsymbol{w})$ and $\boldsymbol{w}_{f+B} = \operatorname{argmin}_{\boldsymbol{w}\in\Omega} f(\boldsymbol{w}) + B(\boldsymbol{w})$. Then*

$$\|\boldsymbol{w}_f - \boldsymbol{w}_{f+B}\| \leq \frac{L}{\lambda}.$$

Proof. Let $\boldsymbol{w}' \in \Omega$ be such that $\|\boldsymbol{w}' - \boldsymbol{w}_f\| > \frac{L}{\lambda}$. Let $\boldsymbol{w}_\alpha = \alpha\boldsymbol{w}_f + (1-\alpha)\boldsymbol{w}'$ for $0 < \alpha < 1$. From the fact that $\boldsymbol{w}_f$ is the minimizer of f on Ω and that f is strongly convex, we have that

$$f(\boldsymbol{w}') \geq f(\boldsymbol{w}_f) + \frac{\lambda\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2}{2}.$$

Furthermore, the strong-convexity of f implies that

$$\begin{aligned}
f(\boldsymbol{w}_\alpha) &\leq \alpha f(\boldsymbol{w}_f) + (1-\alpha) f(\boldsymbol{w}') - \frac{\alpha(1-\alpha)\lambda}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2 \\
&= f(\boldsymbol{w}') + \alpha(f(\boldsymbol{w}_f) - f(\boldsymbol{w}')) - \frac{\alpha(1-\alpha)\lambda}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2 \\
&\leq f(\boldsymbol{w}') - \alpha\frac{\lambda\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2}{2} - \frac{\alpha(1-\alpha)\lambda}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2 \\
&= f(\boldsymbol{w}') - \lambda\alpha\left(1-\frac{\alpha}{2}\right)\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2 \qquad (18)
\end{aligned}$$

Since B is L-Lipschitz, we have

$$B(\boldsymbol{w}_\alpha) \leq B(\boldsymbol{w}') + L\alpha\left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|. \qquad (19)$$

Adding (18) and (19), we get

$$\begin{aligned} f(\boldsymbol{w}_\alpha) + B(\boldsymbol{w}_\alpha) &\leq f(\boldsymbol{w}') + B(\boldsymbol{w}') + L\alpha \left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\| - \lambda\alpha \left(1 - \frac{\alpha}{2}\right) \left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\|^2 \\ &= f(\boldsymbol{w}') + B(\boldsymbol{w}') + \alpha\lambda \left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\| \left(\frac{L}{\lambda} - \left(1 - \frac{\alpha}{2}\right) \left\|\boldsymbol{w}' - \boldsymbol{w}_f\right\| \right) \end{aligned}$$

By the assumption that $\|\boldsymbol{w}' - \boldsymbol{w}_f\| > \frac{L}{\lambda}$, the term $\frac{L}{\lambda} - \left(1 - \frac{\alpha}{2}\right) \|\boldsymbol{w}' - \boldsymbol{w}_f\|$ will be negative for sufficiently small α. This in turn implies that $f(\boldsymbol{w}_\alpha) + B(\boldsymbol{w}_\alpha) < f(\boldsymbol{w}') + B(\boldsymbol{w}')$ for such α. Consequently $\boldsymbol{w}'$ is not a minimizer of $f + B$ for any $\boldsymbol{w}'$ such that $\|\boldsymbol{w}' - \boldsymbol{w}_f\| > \frac{L}{\lambda}$. The conclusion follows.

□

We are now ready to prove Lemma 4. Fix any $\boldsymbol{W} \in \Omega$ and let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \dots \\ \boldsymbol{a}_k^\top \end{bmatrix} \in \mathbb{R}^{k \times d}$ be such that $\max_{i\in[k]} \|\boldsymbol{a}_i - \boldsymbol{a}_i^*\|_2 = \sigma^2 \varepsilon_1$ for some $\varepsilon_1 > 0$. Let

$$\boldsymbol{W}' = M(\boldsymbol{W}, \boldsymbol{A}), \quad (\boldsymbol{W}')^* = M(\boldsymbol{W}, \boldsymbol{A}^*),$$

where,

$$M(\boldsymbol{W}, \boldsymbol{A}) = \arg \max_{\boldsymbol{W}' \in \Omega} Q(\boldsymbol{W}' | \boldsymbol{W}, \boldsymbol{A}),$$

and,

$$Q(\boldsymbol{W}' | \boldsymbol{W}, \boldsymbol{A}) = \mathbb{E} \left[\sum_{i\in[k-1]} p^{(i)}(\boldsymbol{W}, \boldsymbol{A})((\boldsymbol{W}'_i)^\top \boldsymbol{x}) - \log \left(1 + \sum_{i\in[k-1]} e^{(\boldsymbol{W}'_i)^\top \boldsymbol{x}} \right) \right].$$

Here $p^{(i)}(\boldsymbol{A}, \boldsymbol{W}) \triangleq \frac{p_i(\boldsymbol{x}) N_i}{\sum_{i\in[k]} p_i(\boldsymbol{x}) N_i}$ denotes the posterior probability of choosing the i^{th} expert, where

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

Since both $Q(\cdot | \boldsymbol{W}, \boldsymbol{A})$ and $Q(\cdot | \boldsymbol{W}, \boldsymbol{A}^*)$ are strongly concave functions over Ω with some strong-concavity parameter λ, Lemma 5 implies that

$$\|M(\boldsymbol{W}, \boldsymbol{A}) - M(\boldsymbol{W}, \boldsymbol{A}^*)\| \leq \frac{L}{\lambda},$$

where L is the Lipschitz-constant for the function $l(\cdot) \triangleq Q(\cdot | \boldsymbol{W}, \boldsymbol{A}) - Q(\cdot | \boldsymbol{W}, \boldsymbol{A}^*)$. We have that

$$l(\boldsymbol{W}') = \sum_{i\in[k-1]} \mathbb{E}[(p^{(i)}(\boldsymbol{W}, \boldsymbol{A}) - p^{(i)}(\boldsymbol{W}, \boldsymbol{A}^*)(\boldsymbol{W}'_i)^\top \boldsymbol{x})]$$

Without loss of generality let $i = 1$. Since $l(\cdot)$ is linear in $\boldsymbol{W}'$, it suffices to show for each i that

$$\left\| \mathbb{E}[(p^{(1)}(\boldsymbol{W}, \boldsymbol{A}) - p^{(1)}(\boldsymbol{W}, \boldsymbol{A}^*)\boldsymbol{x}] \right\| \leq L,$$

We show that $L = \kappa\varepsilon_1$, or equivalently,

$$\left\| \mathbb{E}[(p^{(1)}(\boldsymbol{W}, \boldsymbol{A}) - p^{(1)}(\boldsymbol{W}, \boldsymbol{A}^*)\boldsymbol{x}] \right\| \leq \kappa\varepsilon_1,$$

Let

$$\boldsymbol{A}_t = \boldsymbol{A}^* + t\Delta, \quad \Delta = \boldsymbol{A} - \boldsymbol{A}^* \in \mathbb{R}^{k\times d}.$$

By hypothesis, we have that $\|\Delta_i\|_2 \leq \sigma^2 \varepsilon_1$ for all $i \in [k]$. Thus in order to show that

$$\left\|\mathbb{E}[(p^{(1)}(\boldsymbol{A}, \boldsymbol{W}) - p^{(1)}(\boldsymbol{A}^*, \boldsymbol{W}))\boldsymbol{x}]\right\|_2 \leq \kappa \varepsilon_1,$$

it suffices to show that

$$\langle \mathbb{E}[(p^{(1)}(\boldsymbol{A}, \boldsymbol{W}) - p^{(1)}(\boldsymbol{A}^*, \boldsymbol{W}))\boldsymbol{x}], \tilde{\Delta} \rangle \leq \kappa \left\|\Delta/\sigma^2\right\|_2 \|\tilde{\Delta}\|_2, \quad \text{for all } \tilde{\Delta} \in \mathbb{R}^d.$$

Or equivalently,

$$\mathbb{E}[(p^{(1)}(\boldsymbol{A}, \boldsymbol{W}) - p^{(1)}(\boldsymbol{A}^*, \boldsymbol{W}))\langle \boldsymbol{x}, \tilde{\Delta} \rangle] \leq \kappa \left\|\Delta/\sigma^2\right\|_2 \|\tilde{\Delta}\|_2.$$

We can rewrite the difference of the posteriors as

$$p^{(1)}(\boldsymbol{A}, \boldsymbol{W}) - p^{(1)}(\boldsymbol{A}^*, \boldsymbol{W}) = \int_0^1 \frac{d}{dt} p^{(1)}(\boldsymbol{A}^* + t\Delta, \boldsymbol{W}) dt = \sum_{i \in [k]} \int_0^1 \langle \nabla_{\boldsymbol{a}_i} p^{(1)}(\boldsymbol{A}_t, \boldsymbol{W}), \Delta_i \rangle dt. \tag{20}$$

Since $N_i = \mathcal{N}(y | g(\boldsymbol{a}_i^\top \boldsymbol{x}), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y - g(\boldsymbol{a}_1^\top \boldsymbol{x}))^2 / 2\sigma^2}$, we have that

$$\nabla_{\boldsymbol{a}_i} N_i = N_i \left(\frac{y - g(\boldsymbol{a}_i^\top \boldsymbol{x})}{\sigma^2} \right) g'(\boldsymbol{a}_i^\top \boldsymbol{x}).$$

Thus,

$$\begin{aligned} \nabla_{\boldsymbol{a}_i} p^{(1)}(\boldsymbol{A}_t, \boldsymbol{W}) &= \nabla_{\boldsymbol{a}_i} \left(\frac{p_1(\boldsymbol{x}) N_1}{\sum_{i \in [k]} p_i(\boldsymbol{x}) N_i} \right) \\ &= \begin{cases} \frac{(\sum_{i \neq 1} p_i(\boldsymbol{x}) N_i) p_1(\boldsymbol{x}) N_1}{(\sum_i p_i(\boldsymbol{x}) N_i)^2} \left(\frac{y - g(\boldsymbol{a}_1^\top \boldsymbol{x})}{\sigma^2} \right) g'(\boldsymbol{a}_1^\top \boldsymbol{x}) \boldsymbol{x}, & \text{if } i = 1 \\ \frac{-p_i(\boldsymbol{x}) p_1(\boldsymbol{x}) N_i N_1}{(\sum_i p_i(\boldsymbol{x}) N_i)^2} \left(\frac{y - g(\boldsymbol{a}_i^\top \boldsymbol{x})}{\sigma^2} \right) g'(\boldsymbol{a}_i^\top \boldsymbol{x}) \boldsymbol{x}, & \text{if } i \neq 1 \end{cases} \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}[(p^{(1)}(\boldsymbol{A}, \boldsymbol{W}) - p^{(1)}(\boldsymbol{A}^*, \boldsymbol{W}))\langle \boldsymbol{x}, \tilde{\Delta} \rangle] &= \sum_{i \in [k]} \int_0^1 \mathbb{E}[\langle \nabla_{\boldsymbol{a}_i} p^{(1)}(\boldsymbol{A}_t, \boldsymbol{W}), \Delta_i \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle] dt \qquad (21) \\ &= \int_0^1 \mathbb{E} \left[\frac{(\sum_{i \neq 1} p_i(\boldsymbol{x}) N_i) p_1(\boldsymbol{x}) N_1}{(\sum_i p_i(\boldsymbol{x}) N_i)^2} \left(\frac{y - g(\boldsymbol{a}_1^\top \boldsymbol{x})}{\sigma^2} \right) g'(\boldsymbol{a}_1^\top \boldsymbol{x}) \langle \boldsymbol{x}, \Delta_1 \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle \right] dt \qquad (22) \\ &\quad + \sum_{i \neq 1} \int_0^1 \mathbb{E} \left[\frac{-p_i(\boldsymbol{x}) p_1(\boldsymbol{x}) N_i N_1}{(\sum_i p_i(\boldsymbol{x}) N_i)^2} \left(\frac{y - g(\boldsymbol{a}_i^\top \boldsymbol{x})}{\sigma^2} \right) g'(\boldsymbol{a}_i^\top \boldsymbol{x}) \langle \boldsymbol{x}, \Delta_i \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle \right] dt, \qquad (23) \end{aligned}$$

where we denoted $(\boldsymbol{a}_i)_t$ by $\boldsymbol{a}_i$ in the integrals above(with a slight abuse of notation) for the sake of notational simplicity. For any $i \neq 1$, we have that

$$\begin{aligned} &\left| \frac{-p_i(\boldsymbol{x}) p_1(\boldsymbol{x}) N_i N_1}{(\sum_i p_i(\boldsymbol{x}) N_i)^2} \left(\frac{y - g(\boldsymbol{a}_i^\top \boldsymbol{x})}{\sigma^2} \right) g'(\boldsymbol{a}_i^\top \boldsymbol{x}) \langle \boldsymbol{x}, \Delta_i \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle \right| \\ &\leq \frac{p_i(\boldsymbol{x}) p_1(\boldsymbol{x}) N_i N_1}{(p_1(\boldsymbol{x}) N_1 + p_i(\boldsymbol{x}) N_i)^2} |(y - g(\boldsymbol{a}_i^\top \boldsymbol{x})) g'(\boldsymbol{a}_i^\top \boldsymbol{x}) \langle \boldsymbol{x}, \Delta_i / \sigma^2 \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle| \end{aligned}$$

For $g =$linear, sigmoid and ReLU, we have that $|g'(\cdot)| \leq 1$. Moreover, $\frac{p_i(\boldsymbol{x}) p_1(\boldsymbol{x}) N_i N_1}{(p_1(\boldsymbol{x}) N_1 + p_i(\boldsymbol{x}) N_i)^2} \leq 1/4$. Thus we have

$$\frac{p_i(\boldsymbol{x}) p_1(\boldsymbol{x}) N_i N_1}{(p_1(\boldsymbol{x}) N_1 + p_i(\boldsymbol{x}) N_i)^2} |(y - g(\boldsymbol{a}_i^\top \boldsymbol{x})) g'(\boldsymbol{a}_i^\top \boldsymbol{x}) \langle \boldsymbol{x}, \Delta_i / \sigma^2 \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle| \leq \frac{1}{4} |y - g(\boldsymbol{a}_i^\top \boldsymbol{x})| |\langle \boldsymbol{x}, \Delta_i / \sigma^2 \rangle \langle \boldsymbol{x}, \tilde{\Delta} \rangle|.$$

We thus get

$$\begin{aligned}\mathbb{E}\left[\frac{-p_i(\boldsymbol{x})p_1(\boldsymbol{x})N_iN_1}{(\sum_i p_i(\boldsymbol{x})N_i)^2}\left(\frac{y-g(\boldsymbol{a}_i^\top\boldsymbol{x})}{\sigma^2}\right)g'(\boldsymbol{a}_i^\top\boldsymbol{x})\langle\boldsymbol{x},\Delta_i\rangle\langle\boldsymbol{x},\tilde{\Delta}\rangle\right] &\leq \frac{1}{4}\mathbb{E}[|y-g(\boldsymbol{a}_i^\top\boldsymbol{x})||\langle\boldsymbol{x},\Delta_i/\sigma^2\rangle\langle\boldsymbol{x},\tilde{\Delta}\rangle|] \qquad (24)\\ &\leq \frac{1}{4}\sqrt{\mathbb{E}[(y-g(\boldsymbol{a}_i^\top\boldsymbol{x}))^2]\mathbb{E}[\langle\boldsymbol{x},\Delta_i/\sigma^2\rangle^2\langle\boldsymbol{x},\tilde{\Delta}\rangle^2]} \qquad (25)\\ &\leq \frac{\sqrt{3}}{4}\sqrt{\mathbb{E}[(y-g(\boldsymbol{a}_i^\top\boldsymbol{x}))^2]}\|\Delta_i/\sigma^2\|_2\|\tilde{\Delta}\|_2 \qquad (26)\end{aligned}$$

Now it remains to bound $\sqrt{\mathbb{E}[(y-g(\boldsymbol{a}_i^\top\boldsymbol{x}))^2]}$. Since $\|\boldsymbol{a}_i\|_2 \leq 1$, one can show that $\mathbb{E}[g(\boldsymbol{a}_i^\top\boldsymbol{x})^2] \leq 1$ for the given choice of non-linearities for g. Also, we have that

$$\mathbb{E}[y^2] = \mathbb{E}[\mathbb{E}[y^2|\boldsymbol{x}]] = \mathbb{E}[\sum_{i\in[k]} p_i^*(\boldsymbol{x})g(\langle\boldsymbol{a}_i^*,\boldsymbol{x}\rangle)^2+\sigma^2] = \mathbb{E}[\sum_{i\in[k]} p_i^*(\boldsymbol{x})]\mathbb{E}[g(\langle\boldsymbol{a}_1^*,\boldsymbol{x}\rangle)^2]+\sigma^2 \leq 1+\sigma^2,$$

where we used the following facts: (i) $\langle\boldsymbol{a}_i^*,\boldsymbol{x}\rangle$ is independent of the random variable $p_i^*(\boldsymbol{x})$ for each $i\in[k]$, (ii) $\langle\boldsymbol{a}_i^*,\boldsymbol{x}\rangle \overset{(d)}{=} \langle\boldsymbol{a}_1^*,\boldsymbol{x}\rangle$ and (iii) $\mathbb{E}[g(\langle\boldsymbol{a}_1^*,\boldsymbol{x}\rangle)^2] \leq 1$. Since $\mathbb{E}[(y-g(\boldsymbol{a}_i^\top\boldsymbol{x}))^2] \leq 2\mathbb{E}[y^2]+\mathbb{E}[g(\boldsymbol{a}_i^\top\boldsymbol{x})^2]$, after substituting these bounds in (26), we get

$$\mathbb{E}\left[\frac{-p_i(\boldsymbol{x})p_1(\boldsymbol{x})N_iN_1}{(\sum_i p_i(\boldsymbol{x})N_i)^2}\left(\frac{y-g(\boldsymbol{a}_i^\top\boldsymbol{x})}{\sigma^2}\right)g'(\boldsymbol{a}_i^\top\boldsymbol{x})\langle\boldsymbol{x},\Delta_i\rangle\langle\boldsymbol{x},\tilde{\Delta}\rangle\right] \leq \frac{\sqrt{6(2+\sigma^2)}}{4}\|\Delta_i/\sigma^2\|_2\|\tilde{\Delta}\|_2.$$

Similarly,

$$\mathbb{E}\left[\frac{p_i(\boldsymbol{x})N_ip_1(\boldsymbol{x})N_1}{(\sum_i p_i(\boldsymbol{x})N_i)^2}\left(\frac{y-g(\boldsymbol{a}_1^\top\boldsymbol{x})}{\sigma^2}\right)g'(\boldsymbol{a}_1^\top\boldsymbol{x})\langle\boldsymbol{x},\Delta_1\rangle\langle\boldsymbol{x},\tilde{\Delta}\rangle\right] \leq \frac{\sqrt{6(2+\sigma^2)}}{4}\|\Delta/\sigma^2\|_2\|\tilde{\Delta}\|_2.$$

Substituting the above two inequalities in (23), we obtain that

$$\mathbb{E}[(p^{(1)}(\boldsymbol{A},\boldsymbol{W})-p^{(1)}(\boldsymbol{A}^*,\boldsymbol{W}))\langle\boldsymbol{x},\tilde{\Delta}\rangle] \leq 2(k-1)\frac{\sqrt{6(2+\sigma^2)}}{4}\|\Delta_1/\sigma^2\|_2\|\tilde{\Delta}\|_2.$$

Defining $\kappa \triangleq (k-1)\frac{\sqrt{6(2+\sigma^2)}}{2}$ and using the fact that $\left\|\Delta/\sigma^2\right\|_2 \leq \varepsilon_1$, we thus obtain

$$\left\|\mathbb{E}[(p^{(1)}(\boldsymbol{A},\boldsymbol{W})-p^{(1)}(\boldsymbol{A}^*,\boldsymbol{W}))\boldsymbol{x}]\right\|_2 \leq \kappa\varepsilon_1.$$

F. Proof of Lemma 3

F.1. Proof for $k=2$

Proof. We first prove the lemma for $k=2$. We show that the assumptions in Appendix B hold *globally* in our setting yielding a geometric convergence. Here we simply denote $M(\boldsymbol{W},\boldsymbol{A}^*)$ as $M(\boldsymbol{w})$ dropping the explicit dependence on $\boldsymbol{A}^*$. Recall that

$$Q(\boldsymbol{w}|\boldsymbol{w}_t) = \mathbb{E}_{p_{\boldsymbol{w}^*}(\boldsymbol{x},y)}\left[p_1(\boldsymbol{x},y,\boldsymbol{w}_t)\cdot(\boldsymbol{w}^\top\boldsymbol{x})-\log(1+e^{\boldsymbol{w}^\top\boldsymbol{x}})\right],$$

where

$$p_1(\boldsymbol{x},y,\boldsymbol{w}_t) = \frac{f(\boldsymbol{w}_t^\top\boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_1^\top\boldsymbol{x}),\sigma^2)}{f(\boldsymbol{w}^\top\boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_1^\top\boldsymbol{x}),\sigma^2)+(1-f(\boldsymbol{w}^\top\boldsymbol{x}))\mathcal{N}(y|g(\boldsymbol{a}_2^\top\boldsymbol{x}),\sigma^2)}. \qquad (27)$$

For simplicity we drop the subscript in the above expectation with respect to the distribution $p_{\boldsymbol{w}^*}(\boldsymbol{x},y)$. Now we verify each of the assumptions.

- Convexity of Ω easily follows from its definition.

- We have that

$$Q(\boldsymbol{w}|\boldsymbol{w}^*) = \mathbb{E}\left[p_1(\boldsymbol{x}, y, \boldsymbol{w}^*) \cdot (\boldsymbol{w}^\top \boldsymbol{x}) - \log(1 + e^{\boldsymbol{w}^\top \boldsymbol{x})}\right].$$

  Note that the strong-concavity of $Q(\cdot|\boldsymbol{w}^*)$ is equivalent to the strong-convexity of $-Q(\cdot|\boldsymbol{w}^*)$. Denoting the sigmoid function by f, we have that for all $\boldsymbol{w} \in \Omega$,

$$\begin{aligned} -\nabla^2 Q(\boldsymbol{w}|\boldsymbol{w}^*) &= \mathbb{E}\left[f'(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right], \\ &\overset{\text{(Stein's lemma)}}{=} \mathbb{E}\left[f'''(\boldsymbol{w}^\top \boldsymbol{x})\right] \cdot \boldsymbol{w}\boldsymbol{w}^\top + \mathbb{E}[f'(\boldsymbol{w}^\top \boldsymbol{x})] \cdot I \\ &= \mathbb{E}[f'''(\|\boldsymbol{w}\| \, Z)] \cdot \boldsymbol{w}\boldsymbol{w}^\top + \mathbb{E}[f'(\|\boldsymbol{w}\| \, Z)] \cdot I, \quad Z \sim \mathcal{N}(0,1) \\ &\overset{(a)}{\succcurlyeq} \inf_{0 \le \alpha \le 1} \min\left\{\mathbb{E}[f'(\alpha Z)], \mathbb{E}[f'(\alpha Z)] + \alpha^2 \mathbb{E}[f'''(\alpha Z)]\right\} \cdot I \\ &= \underbrace{0.14}_{\lambda} \cdot I \end{aligned} \tag{28}$$

  where (a) follows from finding the two possible eigenvalues of the positive-definite matrix in the previous step and considering the minimum among them to ensure strong-convexity. Here the value of λ is found numerically to be approximately around 0.1442.

- For any $\boldsymbol{w}, \boldsymbol{w}_t \in \Omega$,

$$\nabla Q(\boldsymbol{w}|\boldsymbol{w}_t) = \mathbb{E}\left[p_1(\boldsymbol{x}, y, \boldsymbol{w}_t) \cdot \boldsymbol{x} - f(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{x}\right].$$

  Thus,

$$\|\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}^*) - \nabla Q(M(\boldsymbol{w})|\boldsymbol{w})\| = \|\mathbb{E}\left[(p_1(\boldsymbol{x}, y, \boldsymbol{w}_t) - p_1(\boldsymbol{x}, y, \boldsymbol{w}^*) \cdot \boldsymbol{x}\right]\| \overset{(a)}{\le} \gamma_\sigma \|\boldsymbol{w} - \boldsymbol{w}^*\|,$$

  where we want to prove in (a) that γ_σ is smaller than 0.14 for all $\boldsymbol{w} \in \Omega$. Intuitively, this means that the posterior probability in (27) is smooth with respect to the parameter $\boldsymbol{w}$. We will now show that this can be achieved in the high-SNR regime when σ is sufficiently small. This will ensure that $\kappa_\sigma \triangleq \frac{\gamma_\sigma}{\lambda} < 1$. In particular, the value of γ_σ is dimension-independent and depends only on the choice of the non-linearity g.

To prove that

$$\|\mathbb{E}\left[(p_1(\boldsymbol{x}, y, \boldsymbol{w}) - p_1(\boldsymbol{x}, y, \boldsymbol{w}^*)) \cdot \boldsymbol{x}\right]\| \le \gamma \|\boldsymbol{w} - \boldsymbol{w}^*\| = \gamma \|\Delta\|,$$

it suffices to show

$$\langle \mathbb{E}\left[(p_1(\boldsymbol{x}, y, \boldsymbol{w}) - p_1(\boldsymbol{x}, y, \boldsymbol{w}^*)) \cdot \boldsymbol{x}\right], \tilde{\Delta} \rangle \le \gamma \|\Delta\| \|\tilde{\Delta}\|, \quad \forall \tilde{\Delta} \in \mathbb{R}^d.$$

Or equivalently,

$$\mathbb{E}\left[(p_1(\boldsymbol{x}, y, \boldsymbol{w}) - p_1(\boldsymbol{x}, y, \boldsymbol{w}^*)) \langle \boldsymbol{x}, \tilde{\Delta} \rangle\right] \le \gamma \|\Delta\| \|\tilde{\Delta}\|.$$

Let $\Delta \triangleq \boldsymbol{w} - \boldsymbol{w}^*$ and $f(u) \triangleq p_1(\boldsymbol{x}, y, \boldsymbol{w}_u)$ where $\boldsymbol{w}_u = \boldsymbol{w}^* + u\Delta, u \in [0,1]$. Thus $f(1) = p_1(\boldsymbol{x}, y, \boldsymbol{w})$ and $f(0) = p_1(\boldsymbol{x}, y, \boldsymbol{w}^*)$. So we get

$$p_1(\boldsymbol{x}, y, \boldsymbol{w}) - p_1(\boldsymbol{x}, y, \boldsymbol{w}^*) = f(1) - f(0) = \int_0^1 f'(u) du = \int_0^1 \langle \nabla p_1(\boldsymbol{x}, y, \boldsymbol{w}_u), \Delta \rangle du,$$

where the gradient is evaluated with respect to $\boldsymbol{w}_u$. Differentiating (27) with respect to $\boldsymbol{w}$, we get that

$$\begin{aligned} \nabla_{\boldsymbol{w}} p_1(\boldsymbol{x}, y, \boldsymbol{w}) &= \frac{f(\boldsymbol{w}^\top \boldsymbol{x})(1 - f(\boldsymbol{w}^\top \boldsymbol{x}))\mathcal{N}(y|g(\boldsymbol{a}_1^\top \boldsymbol{x}), \sigma^2)\mathcal{N}(y|g(\boldsymbol{a}_2^\top \boldsymbol{x}), \sigma^2)}{(f(\boldsymbol{w}^\top \boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_1^\top \boldsymbol{x}), \sigma^2) + (1 - f(\boldsymbol{w}^\top \boldsymbol{x}))\mathcal{N}(y|g(\boldsymbol{a}_2^\top \boldsymbol{x}), \sigma^2))^2} \cdot \boldsymbol{x} \\ &\triangleq R(\boldsymbol{x}, y, \boldsymbol{w}, \sigma) \cdot \boldsymbol{x}. \end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[(p_1(\boldsymbol{x},y,\boldsymbol{w})-p_1(\boldsymbol{x},y,\boldsymbol{w}^*))\,\langle \boldsymbol{x},\tilde{\Delta}\rangle\right] &= \mathbb{E}\left[\left(\int_0^1 R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma)\langle \boldsymbol{x},\Delta\rangle du\right)\langle \boldsymbol{x},\tilde{\Delta}\rangle\right]\\
&= \int_0^1 \mathbb{E}\left[R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma)\langle \boldsymbol{x},\Delta\rangle\langle \boldsymbol{x},\tilde{\Delta}\rangle\right] du\\
&\le \left(\int_0^1 \sqrt{\mathbb{E}[R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma)^2]}du\right)\sqrt{\mathbb{E}\left[\langle \boldsymbol{x},\Delta\rangle^2\langle \boldsymbol{x},\tilde{\Delta}\rangle^2\right]}\\
&\le \underbrace{\sqrt{3}\left(\int_0^1 \sqrt{\mathbb{E}[R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma)^2]}du\right)}_{\gamma_\sigma}\|\Delta\|\,\|\tilde{\Delta}\|\\
&= \gamma_\sigma\,\|\Delta\|\,\|\tilde{\Delta}\|,
\end{aligned}$$

where the last inequality follows from Lemma 5 of (Balakrishnan et al., 2017). Our goal is to now prove that $\gamma_\sigma \to 0$ as $\sigma \to 0$. First observe that

$$\begin{aligned}
R(\boldsymbol{x},y,\boldsymbol{w},\sigma) &= \frac{f(\boldsymbol{w}^\top\boldsymbol{x})(1-f(\boldsymbol{w}^\top\boldsymbol{x})e^{-(y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))/2\sigma^2}e^{-(y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))/2\sigma^2}}{(f(\boldsymbol{w}^\top\boldsymbol{x})e^{-(y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))/2\sigma^2}+(1-f(\boldsymbol{w}^\top\boldsymbol{x}))e^{-(y-g(\boldsymbol{a}_2^\top\boldsymbol{x}))/2\sigma^2})^2} \le \frac{1}{4}(\text{ since } \frac{ab}{(a+b)^2}\le 1/4)\\
&= \frac{f(1-f)e^{\frac{(y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))^2-(y-g(\boldsymbol{a}_2^\top\boldsymbol{x}))^2}{2\sigma^2}}}{\left(f+(1-f)e^{\frac{(y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))^2-(y-g(\boldsymbol{a}_2^\top\boldsymbol{x}))^2}{2\sigma^2}}\right)^2} \to 0 \text{ as } \sigma\to 0,
\end{aligned}$$

where the key observation is that irrespective of the sign of $(y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))^2-(y-g(\boldsymbol{a}_2^\top\boldsymbol{x}))^2$, the ratio still goes to zero and hence by dominated convergence theorem $\mathbb{E}[R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma)^2]\to 0$ for each $u\in[0,1]$. Now we show that this convergence is uniform in u and thus $\gamma_\sigma\to 0$. For simplicity, define

$$\Delta_1 \triangleq (y-g(\boldsymbol{a}_1^\top\boldsymbol{x}))^2,\quad \Delta_2 \triangleq (y-g(\boldsymbol{a}_2^\top\boldsymbol{x}))^2 \text{ and } \sigma=\frac{1}{n}. \tag{29}$$

Thus,

$$\begin{aligned}
R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma) &= \frac{f(1-f)e^{\frac{n^2}{2}(\Delta_1-\Delta_2)}}{\left(f+(1-f)e^{\frac{n^2}{2}(\Delta_1-\Delta_2)}\right)^2} &(30)\\
&\le \frac{f(1-f)e^{\frac{n^2}{2}(\Delta_1-\Delta_2)}}{\left((1-f)e^{\frac{n^2}{2}(\Delta_1-\Delta_2)}\right)^2} = \frac{f}{1-f}e^{-\frac{n^2}{2}(\Delta_1-\Delta_2)}. &(31)
\end{aligned}$$

Similarly,

$$R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma) \le \frac{1-f}{f}e^{-\frac{n^2}{2}(\Delta_2-\Delta_1)}. \tag{32}$$

Thus, we get

$$R(\boldsymbol{x},y,\boldsymbol{w}_u,\sigma) \le \max\left(\frac{1-f}{f},\frac{f}{1-f}\right)e^{-\frac{n^2}{2}(|\Delta_1-\Delta_2|)}. \tag{33}$$

Hence

$$\begin{aligned}
\frac{\gamma_\sigma}{\sqrt{3}} &= \int_0^1 \sqrt{\mathbb{E}[\mathrm{Ratio}(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2]}du && (34)\\
&\leq \int_0^1 \sqrt{\mathbb{E}\left[\max\left(\frac{1-f}{f}, \frac{f}{1-f}\right)^2 e^{-n^2|\Delta_1 - \Delta_2|}\right]}du && (35)\\
&\leq \int_0^1 \sqrt{\mathbb{E}\left[\left(\frac{1-f}{f}\right)^2 e^{-n^2|\Delta_1-\Delta_2|} + \left(\frac{f}{1-f}\right)^2 e^{-n^2|\Delta_1-\Delta_2|}\right]}du && (36)\\
&= \int_0^1 \sqrt{2\mathbb{E}\left[e^{2\boldsymbol{w}_u^\top \boldsymbol{x}} e^{-n^2|\Delta_1-\Delta_2|}\right]}du && (37)\\
&\leq \int_0^1 \sqrt{2\sqrt{\mathbb{E}[e^{4\boldsymbol{w}_u^\top \boldsymbol{x}}]\mathbb{E}[e^{-2n^2|\Delta_1-\Delta_2|}]}}du && (38)\\
&\overset{(a)}{\leq} \sqrt{2e^4\sqrt{\mathbb{E}[e^{-2n^2|\Delta_1-\Delta_2|}]}}, && (39)
\end{aligned}$$

where (a) follows from the fact $\|\boldsymbol{w}_u\| \leq 1$ and $\mathbb{E}[e^{4\boldsymbol{w}_u^\top \boldsymbol{x}}] = e^{8\|\boldsymbol{w}_u\|^2} \leq e^8$, for each $u \in [0,1]$. Now we analyze the convergence rate of the last term $\mathbb{E}[e^{-2n^2|\Delta_1-\Delta_2|}]$ for the case of linear regression, i.e. $g(z) = z$. Notice that for the two-mixtures, we have

$$y \overset{(d)}{=} Z(\boldsymbol{a}_1^\top \boldsymbol{x}) + (1-Z)\boldsymbol{a}_2^\top \boldsymbol{x} + \sigma N = Z(\boldsymbol{a}_1^\top \boldsymbol{x}) + (1-Z)\boldsymbol{a}_2^\top \boldsymbol{x} + \frac{N}{n}, \quad Z|\boldsymbol{x} \sim \mathrm{Bern}(f(\boldsymbol{w}_*^\top \boldsymbol{x})). \tag{40}$$

Thus,

$$\begin{aligned}
\Delta_1 - \Delta_2 &\overset{(d)}{=} (y - \boldsymbol{a}_1^\top \boldsymbol{x})^2 - (y - \boldsymbol{a}_2^\top \boldsymbol{x})^2 && (41)\\
&= (\boldsymbol{a}_1^\top \boldsymbol{x} - \boldsymbol{a}_2^\top \boldsymbol{x})^2(1-2Z) + \frac{2N}{n}(\boldsymbol{a}_2^\top \boldsymbol{x} - \boldsymbol{a}_1^\top \boldsymbol{x}) && (42)\\
&= \langle \boldsymbol{x}, \boldsymbol{v}\rangle^2(1-2Z) + \frac{2N}{n}\langle \boldsymbol{x}, \boldsymbol{v}\rangle, \quad \boldsymbol{v} = \boldsymbol{a}_1 - \boldsymbol{a}_2. && (43)
\end{aligned}$$

Since Z can equal either 0 or 1, we have

$$\begin{aligned}
\gamma_\sigma &\leq \sqrt{3}\sqrt{2e^4}\left(\mathbb{E}[e^{-2n^2\left|\langle \boldsymbol{x},\boldsymbol{v}\rangle^2(1-2Z) + \frac{2N}{n}\langle \boldsymbol{x},\boldsymbol{v}\rangle\right|}]\right)^{1/4} && (44)\\
&\leq \sqrt{6e^4}\left(\mathbb{E}\left[\max\left(e^{-2n^2\left|\langle \boldsymbol{x},\boldsymbol{v}\rangle^2 + \frac{2N}{n}\langle \boldsymbol{x},\boldsymbol{v}\rangle\right|}, e^{-2n^2\left|-\langle \boldsymbol{x},\boldsymbol{v}\rangle^2 + \frac{2N}{n}\langle \boldsymbol{x},\boldsymbol{v}\rangle\right|}\right)\right]\right)^{1/4} && (45)\\
&\leq \sqrt{6\sqrt{2}e^4}\left(\mathbb{E}\left[e^{-2n^2|\langle \boldsymbol{x},\boldsymbol{v}\rangle^2 + \frac{2N}{n}\langle \boldsymbol{x},\boldsymbol{v}\rangle|}\right]\right)^{1/4} && (46)\\
&= \sqrt{6\sqrt{2}e^4}\left(\mathbb{E}\left[e^{-2n^2|Z^2 + \frac{2ZN}{n}|}\right]\right)^{1/4}, \quad Z \sim \mathcal{N}(0, \|\boldsymbol{a}_1 - \boldsymbol{a}_2\|), N \sim \mathcal{N}(0,1). && (47)\\
&= O\left(\sqrt{6\sqrt{2}e^4}\left(\mathbb{E}[e^{-2n^2Z^2}]\right)^{1/4}\right) && (48)\\
&= \sqrt{6\sqrt{2}e^4}\left(\sqrt{\frac{1}{4n^2\left\|\boldsymbol{a}_1 - \boldsymbol{a}_2\right\|^2 + 1}}\right)^{1/4} && (49)\\
&= O\left(\frac{1}{(n\left\|\boldsymbol{a}_1 - \boldsymbol{a}_2\right\|)^{1/4}}\right) && (50)\\
&= O\left(\left(\frac{\sigma}{\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|}\right)^{1/4}\right). && (51)
\end{aligned}$$

□

F.2. Proof for general k

Proof. The proof strategy for general k is similar. First let $\varepsilon_1 = 0$. Our task is to show that the assumptions of Appendix B hold globally in our setting. The domain Ω is clearly convex since

$$\Omega = \{\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{k-1}) : \|\boldsymbol{w}_i\| \leq 1, \forall i \in [k-1]\}.$$

Now we verify Assumption 2. The function $Q(.|\boldsymbol{w}_t)$ is given by

$$Q(\boldsymbol{w}|\boldsymbol{w}_t) = \mathbb{E}\left[\sum_{i\in[k-1]} p^{(i)}_{\boldsymbol{w}_t}(\boldsymbol{w}_i^\top \boldsymbol{x}) - \log\left(1 + \sum_{i\in[k-1]} e^{\boldsymbol{w}_i^\top \boldsymbol{x}}\right)\right],$$

where $p^{(i)}_{\boldsymbol{w}_t} \triangleq \mathbb{P}\left[z = i|\boldsymbol{x}, y, \boldsymbol{w}_t\right]$ corresponds to the posterior probability for the i^{th} expert, given by

$$p^{(i)}_{\boldsymbol{w}_t} = \frac{p_{i,t}(\boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_i^\top \boldsymbol{x}), \sigma^2)}{\sum_{j\in[k]} p_{j,t}(\boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_j^\top \boldsymbol{x}), \sigma^2)}, \quad p_{i,t}(\boldsymbol{x}) = \frac{e^{(\boldsymbol{w}_t)_i^\top \boldsymbol{x}}}{1 + \sum_{j\in[k-1]} e^{(\boldsymbol{w}_t)_j^\top \boldsymbol{x}}}.$$

Throughout we follow the convention that $\boldsymbol{w}_k = 0$. Thus the gradient of Q with respect to the i^{th} gating parameter $\boldsymbol{w}_i$ is given by

$$\nabla_{\boldsymbol{w}_i} Q(\boldsymbol{w}|\boldsymbol{w}_t) = \mathbb{E}\left[\left(p^{(i)}_{\boldsymbol{w}_t} - \frac{e^{\boldsymbol{w}_i^\top \boldsymbol{x}}}{1 + \sum_{j\in[k-1]} e^{\boldsymbol{w}_j^\top \boldsymbol{x}}}\right) \cdot \boldsymbol{x}\right], \quad i \in [k-1].$$

Thus the $(i,j)^{\text{th}}$ block of the negative Hessian $-\nabla^{(2)}_{\boldsymbol{w}} Q(\boldsymbol{w}|\boldsymbol{w}^*) \in \mathbb{R}^{d(k-1)\times d(k-1)}$ is given by

$$-\nabla_{\boldsymbol{w}_i,\boldsymbol{w}_j} Q(\boldsymbol{w}|\boldsymbol{w}^*) = \begin{cases} \mathbb{E}[p_i(\boldsymbol{x})(1 - p_i(\boldsymbol{x})) \cdot \boldsymbol{x}\boldsymbol{x}^\top], & j = i \\ \mathbb{E}[-p_i(\boldsymbol{x})p_j(\boldsymbol{x}) \cdot \boldsymbol{x}\boldsymbol{x}^\top], & j \neq i \end{cases}, \tag{52}$$

where $p_i(\boldsymbol{x}) = \frac{e^{\boldsymbol{w}_i^\top \boldsymbol{x}}}{1+\sum_{j\in[k-1]} e^{\boldsymbol{w}_j^\top \boldsymbol{x}}}$. It is clear from (52) that $-\nabla^{(2)}_{\boldsymbol{w}} Q(\boldsymbol{w}|\boldsymbol{w}^*)$ is positive semi-definite. Since we are interested in the strong convexity of $-Q(\cdot|\boldsymbol{w}^*)$ which is equivalent to positive definiteness of the negative Hessian, it suffices to show that

$$\lambda \triangleq \inf_{w\in\Omega} \lambda_{\min}\left(-\nabla^{(2)}_{\boldsymbol{w}} Q(\boldsymbol{w}|\boldsymbol{w}^*)\right) > 0.$$

Since the Hessian is continuous with respect to $\boldsymbol{w}$ and consequently the minimum eigenvalue of it, there exists a $\boldsymbol{w}' \in \Omega$ such that

$$\lambda = \lambda_{\min}\left(-\nabla^{(2)}_{\boldsymbol{w}'} Q(\boldsymbol{w}'|\boldsymbol{w}^*)\right) = \inf_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^\top \left(-\nabla^{(2)}_{\boldsymbol{w}'} Q(\boldsymbol{w}'|\boldsymbol{w}^*)\right)\boldsymbol{a},$$

where $\boldsymbol{a} = (\boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_{k-1}^\top)^\top \in \mathbb{R}^{d(k-1)}$. In view of (52), the above equation can be further simplified to

$$\lambda = \inf_{\|\boldsymbol{a}\|=1} \mathbb{E}[\boldsymbol{a}_{\boldsymbol{x}}^\top M_{\boldsymbol{x}} \boldsymbol{a}_{\boldsymbol{x}}], \tag{53}$$

where $\boldsymbol{a}_{\boldsymbol{x}} = (\boldsymbol{a}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{a}_{k-1}^\top \boldsymbol{x})^\top \in \mathbb{R}^{k-1}$ and $M_{\boldsymbol{x}}$ is given by

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

Let the infimum in (53) is attained by $\boldsymbol{a}^*$, i.e. $\lambda = \mathbb{E}[(\boldsymbol{a}^*_{\boldsymbol{x}})^\top M_{\boldsymbol{x}} \boldsymbol{a}^*_{\boldsymbol{x}}]$. For each $\boldsymbol{x}$, $M_{\boldsymbol{x}}$ is strictly diagonally dominant since $|M_{\boldsymbol{x}}(i,i)| = p_i(\boldsymbol{x})(1 - p_i(\boldsymbol{x})) = p_i(\boldsymbol{x})\left(\sum_{j\neq i, j\in[k]} p_j(\boldsymbol{x})\right) > p_i(\boldsymbol{x})\left(\sum_{j\neq i, j\in[k-1]} p_j(\boldsymbol{x})\right) = \sum_{j\neq i} M(i,j)$. Thus $M_{\boldsymbol{x}}$ is positive-definite and $(\boldsymbol{a}^*_{\boldsymbol{x}})^\top M_{\boldsymbol{x}} \boldsymbol{a}^*_{\boldsymbol{x}} > 0$ whenever $\boldsymbol{a}^*_{\boldsymbol{x}} \neq 0$. Since x follows a continuous distribution it follows that $\boldsymbol{a}^*_{\boldsymbol{x}} \neq 0$ with probability 1 and thus $\lambda = \mathbb{E}[(\boldsymbol{a}^*_{\boldsymbol{x}})^\top M_{\boldsymbol{x}} \boldsymbol{a}^*_{\boldsymbol{x}}] > 0$.

Now it remains to show that Assumption 3 too holds, i.e.

$$\|\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}^*) - \nabla Q(M(\boldsymbol{w})|\boldsymbol{w})\| \leq \gamma \|\boldsymbol{w} - \boldsymbol{w}^*\|.$$

Note that $\boldsymbol{w} = (\boldsymbol{w}_1^\top, \ldots, \boldsymbol{w}_{k-1}^\top)^\top \in \mathbb{R}^{d(k-1)}$. We will show that

$$\|(\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}^*))_i - (\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}))_i\| \leq \gamma_\sigma \|\boldsymbol{w} - \boldsymbol{w}^*\|, \quad i \in [k-1],$$

where $(\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}))_i \in \mathbb{R}^d$ refers to the i^{th} block of the gradient and $\gamma_\sigma \to 0$. Observe that

$$(\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}^*))_i - (\nabla Q(M(\boldsymbol{w})|\boldsymbol{w}))_i = \mathbb{E}\left[(p_{\boldsymbol{w}}^{(i)} - p_{\boldsymbol{w}^*}^{(i)}) \cdot \boldsymbol{x}\right]$$

Let $\Delta = \boldsymbol{w} - \boldsymbol{w}^*$ and correspondingly $\Delta = (\Delta_1^\top, \ldots, \Delta_{k-1}^\top)^\top$ where $\Delta_i = \boldsymbol{w}_i - \boldsymbol{w}_i^*$. Thus it suffices to show that

$$\left\|\mathbb{E}[(p_{\boldsymbol{w}}^{(i)} - p_{\boldsymbol{w}^*}^{(i)}) \cdot \boldsymbol{x}]\right\| \leq \gamma_\sigma \|\Delta\|.$$

Or equivalently,

$$\mathbb{E}[(p_{\boldsymbol{w}}^{(i)} - p_{\boldsymbol{w}^*}^{(i)})\langle \boldsymbol{x}, \tilde{\Delta}\rangle] \leq \gamma_\sigma \|\Delta\| \|\tilde{\Delta}\|, \quad \forall \tilde{\Delta} \in \mathbb{R}^d.$$

We consider the case $i = 1$. The proof for the other cases is similar. Recall that

$$p_{\boldsymbol{w}}^{(1)} = \frac{p_1(\boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_1^\top \boldsymbol{x}), \sigma^2)}{\sum_{j\in[k]} p_j(\boldsymbol{x})\mathcal{N}(y|g(\boldsymbol{a}_j^\top \boldsymbol{x}), \sigma^2)}, \quad p_i(\boldsymbol{x}) = \frac{e^{\boldsymbol{w}_i^\top \boldsymbol{x}}}{1 + \sum_{j\in[k-1]} e^{\boldsymbol{w}_j^\top \boldsymbol{x}}}, \quad i \in [k-1].$$

For simplicity we define $N_i = \mathcal{N}(y|g(\boldsymbol{a}_1^\top \boldsymbol{x}), \sigma^2)$. It is straightforward to verify that

$$\nabla_{\boldsymbol{w}_j} p_i(\boldsymbol{x}) = \begin{cases} p_i(\boldsymbol{x})(1 - p_i(\boldsymbol{x})) \cdot \boldsymbol{x}, & j = i \\ -p_i(\boldsymbol{x})p_j(\boldsymbol{x}) \cdot \boldsymbol{x}, & j \neq i \end{cases}$$

Thus

$$\begin{aligned}
\nabla_{\boldsymbol{w}_1}(p_{\boldsymbol{w}}^{(1)}) &= \nabla_{\boldsymbol{w}_1}\left(\frac{p_1(\boldsymbol{x})N_1}{\sum_{i=1}^N p_i(\boldsymbol{x})N_i}\right) \\
&= \frac{\left(\sum_{i=1}^N p_i(\boldsymbol{x})N_i\right) p_1(\boldsymbol{x})(1 - p_1(\boldsymbol{x}))N_1 - p_1(\boldsymbol{x})N_1\left(-\sum_{j\neq 1} p_j(\boldsymbol{x})p_1(\boldsymbol{x})N_j + p_1(\boldsymbol{x})(1 - p_1(\boldsymbol{x}))N_1\right)}{\left(\sum_{i=1}^N p_i(\boldsymbol{x})N_i\right)^2} \cdot \boldsymbol{x} \\
&= \frac{p_1(\boldsymbol{x})N_1\left(\sum_{j\geq 2} p_j(\boldsymbol{x})N_j\right)}{\left(\sum_{i=1}^N p_i(\boldsymbol{x})N_i\right)^2} \cdot \boldsymbol{x} \\
&\triangleq R_1(\boldsymbol{x}, y, \boldsymbol{w}, \sigma) \cdot \boldsymbol{x}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\nabla_{\boldsymbol{w}_i}(p_{\boldsymbol{w}}^{(1)}) &= \frac{p_1(\boldsymbol{x})p_i(\boldsymbol{x})N_1N_i}{\left(\sum_{i=1}^N p_i(\boldsymbol{x})N_i\right)^2} \cdot \boldsymbol{x}, \quad i \neq 1, \\
&\triangleq R_i(\boldsymbol{x}, y, \boldsymbol{w}, \sigma) \cdot \boldsymbol{x}.
\end{aligned}$$

Let $\boldsymbol{w}_u \triangleq \boldsymbol{w}^* + u\Delta, u \in [0, 1]$ and $f(u) \triangleq p_{\boldsymbol{w}_u}^{(1)}$. Thus

$$\begin{aligned}
p_{\boldsymbol{w}}^{(1)} - p_{\boldsymbol{w}^*}^{(1)} = f(1) - f(0) &= \int_0^1 f'(u)du \\
&= \int_0^1 \left(\sum_{i\in[k-1]} \langle \nabla_{\boldsymbol{w}_i}(p_{\boldsymbol{w}_u}^{(1)}), \Delta_i\rangle\right) du \\
&= \sum_{i\in[k-1]} \int_0^1 R_i(\boldsymbol{x}, y, \boldsymbol{w}, \sigma)\langle \boldsymbol{x}, \Delta_i\rangle du.
\end{aligned}$$

So we get

$$\begin{aligned}
\mathbb{E}[(p^{(1)}_{\boldsymbol{w}} - p^{(1)}_{\boldsymbol{w}^*})\langle \boldsymbol{x}, \tilde{\Delta}\rangle] &= \sum_{i\in[k-1]} \int_0^1 \mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)\langle \boldsymbol{x}, \Delta_i\rangle \langle \boldsymbol{x}, \tilde{\Delta}\rangle] du \\
&\leq \sum_{i\in[k-1]} \int_0^1 \sqrt{\mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2]\mathbb{E}[\langle \boldsymbol{x}, \Delta_i\rangle^2 \langle \boldsymbol{x}, \tilde{\Delta}\rangle^2]} du \\
&\leq \sum_{i\in[k-1]} \int_0^1 \sqrt{\mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2]} \left(\sqrt{3}\, \|\Delta_i\|\, \|\tilde{\Delta}\|\right) du \\
&\leq \sum_{i\in[k-1]} \int_0^1 \sqrt{\mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2]} \left(\sqrt{3}\, \|\Delta\|\, \|\tilde{\Delta}\|\right) du \\
&= \underbrace{\left(\sum_{i\in[k-1]} \int_0^1 \sqrt{\mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2]} du\right)}_{\gamma^{(1)}_\sigma} \left(\sqrt{3}\, \|\Delta\|\, \|\tilde{\Delta}\|\right)
\end{aligned}$$

Now our goal is to show that $\mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2] \to 0$ as $\sigma \to 0$. For $i = 1$, we have

$$R_1(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2 = \left(\frac{\sum_{j\geq 2} p_1(\boldsymbol{x})p_j(\boldsymbol{x})N_1 N_j}{\left(\sum_{i=1}^N p_i(\boldsymbol{x})N_i\right)^2}\right)^2 \leq k \sum_{j\geq 2} \left(\frac{p_1(\boldsymbol{x})p_j(\boldsymbol{x})N_1 N_j}{\left(\sum_{i=1}^N p_i(\boldsymbol{x})N_i\right)^2}\right)^2 \leq k \sum_{j\geq 2} \left(\frac{p_1(\boldsymbol{x})p_j(\boldsymbol{x})N_1 N_j}{(p_1(\boldsymbol{x})N_1 + p_j(\boldsymbol{x})N_j)^2}\right)^2$$

Similarly,

$$R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2 \leq \left(\frac{p_1(\boldsymbol{x})p_i(\boldsymbol{x})N_1 N_i}{(p_1(\boldsymbol{x})N_1 + p_i(\boldsymbol{x})N_i)^2}\right)^2, \quad \forall i \neq 1, i \in [k-1].$$

For $\boldsymbol{w} = \boldsymbol{w}_u$ and $i \neq 1$, we have that

$$\begin{aligned}
\frac{p_1(\boldsymbol{x})p_i(\boldsymbol{x})N_1 N_i}{(p_1(\boldsymbol{x})N_1 + p_i(\boldsymbol{x})N_i)^2} &= \frac{e^{\boldsymbol{w}_1^\top \boldsymbol{x}} e^{\boldsymbol{w}_i^\top \boldsymbol{x}} e^{-\frac{(y-g(\boldsymbol{a}_1^\top \boldsymbol{x}))^2}{2\sigma^2}} e^{-\frac{(y-g(\boldsymbol{a}_i^\top \boldsymbol{x}))^2}{2\sigma^2}}}{\left(e^{\boldsymbol{w}_1^\top \boldsymbol{x}} e^{-\frac{(y-g(\boldsymbol{a}_1^\top \boldsymbol{x}))^2}{2\sigma^2}} + e^{\boldsymbol{w}_i^\top \boldsymbol{x}} e^{-\frac{(y-g(\boldsymbol{a}_i^\top \boldsymbol{x}))^2}{2\sigma^2}}\right)^2} \leq \frac{1}{4} \\
&= \frac{e^{\boldsymbol{w}_1^\top \boldsymbol{x}} e^{\boldsymbol{w}_i^\top \boldsymbol{x}} e^{\frac{(y-g(\boldsymbol{a}_1^\top \boldsymbol{x}))^2 - (y-g(\boldsymbol{a}_i^\top \boldsymbol{x}))^2}{2\sigma^2}}}{\left(e^{\boldsymbol{w}_1^\top \boldsymbol{x}} + e^{\boldsymbol{w}_i^\top \boldsymbol{x}} e^{\frac{(y-g(\boldsymbol{a}_1^\top \boldsymbol{x}))^2 - (y-g(\boldsymbol{a}_i^\top \boldsymbol{x}))^2}{2\sigma^2}}\right)^2} \\
&\xrightarrow{\sigma \to 0} 0.
\end{aligned}$$

Thus, by Dominated Convergence Theorem, $\mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2] \to 0$ for each $u \in [0, 1]$. To show that $\int_0^1 \mathbb{E}[R_i(\boldsymbol{x}, y, \boldsymbol{w}_u, \sigma)^2] du \to 0$, we can now follow the same analysis as in the proof of Theorem 2 from (29) on-wards (replacing $\boldsymbol{w}$ there with $\boldsymbol{w}_1 - \boldsymbol{w}_i$) which ensures that $\gamma^{(1)}_\sigma$ in our case converges to zero. Similarly for other $i \in [k-1]$, we get that $\gamma^{(i)} \to 0$. Taking $\gamma_\sigma = \gamma^{(1)}_\sigma + \ldots + \gamma^{(k-1)}_\sigma$ and $\kappa_\sigma = \frac{\gamma_\sigma}{\lambda}$ completes the proof.

□

G. Gradient EM algorithm

In this section, we provide the convergence guarantees for the gradient EM algorithm. For simplicity, we prove the results for $k = 2$ and $(\boldsymbol{a}_1, \boldsymbol{a}_2) = (\boldsymbol{a}_1^*, \boldsymbol{a}_2^*)$. Thus we want to learn the gating parameter $\boldsymbol{w}^*$ in this setting. The results for the general case follow essentially the same proof as that of Theorem 2. In particular, our Theorem 5 can be viewed as a generalization of Lemma 3. Together with Lemma 4, extension to general k is straightforward.

Note that in the M-step of the EM algorithm, instead of maximizing $Q(\cdot|\boldsymbol{w}_t)$, we can chose an iterate so that it increases the Q value instead of fully maximizing it, i.e. $Q(\boldsymbol{w}_{t+1}|\boldsymbol{w}_t) \geq Q(\boldsymbol{w}_t|\boldsymbol{w}_t)$. Such a procedure is termed as *generalized EM*. *Gradient EM* is an example of generalized EM in which we take an ascent step in the direction of the gradient of $Q(\cdot|\boldsymbol{w}_t)$ to produce the next iterate, i.e.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha \nabla Q(\boldsymbol{w}_t|\boldsymbol{w}_t),$$

where $\alpha > 0$ is a suitably chosen step size and the gradient is with respect to the first argument. To account for the constrained optimization, we can include a projection step. Mathematically,

$$\boldsymbol{w}_{t+1} = G(\boldsymbol{w}_t), \quad G(\boldsymbol{w}) = \Pi_\Omega(\boldsymbol{w} + \alpha \nabla Q(\boldsymbol{w}|)\boldsymbol{w}),$$

where Π_Ω refers to the projection operator. Our next result establishes that the iterates of the gradient EM algorithm too converge geometrically for an appropriately chosen step size α.

Theorem 5. *Suppose that the domain $\Omega = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \leq 1\}$ and $(\boldsymbol{a}_1, \boldsymbol{a}_2) = (\boldsymbol{a}_1^*, \boldsymbol{a}_2^*)$. Then there exist constants $\alpha_0 > 0$ and $\sigma_0 > 0$ such that for any step size $0 < \alpha \leq \alpha_0$ and noise variance $\sigma < \sigma_0$, the gradient EM updates on the gating parameter $\{\boldsymbol{w}\}_{t \geq 0}$ converge geometrically to the true parameter $\boldsymbol{w}^*$, i.e.*

$$\|\boldsymbol{w}_t - \boldsymbol{w}^*\| \leq (\rho_\sigma)^t \|\boldsymbol{w}_0 - \boldsymbol{w}^*\|,$$

where ρ_σ is a dimension-independent constant depending on g and σ.

Remark 3. The condition $\sigma < \sigma_0$ ensures that the Lipschitz constant ρ_σ for the map G is strictly less than 1. The constant α_0 depends only on two universal constants which are nothing but the strong-concavity and the smoothness parameters for the function $Q(\cdot|\boldsymbol{w}^*)$.

Proof. In addition to the assumptions of Appendix B, if we can ensure that the map $-Q(\cdot|\boldsymbol{w}^*)$ is μ-smooth, then the proof follows from Theorem 3 of (Balakrishnan et al., 2017) if we choose $\alpha_0 = \frac{2}{\mu+\lambda}$ where λ is the strong-convexity parameter of $-Q(\cdot|\boldsymbol{w}^*)$. The strong-convexity is already established in Appendix D.3. To find the smoothness parameter, note that

$$\begin{aligned}
-\nabla^2 Q(\boldsymbol{w}|\boldsymbol{w}^*) &= \mathbb{E}\left[f'(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \boldsymbol{x}\boldsymbol{x}^\top\right], \\
&= \mathbb{E}\left[f'''(\boldsymbol{w}^\top \boldsymbol{x})\right] \cdot \boldsymbol{w}\boldsymbol{w}^\top + \mathbb{E}[f'(\boldsymbol{w}^\top \boldsymbol{x})] \cdot I \\
&= \mathbb{E}[f'''(\|\boldsymbol{w}\| Z)] \cdot \boldsymbol{w}\boldsymbol{w}^\top + \mathbb{E}[f'(\|\boldsymbol{w}\| Z)] \cdot I, \quad Z \sim \mathcal{N}(0,1) \\
&\preceq \sup_{0 \leq \alpha \leq 1} \min \left\{\mathbb{E}[f'(\alpha Z)], \mathbb{E}[f'(\alpha Z)] + \alpha^2 \mathbb{E}[f'''(\alpha Z)]\right\} \cdot I \\
&= \underbrace{0.25}_{\mu} \cdot I.
\end{aligned}$$

The contraction parameter is then given by

$$\rho_\sigma = 1 - \frac{2\lambda + 2\gamma_\sigma}{\mu + \lambda}.$$

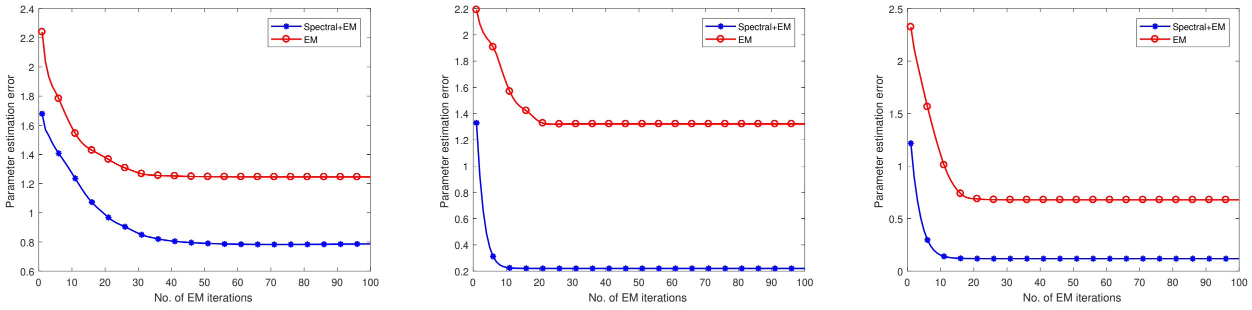
Since $\gamma_\sigma \xrightarrow{\sigma \to 0} 0$, $\rho_\sigma < 1$ whenever $\sigma < \sigma_0$ for a constant σ_0. □

H. Additional experiments

H.1. Synthetic data

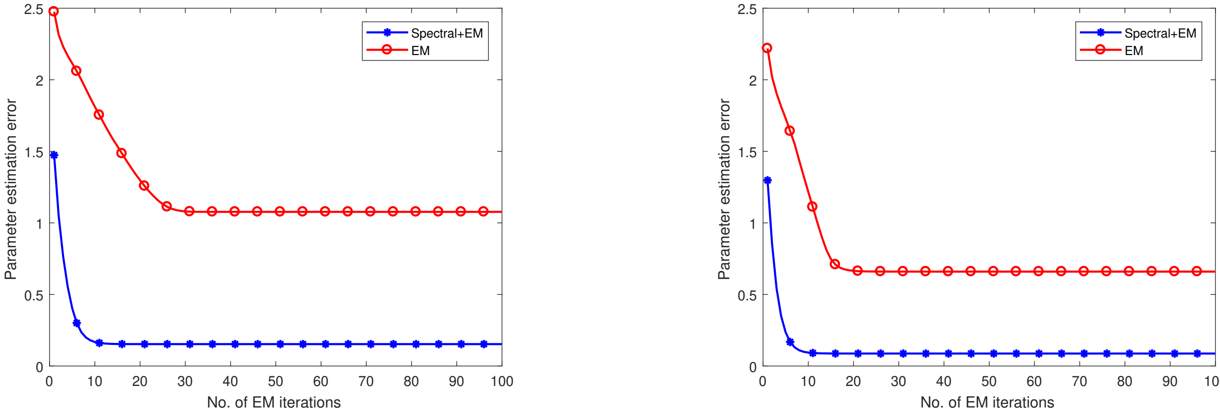
In Figure 4, we varied the number of samples our data set and fixed the other set of parameters to $k = 3, d = 5, \sigma = 0.5$.

In Figure 5 we repeated our experiments for the choice of $n = 10000, d = 5, k = 3$ for two different popular choices of non-linearities: sigmoid and ReLU. The same conclusion as in the linear setting holds in this case too with our algorithm outperforming the EM consistently.



(a) (b) (c)

Figure 4: Plot of parameter estimation error with varying number of samples(n): (a) $n = 1000$ (b) $n = 5000$. (c) $n = 10000$.



(a) (b)

Figure 5: Parameter estimation error for the sigmoid and ReLU nonlinearities respectively.

H.2. Real data

For real data experiments, we choose the 3 standard regression data sets from the UCI Machine Learning Repository: Concrete Compressive Strength Data Set, Stock portfolio performance Data Set, and Airfoil Self-Noise Data Set (Yeh, 1998; Liu & Yeh, 2017; Brooks et al., 1989). In all the three tasks, the goal is to predict the outcome or the response y for each input $\boldsymbol{x}$, which typically contains some task specific attributes. For example, in the concrete compressive strength, the task is to predict the compressive strength of the concrete given its various attributes such as the component of cement, water, age, etc. For this data, the input $\boldsymbol{x} \in \mathbb{R}^8$ corresponds to 8 different attributes of the concrete and the output $y \in \mathbb{R}$ corresponds to its concrete strength. Similarly, for the stock portfolio data set the input $\boldsymbol{x} \in \mathbb{R}^6$ contains the weights of several stock-picking concepts such as weight of the Large S/P concept, weight of the Small systematic Risk concept, etc,. and the output y is the corresponding excess return. The airfoil data set is obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections and the goal is predict the scaled sound pressure level (in dB) given the frequency, angle of attack, etc,. For all the tasks, we pre-processed the data by whitening the input and scaling the output to lie in $(-1, 1)$. We randomly allotted 75% of the data samples for training and the rest for testing. Our evaluation metric is the prediction error on the test set $(\boldsymbol{x}_i, y_i)_{i=1}^n$ defined as

$$\mathcal{E} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2,$$

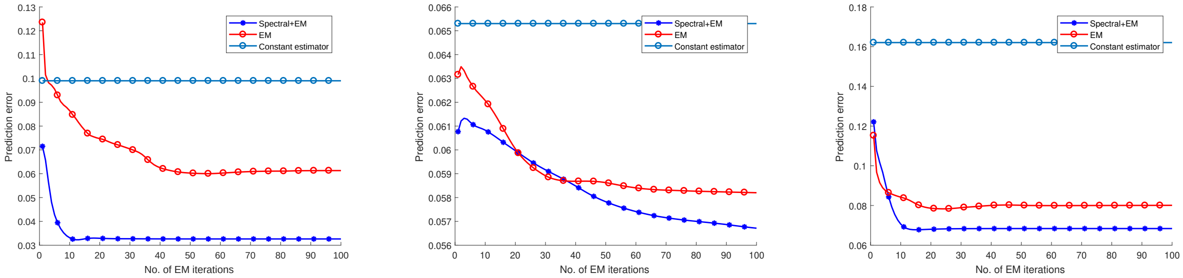


Figure 6: Prediction error for the concrete, stock portfolio and the airfoil data sets respectively.

where $\hat{y}_i$ corresponds to the predicted output response using the learned parameters. In other words,

$$\hat{y} = \sum_{i \in [k]} \frac{e^{\hat{\boldsymbol{w}}_i^\top \boldsymbol{x}}}{\sum_{j \in [k]} e^{\hat{\boldsymbol{w}}_j^\top \boldsymbol{x}}} \cdot g(\hat{\boldsymbol{a}}_i^\top \boldsymbol{x}).$$

We ran the joint-EM algorithm (with 10 different trails) on these tasks with various choices for $k \in \{2, \ldots, 10\}, \sigma \in \{0.1, 0.4, 0.8, 1\}, g \in \{\text{linear}, \text{sigmoid}, \text{ReLU}\}$ and found the best hyper-parameters to be $(k = 3, \sigma = 0.1$ and $g = \text{linear})$, $(k = 3, \sigma = 0.4, g = \text{sigmoid})$ and $(k = 3, \sigma = 0.1, g = \text{linear})$ for the three datasets respectively. For this choice of best hyper-parameters found for joint-EM, we ran our algorithm. Figure 6 highlights the predictive performance of our algorithm as compared to that of the EM. We also plotted the variance of the test data for reference and to gauge the performance of our algorithm. In all the settings our algorithm is able to obtain a better set of parameters resulting in smaller prediction error.