PROVABLY LEARNING MIXTURES OF LINEAR EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) models are widely deployed in modern machine learning. However, few convergence guarantees exist for gradient-based training of MoEs: the objective is nonconvex, and even simple mixtures trap first-order methods in spurious local optima. We analyze a canonical gated mixture featuring linear experts and a gate induced by Gaussian clusters in the input space, and give, to our knowledge, the first global convergence guarantee for jointly training the gate and the experts of such a model by gradient EM on its likelihood. Given a logarithmic overparameterization relative to the true number of experts m, gradient EM converges from randomly initialized gates and least-squares-initialized experts to any prescribed accuracy in polynomially many iterations, under a separation condition on the gate centers. Our analysis establishes a gradient-dominance inequality for the coupled dynamics by decoupling the loss into per-cluster subproblems. Empirical results validate the predicted $m \log m$ phase transition and suggest that the least-squares warm start required by our analysis is not needed in practice: randomly initialized experts perform nearly as well. We also find that leaving the mixing weights unconstrained concentrates routing onto the true number of experts, whereas load-balancing penalties spread it, so that the fitted model can be pruned to a fraction of its width without loss of accuracy.

1 INTRODUCTION

Mixture-of-experts (MoE) models decompose a prediction problem across a collection of specialized *experts*, with a *gate* (or router) that decides which experts are responsible for each input (Jacobs et al., 1991; Jordan & Jacobs, 1994). The idea is now standard in large models: sparsely gated MoE layers let the parameter count of a network grow by orders of magnitude while the compute per input stays fixed (Shazeer et al., 2017; Fedus et al., 2022). Often, many more experts than are necessary are initialized, the gate is initialized at random, and expert load-balancing penalties are implemented to incentivize broad expert utilization. Each of these choices is motivated by experience rather than theory. Why gradient-based training of a gated mixture should succeed at all is only partly understood.

An MoE can be seen as a latent-variable model whose likelihood is nonconvex. The nonconvexity comes from the coupling of gate and experts: the gate decides which expert sees which data, and the experts' errors feed back into the gate. The theory of expectation–maximization (EM) has shed considerable light on the specific latent-variable model of a Gaussian mixture model (GMM). When the GMM has exactly as many components as the ground-truth, EM and its gradient variants enjoy global guarantees only for two components (Xu et al., 2016; Daskalakis et al., 2017) or from a good initialization (Balakrishnan et al., 2017; Yan et al., 2017), and for three or more Gaussian components the population likelihood provably has spurious local maxima that trap EM from random initialization with high probability (Jin et al., 2016). A recent line of work has shown that this obstruction disappears when the model is *over-parameterized*: with $n = \Omega(m \log m)$ fitted components, randomly initialized gradient EM recovers an m-component Gaussian mixture in polynomial time, first for $m = 1$ (Xu et al., 2024) and then for general well-separated mixtures (Zhou et al., 2026). Recent work on MoEs over-provision of experts so that every latent cluster is covered at initialization is the mechanism behind recent gradient-based analyses (Chen et al., 2022; Kawata et al., 2025; Liao & Kyrillidis, 2026). Whether it carries the joint likelihood dynamics of gate and experts has remained open.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper we give a global convergence guarantee for gradient-based training of a Gaussian-gated mixture of linear experts. Data are generated by drawing a latent expert index ℓ with probability π_ℓ^*, then an input $\boldsymbol{x} \sim N(\boldsymbol{\mu}_\ell^*, I_d)$, and finally a response $y = \boldsymbol{x}^\top \boldsymbol{\theta}_\ell^* + \epsilon$ with $\epsilon \sim N(0,1)$ independent of $(\ell, \boldsymbol{x})$. Although this is a generative model, its predictive function is exactly an MoE in the sense used in practice. A quick calculation shows that

$$\mathbf{E}[y \mid \boldsymbol{x}] = \sum_{i=1}^{n} \mathrm{softmax}_i\big(\langle \boldsymbol{x}, \boldsymbol{\mu}_i \rangle + b_i\big)\, \boldsymbol{x}^\top \boldsymbol{\theta}_i, \qquad b_i = \log \pi_i - \tfrac{1}{2}\|\boldsymbol{\mu}_i\|^2, \tag{1}$$

that is, a softmax router with weights $\boldsymbol{\mu}_i$ and biases b_i with linear experts $\boldsymbol{\theta}_i$. The mixing weights π_i, which are usually fixed or ignored in MoE analysis, play the role of learnable router biases. We train this model by gradient EM on the population log-likelihood, which we show coincides with gradient descent on the Kullback–Leibler (KL) divergence between the truth and the model.

Our main result (Theorem 4.2) states that when the fitted model has $n \gtrsim m \log m$ experts, the gates are initialized at random from the data, the experts are initialized by a windowed least-squares fit, and the true gate centers are well separated, gradient EM drives the population KL loss below any target $\varepsilon \geq e^{-c\,dn}$ within a polynomial number of steps. Global guarantees for gradient-based MoE training exist for staged algorithms, purpose-built losses and specific teacher models (Section 2); to our knowledge this is the first for joint training of gate and experts on the likelihood itself, and the first in which the mixing weights are free. The proof establishes a gradient-dominance inequality. A central difficulty is controlling the expert parameters: the expert errors (residuals) $y - \boldsymbol{x}^\top \boldsymbol{\theta}_i$ have to remain bounded within a shrinking budget throughout training. The least-squares warm start supplies the initial bound on the residuals, and an expert-drift argument ensures the budget is not violated.

Our contributions are as follows.

- **A global guarantee for joint gate–expert training** (Theorem 4.2): with $n \gtrsim m \log m$ experts (for balanced mixing weights), random gates and least-squares experts, population gradient EM reaches any accuracy $\varepsilon \geq e^{-c\,dn}$ in polynomially many steps.
- **A proof technique** that decouples the loss into per-cluster problems, controls each by a Gaussian log-Sobolev inequality, and keeps the expert residuals inside a shrinking budget along the whole trajectory.
- **Experiments** showing (i) a sharp success transition along $n \asymp m \log m$ (Figure 1), (ii) that randomly initialized experts succeed almost as often as least-squares experts once the gates cover the clusters (Figure 2), and (iii) that free mixing weights concentrate routing on about m experts and make the model prunable, while load-balancing penalties destroy both properties (Figures 3 and 4).

2 Related Work

EM for exactly parameterized mixtures. The convergence of EM for Gaussian mixtures was studied early by Xu & Jordan (1996). Modern convergence theory begins with Balakrishnan et al. (2017), who showed that population EM converges locally at a geometric rate, and verified this for two-component Gaussian mixtures and two-component mixed linear regression. Subsequent work extended these guarantees to mixtures of many linear regressions, to two-component mixed linear regression from random initialization, and to well-separated Gaussian mixtures (Kwon & Caramanis, 2020b; Kwon et al., 2024; Kwon & Caramanis, 2020a). For three or more components the picture changes: Jin et al. (2016) showed that the population log-likelihood has spurious local maxima to which EM from random initialization converges with high probability.

Over-parameterized mixtures. Fitting more components than the data contain changes the picture. Dwivedi et al. (2020) showed that over-specification creates singularities in the Fisher information and slows the local rate of EM from geometric to polynomial, and Xu et al. (2024) then proved the first global result in this regime: gradient EM fitting an n-component mixture to a single Gaussian converges from random initialization at rate $O(1/\sqrt{t})$. Zhou et al. (2026) extended this to a general m-component, well-separated Gaussian mixture, showing that $n = \Omega(m \log m)$ components suffice for randomly initialized gradient EM to converge globally in polynomial time and with polynomially many samples. Our analysis builds on their framework; the experts add a

second parameter set whose errors enter the gate through the regression residuals, and controlling that coupling is the main new difficulty.

Theory of mixtures of experts. Mixtures of experts were introduced by Jacobs et al. (1991) and Jordan & Jacobs (1994), who already proposed EM for training them. A recent body of work analyzes the *statistical* properties of MoE estimators: convergence rates of the maximum-likelihood estimator for Gaussian MoEs (Ho et al., 2022) along with the effect of the softmax gate on identifiability and rates (Nguyen et al., 2023; 2024a). The Gaussian-gated model of Nguyen et al. (2024b) is closest to ours; however, these results concern the estimator rather than the algorithm that computes it. On the algorithmic side, Fruytier et al. (2025) cast EM for mixtures of experts as mirror descent and obtain local guarantees. Makkuva et al. (2019) recover the experts by a moment method and then the gate by EM, and Makkuva et al. (2020) obtain global guarantees for gradient descent by designing separate losses for the gate and the experts. Closest to us are analyses of gradient-based training under a latent cluster structure: Chen et al. (2022) (top-1 routing and classification, where the nonlinearity of the experts is essential), Kawata et al. (2025) (Gaussian clusters with single-index targets, trained by a multi-phase SGD schedule that updates the router and the experts separately), and Liao & Kyrillidis (2026) (joint gradient flow on the population squared error for a normalized sigmoid router with Hermite experts, an orthonormal teacher and $d \geq \mathrm{poly}(m)$, followed by pruning). All three provision $\Omega(m \log m)$ experts so that every cluster is covered at initialization, as we do. We differ in analyzing the likelihood objective directly, with gate, experts and mixing weights trained jointly by one gradient method, for arbitrary well-separated centers and mixing weights and without a high-dimensional regime.

Routing and load balancing in practice. Sparsely gated MoE layers (Shazeer et al., 2017) and their simplifications (Fedus et al., 2022) route each input to a few experts and add an auxiliary loss that penalizes imbalanced expert utilization, since imbalanced routing wastes capacity and complicates parallel execution. In the idealized case where the fraction of inputs dispatched to expert i equals its average routing probability $\bar{\rho}_i$, the Switch auxiliary loss (Fedus et al., 2022) equals $n \sum_{i=1}^{n} \bar{\rho}_i^2 = n/n_{\text{eff}}$, where $n_{\text{eff}} := \left(\sum_{i=1}^{n} \bar{\rho}_i^2\right)^{-1}$ is the effective number of experts the router uses; minimizing the penalty therefore pushes utilization toward the full fitted width n. More recently, balancing has also been pursued through per-expert bias terms that are added to the routing scores used for expert selection and adjusted by a rule during training, instead of through an auxiliary loss (Wang et al., 2024; DeepSeek-AI, 2024). These biases still aim at balanced load, but they leave the gradient objective untouched.

3 Learning a Mixture of Experts

3.1 Problem Statement

Our data model is an m-component mixed linear regression in which each regression vector has its own Gaussian covariate cluster. A sample $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ is generated by

$$\ell \sim \mathrm{Cat}(\pi_1^*, \ldots, \pi_m^*), \qquad \boldsymbol{x} \mid \ell \sim N(\boldsymbol{\mu}_\ell^*, I_d), \qquad y \mid \boldsymbol{x}, \ell \sim N(\boldsymbol{x}^\top \boldsymbol{\theta}_\ell^*, 1).$$

We can now define the mixture distribution p_* over $\boldsymbol{z}$. Let $\varphi(\boldsymbol{x}; \boldsymbol{\alpha}) := (2\pi)^{-d/2} e^{-\|\boldsymbol{x}-\boldsymbol{\alpha}\|^2/2}$ denote the p.d.f. of a Gaussian with mean $\boldsymbol{\alpha} \in \mathbb{R}^d$ and unit covariance and write the one dimensional case as $\varphi_1(y; a)$. Also, let $\mathcal{P}^k \subset \mathbb{R}^{k+1}$ denote the probability simplex. For $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ the joint density of one component is $\varphi(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\theta}) := \varphi(\boldsymbol{x}; \boldsymbol{\mu})\, \varphi_1(y; \boldsymbol{x}^\top \boldsymbol{\theta})$. The ground truth is then the m-component joint density

$$p_*(\boldsymbol{z}) = \sum_{\ell=1}^{m} \pi_\ell^*\, \varphi(\boldsymbol{z}; \boldsymbol{\mu}_\ell^*, \boldsymbol{\theta}_\ell^*), \quad \boldsymbol{\pi}^* \in \mathcal{P}^{m-1} \tag{2}$$

and the over-parameterized model has $n \geq m$ experts,

$$p(\boldsymbol{z}) = p(\boldsymbol{z}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \pi_i\, \varphi(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\theta}_i), \qquad \boldsymbol{\pi} \in \mathcal{P}^{n-1}. \tag{3}$$

We write $\varphi_i(\boldsymbol{z}) := \varphi(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\theta}_i)$ and $\varphi_\ell^*(\boldsymbol{z}) := \varphi(\boldsymbol{z}; \boldsymbol{\mu}_\ell^*, \boldsymbol{\theta}_\ell^*)$, and drop parameter arguments when they are clear from context. Define $\pi_{\min} := \min_\ell \pi_\ell^*$. We write $\mathbf{E}_\ell := \mathbf{E}_{\boldsymbol{z} \sim \varphi_\ell^*}$ for expectation

under the ℓth true component, i.e., $\boldsymbol{x} \mid \ell \sim N(\boldsymbol{\mu}_\ell^*, I_d)$ and then $y \mid \boldsymbol{x}, \ell \sim N(\boldsymbol{x}^\top \boldsymbol{\theta}_\ell^*, 1)$, so that $\mathbf{E}_{p_*} = \sum_\ell \pi_\ell^* \mathbf{E}_\ell$.

Let $S_\ell := \{i \in [n] : \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| \le \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^*\| \; \forall j\}$ (ties broken by a fixed rule) be the Voronoi group of $\boldsymbol{\mu}_\ell^*$, $\ell(i)$ the group of i, $\hat{\pi}_\ell := \sum_{i \in S_\ell} \pi_i$ and

$$U_\mu(\boldsymbol{\mu}) := \sum_\ell \sum_{i \in S_\ell} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\|^2, \qquad U_\theta(\boldsymbol{\mu}, \boldsymbol{\theta}) := \sum_\ell \sum_{i \in S_\ell} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*\|^2, \qquad U := U_\mu + U_\theta \quad (4)$$

measure how far the fitted components sit from the truth of their group.

3.2 EM Algorithm

EM iteratively updates the parameters $\boldsymbol{\mu} := (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_n^\top)^\top$, $\boldsymbol{\theta} := (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_n^\top)^\top$ and $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)$ of p. At the current iterate $(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')$, the **E-step** computes the posterior probability that component i generated $\boldsymbol{z}$,

$$\psi_i(\boldsymbol{z}; \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}') := \frac{\pi_i' \, \varphi(\boldsymbol{z}; \boldsymbol{\mu}_i', \boldsymbol{\theta}_i')}{p(\boldsymbol{z}; \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')},$$

written $\psi_i(\boldsymbol{z})$ when the parameters are clear, and the **M-step** maximizes, with these posteriors frozen, the expected complete-data log-likelihood (up to an additive constant)

$$Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta} \mid \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}') := \mathbf{E}_{p_*}\Big[\sum_{i=1}^n \psi_i(\boldsymbol{z}; \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')\big(\log \pi_i - \tfrac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 - \tfrac{1}{2}(y - \boldsymbol{x}^\top \boldsymbol{\theta}_i)^2\big)\Big].$$

The expectation over p_* makes this *population* EM. We analyze the *gradient EM* variant in which the M-step for $(\boldsymbol{\mu}, \boldsymbol{\theta})$ is replaced by one gradient step on Q, while the mixing weights are updated by approximately minimizing the loss $\mathcal{L}$ below over the simplex, a convex problem (Algorithm 1). By Claim 3.1, the gradient step on Q is a gradient step on $\mathcal{L}$.

3.3 Loss

We use the KL divergence loss, as in Xu et al. (2024); Zhou et al. (2026),

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) := D_{\mathrm{KL}}(p_* \| p) = \mathbf{E}_{p_*}[-\log p(\boldsymbol{z})] - \mathbf{E}_{p_*}[-\log p_*(\boldsymbol{z})] \quad (5)$$

There are two reasons for this choice. First, $\mathcal{L}$ is nonnegative and vanishes exactly when $p = p_*$, i.e., when the model reproduces the data density and hence the true regression function $\mathbf{E}_{p_*}[y \mid \boldsymbol{x}]$. Second, gradient EM is gradient descent on $\mathcal{L}$:

Claim 3.1. *For any fixed* $(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')$, $\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta} \mid \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')\big|_{(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')} = -\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')$.

As is common when training and deploying MoE models, we are concerned with minimizing the loss $\mathcal{L}$ via a gradient descent procedure. Exact parameter recovery is not the focus of this work.

4 Results

4.1 Assumptions

As is standard, our guarantees are stated in terms of the separation and spread of the true parameters:

$$\Delta := \min_{\ell \neq j} \|\boldsymbol{\mu}_\ell^* - \boldsymbol{\mu}_j^*\|, \quad D_{\max} := \max_{\ell \neq j} \|\boldsymbol{\mu}_\ell^* - \boldsymbol{\mu}_j^*\|, \quad D_\theta := \max_{\ell \neq j} \|\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_j^*\|$$

The loss and the updates are translation equivariant, so we may assume $\max_\ell \|\boldsymbol{\mu}_\ell^*\| \le D_{\max}$ (translate so that one true center is the origin). With $\Lambda := d + D_{\max}^2$ this gives $\mathbf{E}_\ell \|\boldsymbol{x}\|^2 \le 2\Lambda$ for each ℓ. We then make the following well-separatedness assumption on *only* the minimum distance between the centers, where $\delta \in (0, 1)$ is the failure probability of Lemma 4.1 and Theorem 4.2.

Assumption 1 (Well-separated gates)**.** *$m \ge 2$ and $\Delta \ge C_{\mathrm{sep}} \max\{\sqrt{\log(dnm(1 + D_{\max})(1 + D_\theta)/\pi_{\min})}, \sqrt{n(d + \log(n/\delta))}\}$ for a large universal constant C_{sep}. (When $m = 1$ all cross-cluster terms below vanish and the separation requirements are vacuous.)*

The term $\sqrt{n(d+\log(n/\delta))} \asymp B$ is the radius of the region $U \le B^2/2$ that the iteration never leaves, and the proof needs every fitted gate to stay far from every foreign true center, i.e., $\Delta \gg B$. We highlight that a sharper separation condition defined in the appendix, (15), is used in the proofs. Throughout the text, $c, C > 0$ denote universal constants whose values may change from line to line, $a \lesssim b$ means $a \le Cb$, and we set

$$\rho_0 := C\sqrt{d+\log(4n/\delta)}, \qquad B^2 := 4n\rho_0^2, \qquad L_0 := \log n + 2(d+1),$$

and, for a target accuracy ε, $B_\theta := c\min\{\varepsilon/B,\, 1/((1+D_\theta)\Lambda)\}$. Here ρ_0 bounds the distance of every initial gate from the true center it was drawn from, B is the radius of the region in which the analysis takes place, L_0 bounds the initial loss, and B_θ is the expert-error budget. Finally, $\mathsf{A}_{\mathrm{g}}, A = \mathrm{poly}\big(d, m, n, \frac{1}{\pi_{\min}}, B, 1+D_{\max}, 1+D_\theta\big)$ denote polynomial prefactors whose degree is universal and which may be enlarged from line to line (exact forms in Appendix A.2).

4.2 Initialization & Algorithm

Our analysis needs the experts to start close to the truth. We obtain this by a windowed least-squares fit: each $\boldsymbol{\theta}_i^{(0)}$ regresses y on $\boldsymbol{x}$ using only data within distance R of $\boldsymbol{\mu}_i^{(0)}$,

$$\boldsymbol{\theta}_i^{(0)} := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbf{E}_{p_*}\Big[\mathbb{1}\{\|\boldsymbol{x}-\boldsymbol{\mu}_i^{(0)}\| \le R\}\,(y-\boldsymbol{x}^\top\boldsymbol{\theta})^2\Big]. \tag{6}$$

When Δ is large, the window contains almost only data from the cluster that $\boldsymbol{\mu}_i^{(0)}$ was drawn from, so (6) nearly returns that cluster's regression vector; this is item 3 of the following lemma:

Lemma 4.1 (Initialization (Informal)). *Let Assumption 1 hold, fix $\delta \in (0,1)$, suppose $n \ge \frac{C}{\pi_{\min}}\log\frac{2m}{\delta}$, and let $R := \rho_0 + 2\sqrt{d}$ (so that $\Delta \ge 4(R+\rho_0)$ under Assumption 1). Write J_i for the cluster from which $\boldsymbol{\mu}_i^{(0)}$ was drawn in Algorithm 1. Then with probability at least $1-\delta$:*

1. *for each $\ell \in [m]$ there is i with $J_i = \ell$ and $\|\boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}_\ell^*\| \lesssim \sqrt{d}$;*
2. *for every i, $\|\boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}_{J_i}^*\| \le \rho_0$; $\|\boldsymbol{\theta}_i^{(0)} - \boldsymbol{\theta}_{J_i}^*\| \le \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$;*
3. *the initial loss satisfies $\mathcal{L}^{(0)} \le L_0$ (using item 3).*

Item 1 guarantees that every true cluster receives a gate, item 2 bounds the distance of every gate to the center it was drawn from, item 3 is the accuracy of the warm start, and item 4 bounds the initial loss. For $\pi_{\min} \gtrsim 1/m$ (in particular for balanced weights) the requirement on n reads $n \gtrsim m\log(m/\delta)$: this is where the $m\log m$ over-parameterization comes from, and it is a coupon-collector bound, since each of the m true clusters must be hit by at least one of n random gate draws.

Algorithm 1 is gradient EM with extra care for the mixing weights $\boldsymbol{\pi}$, which are constrained to the simplex $\mathcal{P}^{n-1}$. The weight step is ε_π-optimal, i.e., it has simplex dual gap at most ε_π and does not increase the loss (Appendix A.5); the EM fixed-point iteration $\pi_i \leftarrow \mathbf{E}_{p_*}\psi_i$ run to that tolerance qualifies. Near-optimality is what controls the group masses $\hat{\pi}_\ell$ (Lemma 4.4); we do not analyze a projected gradient step on $\boldsymbol{\pi}$.

4.3 Main Result

We now state an informal version of the main result and sketch the argument; full proofs are in the appendix.

Theorem 4.2 (Convergence of gradient EM (Informal)). *Let Assumption 1 hold, fix $\delta \in (0,1)$, and fix a target accuracy ε with $e^{-c\,dn} \le \varepsilon \le c\,\pi_{\min}$. Run Algorithm 1 with $\eta \le c\min\{1/A,\ B_\theta^2/L_0\}$ and $\varepsilon_\pi \le c\,\varepsilon/\mathsf{A}_{\mathrm{g}}$. Then with probability at least $1-\delta$ over the random initialization,*

$$\mathcal{L}(\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}) \le \varepsilon \quad \textit{for some } t \le T := \frac{CB^2}{\eta\varepsilon},$$

and the loss is non-increasing up to that time. For the largest admissible step size, $T = O(L_0 B^4/\varepsilon^3)$. Moreover, for every $t \le T$ before the target is reached, with B and B_θ as above,

$$U(\boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}) \le \frac{B^2}{2}, \qquad U_\theta(\boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}) \le B_\theta^2, \qquad \textit{and the groups } S_\ell \textit{ do not change.}$$

Algorithm 1 Population gradient EM for the MoE with least-squares warm start

1: **Input:** fitted width n, step size η, horizon T, weight-subproblem tolerance ε_π, window radius R
2: **(Random) gate initialization:** draw $\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_n^{(0)} \overset{\text{i.i.d.}}{\sim} p_*^{\boldsymbol{x}}$ (the $\boldsymbol{x}$-marginal of p_*) and set $\pi_i^{(0)} = 1/n$
3: **Expert initialization (windowed least squares):** set each $\boldsymbol{\theta}_i^{(0)}$ as in (6)
4: **for** $t = 0$ to $T-1$ **do**
5: $\quad \boldsymbol{\pi}^{(t+1)} \leftarrow$ an ε_π-optimal, loss non-increasing solution of the convex problem $\min_{\boldsymbol{\pi} \in \mathcal{P}^{n-1}} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)})$
6: $\quad (\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}) \leftarrow (\boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}) - \eta \, \nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}^{(t+1)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)})$
7: **end for**
8: **Output:** $(\boldsymbol{\pi}^{(T)}, \boldsymbol{\mu}^{(T)}, \boldsymbol{\theta}^{(T)})$

We prove Theorem 4.2 by establishing a gradient-dominance inequality $\|\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}\| \gtrsim \mathcal{L}/B$ on the region $U \le B^2/2$, $U_\theta \le B_\theta^2$ (a Łojasiewicz inequality). We obtain it by lower bounding the projection of the gradient onto the *intuitive* direction that moves every gate toward the true center of its group and every expert toward the true regression vector of that group:

$$P := \sum_{\ell \in [m]} \sum_{i \in S_\ell} \Big(\langle \nabla_{\boldsymbol{\mu}_i} \mathcal{L}, \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^* \rangle + \langle \nabla_{\boldsymbol{\theta}_i} \mathcal{L}, \boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^* \rangle \Big)$$

When the separation is large, each cluster is essentially an isolated problem: $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| \le B$ for $i \in S_\ell$, while $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j^*\| \ge \Delta - B \gg B$ for $j \ne \ell$. Since Gaussian densities decay exponentially, a component of group ℓ is essentially invisible to data from cluster j:

Lemma 4.3 (Cross-group terms)**.** *For $i \in S_\ell$, $j \ne \ell$ and any nonnegative polynomial weight $W(\boldsymbol{z})$ of bounded degree (Lemma D.4),* $\mathbf{E}_j\big[\psi_i(\boldsymbol{z}) W(\boldsymbol{z})\big] \le \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$.

We now proceed by writing the normalized within-group mixture, its posterior, and the posterior-averaged weights

$$\lambda_i^{(\ell)} := \frac{\pi_i}{\hat{\pi}_\ell}, \qquad q_\ell(\boldsymbol{z}) := \sum_{i \in S_\ell} \lambda_i^{(\ell)} \varphi_i(\boldsymbol{z}), \qquad \psi_i^{(\ell)} := \frac{\lambda_i^{(\ell)} \varphi_i}{q_\ell}, \qquad \bar{\lambda}_i^{(\ell)} := \mathbf{E}_\ell\big[\psi_i^{(\ell)}\big],$$

so that $\mathcal{L}_\ell := D_{\mathrm{KL}}(\varphi_\ell^* \| q_\ell)$ is the loss of the ℓth true component. Grouping the model's components by the S_ℓ writes both densities as m-term mixtures: $p_* = \sum_\ell \pi_\ell^* \varphi_\ell^*$ and $p = \sum_\ell \hat{\pi}_\ell q_\ell$. The log-sum inequality gives $D_{\mathrm{KL}}(p_* \| p) \le \sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\varphi_\ell^* \| q_\ell) + D_{\mathrm{KL}}(\boldsymbol{\pi}^* \| \hat{\boldsymbol{\pi}})$. Separation and the near-optimal $\boldsymbol{\pi}$-step make the weight-mismatch term small:

Lemma 4.4 (Loss vs. within-group KL)**.** $\mathcal{L} \le \sum_\ell \pi_\ell^* \mathcal{L}_\ell + C\varepsilon_\pi + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$.

Turning our attention to the individual clusters, define the within-cluster projection

$$P_\ell := \sum_{i \in S_\ell} \Big(\langle \nabla_{\boldsymbol{\mu}_i} \mathcal{L}_\ell, \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^* \rangle + \langle \nabla_{\boldsymbol{\theta}_i} \mathcal{L}_\ell, \boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^* \rangle \Big). \tag{7}$$

Using Stein's lemma (Stein, 1981) (Lemma A.3) and the Gaussian log-Sobolev inequality (Gross, 1975), we show that within a cluster the projection P_ℓ dominates the cluster loss $\mathcal{L}_\ell$:

Lemma 4.5 (Projection comparison)**.** *Under the assumptions of Theorem 4.2, for every $\ell \in [m]$,*

$$P_\ell \;\ge\; 2\mathcal{L}_\ell - 2D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)} \| \lambda^{(\ell)}) - c\,\varepsilon, \qquad \sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)} \| \lambda^{(\ell)}) \le \varepsilon_\pi + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}.$$

Comparing P_ℓ and P and using Lemma 4.3 to eliminate small cross terms, and using the good initialization of the regression terms, gives the following decoupling to the intuitive direction

Lemma 4.6 (Group decoupling)**.** *Under the assumptions of Theorem 4.2,*

$$P \;\ge\; \sum_{\ell \in [m]} \pi_\ell^* P_\ell - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}.$$

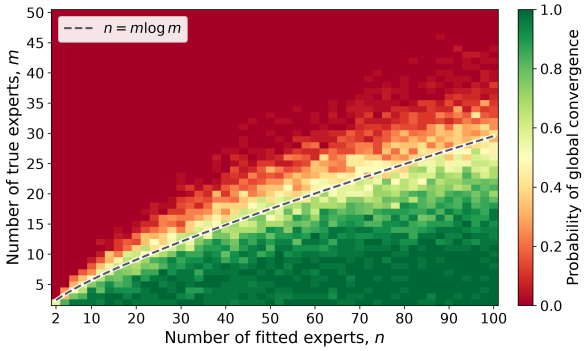

Figure 1: Probability of global convergence across parameterization regimes, with the LS warm start (35 trials per cell). The dashed line is the reference curve $n = m \log m$.

Chaining Lemmas 4.6, 4.5 and 4.4 and absorbing the error terms into $c\varepsilon \le c\mathcal{L}$ gives

Lemma 4.7 (Descent direction)**.** *Under the assumptions of Theorem 4.2,*

$$P \ge c\, \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}).$$

Since $P \le \|\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}\|_F \, B$ when $U \le B^2/2$ by Cauchy–Schwarz, Lemma 4.7 gives $\|\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}\|_F \ge c\mathcal{L}/B$. Two further ingredients close the argument (Appendix B). Local smoothness (Lemma D.1) gives $\mathcal{L}^{(t+1)} \le \mathcal{L}^{(t)} - \frac{\eta}{2}\|\nabla \mathcal{L}\|^2$, which with gradient dominance yields $1/\mathcal{L}^{(t)} \ge 1/\mathcal{L}^{(0)} + c\eta t/B^2$, hence $T \lesssim B^2/(\eta\varepsilon)$. The same projection bound shows that U is non-increasing, and since the expert part of the projection is a nonnegative quadratic form (Proposition E.1), U_θ drifts by only $O(\eta L_0)$ over the run, which the choice of η keeps inside B_θ^2.

5 Experiments

Three ingredients of the analysis are not settled by the theorem: whether the width $n \gtrsim m \log m$ is needed, whether the least-squares warm start is needed, and what free mixing weights buy compared with the balancing that deployed MoEs impose. We test each. Data come from the model (2) and we run finite-sample gradient EM (Algorithm 1 with sample averages) for at most $T = 10{,}000$ updates on $N = 10{,}000$ points. A run succeeds if the joint KL between the truth and the fitted model is at most 0.02, measured on the training sample in Section 5.1 and on the held-out points otherwise. For pruning (Section 5.3) we report the held-out KL of the conditional density of y given $\boldsymbol{x}$, the quantity a deployed MoE is judged on. Full details are in Appendix F.

5.1 Scaling with the number of experts

We run a sweep to test where the failure boundary of n versus m lies. The question is whether the transition follows the diagonal $n = m$ (over-parameterization unnecessary) or the curve $n \asymp m \log m$ (the coverage mechanism of Lemma 4.1). Varying $m = 2, \ldots, 50$ and $n = 2, 4, \ldots, 100$, Figure 1 shows a sharp transition along the reference curve $n \asymp m \log m$, with success near $n = m$ remaining low. The theorem only shows that $n \gtrsim m \log m$ suffices; the sharpness of the empirical transition along that curve, and the low success near $n = m$ for large m, suggest that logarithmic over-provisioning is also necessary for gradient EM from random gates.

5.2 Necessity of the least-squares warm start

Since modern MoE architectures train their experts and gates from random initialization, we compare LS initialization against $\boldsymbol{\theta}_i^{(0)} \sim N(0, I_d)$ with everything else held fixed. We sweep the fitted width $n \in \{4, 8, \ldots, 24\}$ at weak ($\Delta = 2$) and strong ($\Delta = 12$) gate separation. By Lemma 4.1 the warm start is most accurate when Δ is large, so $\Delta = 12$ is where it should help most, and $\Delta = 2$ is a stress test outside the theory.

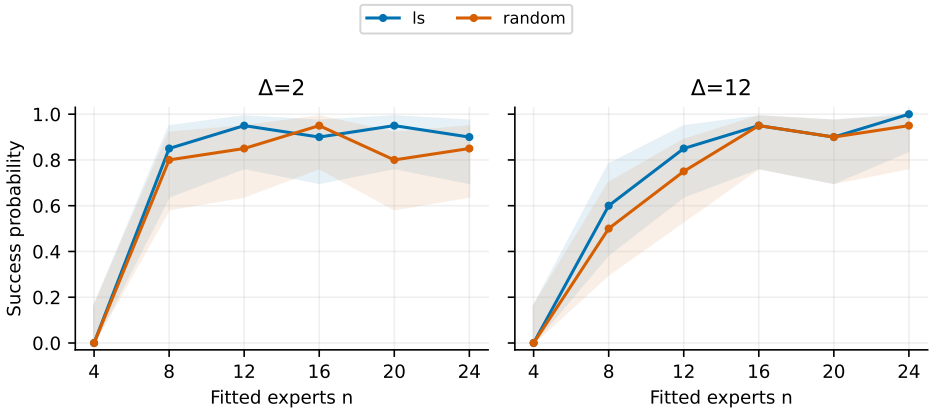

Figure 2: Success probability versus fitted width for LS and random expert initialization. *(Left)* Weak gate separation. *(Right)* Strong gate separation. Shaded regions are confidence intervals over the 20 seeds

.

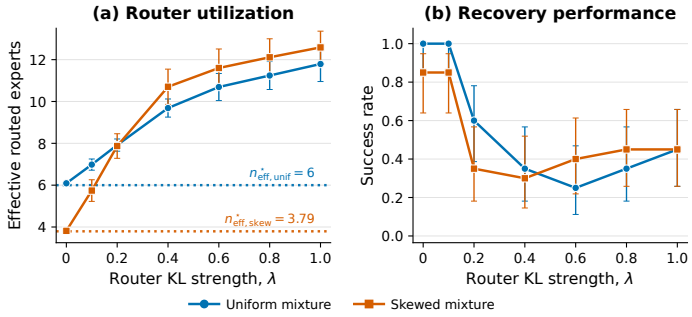

Figure 3: Effect of router load-balancing regularization on expert utilization and recovery. **(a)** Effective number of routed experts, n_{eff}, against the router-KL strength λ for uniform and skewed ground-truth mixture weights. Dotted lines show the corresponding oracle expert counts. **(b)** Recovery success against λ for the same two ground truths. Error bars are 95% confidence intervals.

In both panels of Figure 2 the two initializations fail at $n = 4 < m$, where the truth is not representable, and succeed together once the model is moderately over-parameterized. This indicates that the most important aspect is sufficient coverage by the initialized cluster centers $\boldsymbol{\mu}_i$. Once a gate lands near each true center, the experts follow regardless of where they started. Strong separation is in fact harder at small n ($n = 8$): a gate cannot cross a wide gap, so coverage at initialization is decisive, whereas for $\Delta = 2$ an uncovered cluster can still attract a gate.

5.3 Mixing weights, balancing, and effective experts

Following the discussion in Section 2, we investigate how load balancing affects the benefits of freely learned mixing coefficients. We fit $n = 16$ experts at $\Delta = 8$, comparing unregularized training with the addition of $\lambda D_{\text{KL}}(\bar{\boldsymbol{\rho}} \| \text{Unif}_n)$ to the loss, where $\bar{\boldsymbol{\rho}}$ denotes average input-only router usage and $\lambda \in [0, 1]$. Mixing coefficients remain trainable in both cases; the penalty encourages uniform usage across fitted experts. We consider uniform and skewed true mixture weights, specified in Appendix F.

We measure utilization by the effective expert count $n_{\text{eff}} = \left(\sum_{i=1}^{n} \bar{\rho}_i^2\right)^{-1}$ of Section 2, where $\rho_i(\boldsymbol{x}) = P(i \mid \boldsymbol{x})$ is the responsibility ψ_i with y marginalized out and $\bar{\rho}_i$ is its average over the held-out inputs. For large Δ the routing mass of a point from cluster ℓ falls almost entirely on S_ℓ, so if a single expert carries each group, $\bar{\boldsymbol{\rho}}$ reproduces $\boldsymbol{\pi}^*$ and n_{eff} equals $n^*_{\text{eff}} := \left(\sum_{\ell=1}^{m} (\pi^*_\ell)^2\right)^{-1}$. For a uniform prior this is $n^*_{\text{eff}} = (m \cdot m^{-2})^{-1} = m$ and 3.794 for our skewed one. Free weights recover both almost exactly, at 6.09 and 3.82 (Figure 3). Raising λ increases n_{eff} monotonically

(Figure 3(a)), while the success rate drops from 1.00 and 0.85 for uniform and skewed mixtures, respectively, to between 0.25 and 0.45 for $\lambda \geq 0.4$ (Figure 3(b)). Mild balancing increases effective utilization before reducing observed fitting success: at $\lambda = 0.1$, the empirical success rates are unchanged, while n_{eff} rises to approximately 7.0 and 5.4, respectively. Higher utilization is therefore not a proxy for better training. Encouraging uniform usage across fitted experts can reduce convergence success under the tested optimization budget, even when the true mixture weights are uniform.

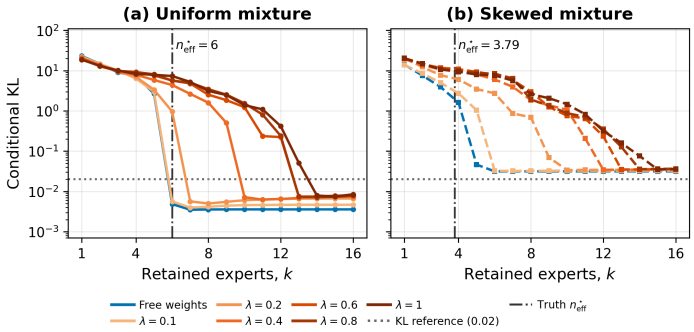

Figure 4: Effect of global expert pruning. Mean held-out conditional KL over all 20 seeds, including unsuccessful training runs, is evaluated after retaining the k experts with the largest learned mixing weights, renormalized without retraining, for **(a)** uniform and **(b)** skewed ground-truth mixture weights. The dotted line at 0.02 is the joint-KL threshold used to define training success.

The same effect decides whether over-parameterization pays off at inference. We prune to the k experts with the largest learned mixing weights after the MoE has trained. Free-weight models reach their accuracy floor once $k \approx m$, while strongly balanced ones need nearly all the fitted experts (Figure 4). Over-provisioning is what makes optimization succeed (Figure 1); whether it also costs inference depends on whether the surplus can be discarded afterward, and that is decided by the router. A balancing penalty does the opposite: it drives n_{eff} towards n by construction and leaves a more evenly used router with nothing to prune.

6 Conclusion and Future Work

We showed that gradient EM, equivalently gradient descent on the population KL loss, provably learns a Gaussian-gated mixture of m linear experts from random gates and least-squares experts, provided the fitted width is $n \gtrsim m \log m$ and the gate centers are well separated. The experiments add two observations. First, the least-squares warm start is not needed in practice: randomly initialized experts converge almost as often once the gates cover the clusters. Second, free mixing weights concentrate routing on about m experts, so the trained model can be pruned to the true width, whereas load-balancing penalties spread routing over most of the n experts and leave little to prune.

Three questions follow. Can the least-squares warm start be removed from the analysis, as the experiments suggest? Furthermore, can the strong separation condition be relaxed as suggested in the same experiment? And does n_{eff} provably track n^*_{eff} at the end of training, as Figure 3 indicates? Both experiments point to the same bottleneck: whether the initial gates cover the true clusters. Characterizing when gates are learnable from scratch, and connecting the optimization view taken here with the statistical rates of Ho et al. (2022); Nguyen et al. (2024b), is the natural next step.

Limitations. Our guarantee is for population gradient EM; a full sample-based analysis, as Zhou et al. (2026) give for Gaussian mixtures, remains to be done. The model has identity covariances, unit noise and linear experts, and the gate is induced by Gaussian clusters rather than a free softmax router. The separation required by Assumption 1 grows with the fitted width, which we believe to be an artifact of the potential U, and the step size must scale with ε^2, giving a $1/\varepsilon^3$ iteration count. Finally, the experts must be warm-started by least squares, although the experiments indicate that this is not needed in practice.

AI USE STATEMENT

All research ideas in this paper are the authors' own: the problem, the high-level proof strategy, and the choice of experiments. Large language models (LLMs: Claude, Gemini, ChatGPT) were used as assistants for the remaining work, always together with manual effort. The authors checked every LLM output and take full responsibility for the content of the paper. LLMs helped refine the exact statements of the mathematical claims and assisted in writing the proofs; the authors verified each such step. The code that generates the synthetic data (from the data model specified by the authors), implements the algorithm, runs the experiments and produces the figures was written largely by LLMs, and was reviewed and tested by the authors. Relevant work was identified both by the authors and through LLM-assisted searches and summaries, and every citation was checked by the authors. LLMs were also used to draft and edit parts of the paper for readability, to give feedback on how the method and experiments are presented, and to format references.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Section 4 and complete proofs are in Appendices A–E. The data-generating process, hyperparameters and evaluation protocol of every experiment are listed in Appendix F.

REFERENCES

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the Mixture-of-Experts layer in deep learning. In *Advances in Neural Information Processing Systems*, 2022.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two Gaussians. In *Conference on Learning Theory*, 2017.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J. Wainwright, Michael I. Jordan, and Bin Yu. Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, 48(6): 3161–3182, 2020.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Quentin Fruytier, Aryan Mokhtari, and Sujay Sanghavi. Learning mixtures of experts with EM: A mirror descent perspective. In *International Conference on Machine Learning*, 2025.

Leonard Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Nhat Ho, Chiao-Yu Yang, and Michael I. Jordan. Convergence rates for Gaussian mixtures of experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, 2016.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

Ryotaro Kawata, Kohsei Matsutani, Yuri Kinoshita, Naoki Nishikawa, and Taiji Suzuki. Mixture of experts provably detect and learn the latent cluster structure in gradient-based learning. In *International Conference on Machine Learning*, 2025.

Jeongyeol Kwon and Constantine Caramanis. The EM algorithm gives sample-optimality for learning mixtures of well-separated Gaussians. In *Conference on Learning Theory*, 2020a.

Jeongyeol Kwon and Constantine Caramanis. EM converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, 2020b.

Jeongyeol Kwon, Wei Qian, Yudong Chen, Constantine Caramanis, Damek Davis, and Nhat Ho. Global optimality of the EM algorithm for mixtures of two-component linear regressions. *IEEE Transactions on Information Theory*, 70(9):6519–6546, 2024.

Fangshuo Liao and Anastasios Kyrillidis. Guided by the experts: Provable feature learning dynamic of soft-routed Mixture-of-Experts. In *International Conference on Artificial Intelligence and Statistics*, 2026.

Ashok Vardhan Makkuva, Pramod Viswanath, Sreeram Kannan, and Sewoong Oh. Breaking the gridlock in Mixture-of-Experts: Consistent and efficient algorithms. In *International Conference on Machine Learning*, 2019.

Ashok Vardhan Makkuva, Sewoong Oh, Sreeram Kannan, and Pramod Viswanath. Learning in gated neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Huy Nguyen, TrungTin Nguyen, and Nhat Ho. Demystifying softmax gating function in Gaussian mixture of experts. In *Advances in Neural Information Processing Systems*, 2023.

Huy Nguyen, Pedram Akbarian, TrungTin Nguyen, and Nhat Ho. A general theory for softmax gating multinomial logistic mixture of experts. In *International Conference on Machine Learning*, 2024a.

Huy Nguyen, TrungTin Nguyen, Khai Nguyen, and Nhat Ho. Towards convergence rates for parameter estimation in Gaussian-gated mixture of experts. In *International Conference on Artificial Intelligence and Statistics*, 2024b.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated Mixture-of-Experts layer. In *International Conference on Learning Representations*, 2017.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for Mixture-of-Experts. *arXiv preprint arXiv:2408.15664*, 2024.

Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems*, 2016.

Lei Xu and Michael I. Jordan. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.

Weihang Xu, Maryam Fazel, and Simon S. Du. Toward global convergence of gradient EM for over-parameterized Gaussian mixture models. In *Advances in Neural Information Processing Systems*, 2024.

Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems*, 2017.

Mo Zhou, Weihang Xu, Maryam Fazel, and Simon S. Du. Global convergence of gradient EM for over-parameterized Gaussian mixtures. *arXiv preprint arXiv:2506.06584*, 2026.

A NOTATION, CONSTANTS, AND PRELIMINARIES

This appendix contains the proofs of all statements in the body of the paper. Throughout, $c, C > 0$ are universal constants whose value may change from line to line, $a \lesssim b$ means $a \le Cb$, and $\mathsf{A}_{\mathrm{g}} = \mathrm{poly}\big(d, m, n, 1/\pi_{\min}, 1+B, 1+D_{\max}, 1+D_\theta\big)$ denotes a polynomial prefactor of universal degree which may be enlarged from line to line. A single accuracy parameter ε is used throughout: it is simultaneously the target accuracy of Theorem 4.2 and the threshold below which the error terms of the analysis are absorbed. The Gaussian observation noise is always written ϵ, as in the data model of the main text.

A.1 NOTATION

Recall $\boldsymbol{z} = (\boldsymbol{x}, y)$, the truth (2), the model (3), the Voronoi groups S_ℓ and the potentials (4). We write

$$\varphi_i(\boldsymbol{z}) := \varphi(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\theta}_i), \qquad \varphi^*_\ell(\boldsymbol{z}) := \varphi(\boldsymbol{z}; \boldsymbol{\mu}^*_\ell, \boldsymbol{\theta}^*_\ell),$$
$$\psi_i(\boldsymbol{z}) := \frac{\pi_i \varphi_i(\boldsymbol{z})}{p(\boldsymbol{z})}, \qquad r_i(\boldsymbol{z}) := y - \boldsymbol{x}^\top \boldsymbol{\theta}_i.$$

for the components, the posterior (membership) weights and the residual of expert i. Sampling from cluster j means

$$\boldsymbol{x} = \boldsymbol{\mu}^*_j + \boldsymbol{Z}, \qquad y = \boldsymbol{x}^\top \boldsymbol{\theta}^*_j + \epsilon, \qquad (\boldsymbol{Z}, \epsilon) \sim N(0, I_{d+1}), \tag{8}$$

and we write $\mathbf{E}_j$ for the corresponding expectation, so that $\mathbf{E}_{p_*} = \sum_j \pi^*_j \mathbf{E}_j$. Relative to cluster j we use the recentered parameters

$$\boldsymbol{h}^{(j)}_k := \boldsymbol{\mu}_k - \boldsymbol{\mu}^*_j, \qquad \beta^{(j)}_k := \boldsymbol{\theta}_k - \boldsymbol{\theta}^*_j, \qquad a^{(j)}_k(\boldsymbol{x}) := \boldsymbol{x}^\top \beta^{(j)}_k, \qquad \text{so that } r_k = \epsilon - a^{(j)}_k. \tag{9}$$

The superscript (j) is dropped whenever the cluster is clear from the context. Finally, for each group we use the normalized within-group mixture and its posterior,

$$\lambda^{(\ell)}_i := \frac{\pi_i}{\hat{\pi}_\ell}, \qquad q_\ell := \sum_{i \in S_\ell} \lambda^{(\ell)}_i \varphi_i, \qquad \psi^{(\ell)}_i := \frac{\lambda^{(\ell)}_i \varphi_i}{q_\ell},$$
$$\bar{\lambda}^{(\ell)}_i := \mathbf{E}_\ell\big[\psi^{(\ell)}_i\big], \qquad \mathcal{L}_\ell := D_{\mathrm{KL}}(\varphi^*_\ell \| q_\ell). \tag{10}$$

A.2 CONSTANTS AND THE SHARP SEPARATION CONDITION

Fix $\delta \in (0,1)$ and a target accuracy ε. With $\rho_0 := C\sqrt{d + \log(4n/\delta)}$ and $R := \rho_0 + 2\sqrt{d}$ as in Algorithm 1, set

$$L_0 := \log n + 2(d+1), \qquad B^2 := 4n\rho_0^2, \tag{11}$$

so that, on the event of Lemma B.1, $\mathcal{L}^{(0)} \le L_0$ and $U(\boldsymbol{\mu}^{(0)}, \boldsymbol{\theta}^{(0)}) \le B^2/2$: item 2 of that lemma gives $U_\mu(\boldsymbol{\mu}^{(0)}) \le n\rho_0^2 = B^2/4$, and (27) gives $U_\theta(\boldsymbol{\mu}^{(0)}, \boldsymbol{\theta}^{(0)}) \le B_\theta^2/4 \le \frac{1}{4}$. Since $n \ge 1$ and $C \ge 1$ in the definition of ρ_0, we also have $B = 2\sqrt{n}\,\rho_0 \ge 2$ and $L_0 \ge 1$. Define the *expert-error budget*, the smoothness constant and the step-size and weight-tolerance constants by

$$B_\theta := c_\theta \min\Big\{\frac{\varepsilon}{B}, \frac{1}{(1+D_\theta)\Lambda}\Big\}, \tag{12}$$

$$A := C(1+\Lambda)^2\Big(1 + d + B^2 + \log\frac{1}{\pi_{\min}} + \frac{1}{\pi_{\min}}\Big), \qquad A_\eta := \max\Big\{A, \frac{L_0}{B_\theta^2}\Big\}, \qquad A_\pi := \frac{\mathsf{A}_{\mathrm{g}}}{c}, \tag{13}$$

with c_θ a small universal constant. Note that A_η depends on the target accuracy through B_θ: for small ε we have $B_\theta = c_\theta \varepsilon / B$, hence $A_\eta = L_0 B^2/(c_\theta^2 \varepsilon^2)$ and the admissible step size scales as ε^2. Since B^2 appears additively inside A with a large constant, and since $B^2 = 4n\rho_0^2 \ge 4C^2\big(d + \log(4n/\delta)\big) \ge \log n + 2(d+1) = L_0$ for $n \ge 1$ and $C \ge 1$, we always have

$$A \;\ge\; B^2 \;\ge\; L_0, \tag{14}$$

a fact used repeatedly to bound η-dependent quantities. The proofs use the following sharp form of Assumption 1:

$$\Delta \;\geq\; C_{\mathrm{sep}}\Big(B + \sqrt{\log \mathsf{A}_{\mathrm{g}}} + \sqrt{\tfrac{L_0+\log(1/\pi_{\min})}{\pi_{\min}}} + \sqrt{\log \tfrac{1}{\varepsilon}} + \sqrt{\log \tfrac{1}{B_\theta}}\Big), \tag{15}$$

for a large universal constant $C_{\mathrm{sep}} \geq 32$. Every term of (15) is of the form $\sqrt{\mathrm{poly}}$ or $\sqrt{\log \mathrm{poly}}$, since A_{g} and $1/B_\theta$ are polynomial in the problem parameters; Assumption 1 is the readable form of this requirement, and the term $\sqrt{\log(1/\varepsilon)}$ is what forces the restriction $\varepsilon \geq e^{-c\,dn}$. Condition (15) is used in exactly two ways: through $\Delta \geq C_{\mathrm{sep}}B$, which drives Lemma A.1, and through the implication

$$\mathsf{A}_{\mathrm{g}}\, e^{-c\Delta^2} \;\leq\; c\,\varepsilon\, B_\theta^2/B^2, \tag{16}$$

which follows by squaring (15) and taking C_{sep} large: the terms $\sqrt{\log \mathsf{A}_{\mathrm{g}}}$, $\sqrt{\log(1/\varepsilon)}$ and $\sqrt{\log(1/B_\theta)}$ contribute exactly the factors $\mathsf{A}_{\mathrm{g}}^{-1}$, ε and B_θ^2 on the right-hand side, and $B \leq \mathsf{A}_{\mathrm{g}}$.

Two standing consequences of the potential bounds will be used without comment:

$$U \leq \tfrac{B^2}{2} \implies \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| \leq B \;\; (i \in S_\ell), \qquad U_\theta \leq B_\theta^2 \implies \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*\| \leq B_\theta \;\; (i \in S_\ell). \tag{17}$$

Finally we abbreviate

$$\kappa := 2B_\theta(D_\theta + 2B_\theta). \tag{18}$$

Lemma A.1 (Elementary consequences of the choice of constants)**.** *Assume Assumption 1 (in particular $m \geq 2$) and* (12)–(15)*. Then*

(i) $B \leq \Delta \leq 2D_{\max}$ *and* $\Lambda + B^2 \leq 5\Lambda$*;*
(ii) $B_\theta \leq c_\theta \leq 1$ *and* $B_\theta^2\Lambda \leq 1$*;*
(iii) $\kappa(\Lambda + B^2) \leq c_0 := 20\, c_\theta$*;*
(iv) $B < \Delta/4$ *and* $\Delta > 2(B+1)$.

Proof. (i) Condition (15) gives $\Delta \geq C_{\mathrm{sep}}B \geq B$. In the working coordinates fixed in the main text, every true center satisfies $\|\boldsymbol{\mu}_\ell^*\| \leq D_{\max}$, so $\Delta \leq \big\|\boldsymbol{\mu}_\ell^* - \boldsymbol{\mu}_j^*\big\| \leq 2D_{\max}$ for any $\ell \neq j$. Hence $B^2 \leq 4D_{\max}^2 \leq 4\Lambda$ because $\Lambda = d + D_{\max}^2 \geq D_{\max}^2$, and adding Λ gives $\Lambda + B^2 \leq 5\Lambda$.

(ii) By (12), $B_\theta \leq c_\theta/\big((1+D_\theta)\Lambda\big) \leq c_\theta \leq 1$, since $D_\theta \geq 0$ and $\Lambda \geq d \geq 1$. The same bound multiplied by Λ gives

$$B_\theta\Lambda \;\leq\; \frac{c_\theta}{1+D_\theta} \;\leq\; c_\theta, \tag{19}$$

so $B_\theta^2\Lambda = B_\theta \cdot (B_\theta\Lambda) \leq 1 \cdot c_\theta \leq 1$.

(iii) By (i) and (19),

$$\kappa(\Lambda + B^2) \;\leq\; 2B_\theta(D_\theta + 2B_\theta)\cdot 5\Lambda \;=\; 10\,(B_\theta\Lambda)(D_\theta + 2B_\theta) \;\leq\; 10\,\frac{c_\theta}{1+D_\theta}\,(D_\theta + 2) \;\leq\; 20\, c_\theta,$$

where we used $B_\theta \leq 1$ in $D_\theta + 2B_\theta \leq D_\theta + 2$ and then $D_\theta + 2 \leq 2(1+D_\theta)$.

(iv) $B \leq \Delta/C_{\mathrm{sep}} \leq \Delta/32 < \Delta/4$. Since $B \geq 2$, $\Delta > 2(B+1)$. □

A.3 Gaussian moments

Lemma A.2 (Moment bounds)**.** *For every $\ell \in [m]$ and every $\boldsymbol{v} \in \mathbb{R}^d$,*

$$\mathbf{E}_\ell \|\boldsymbol{x}\|^2 \leq 2\Lambda, \qquad \mathbf{E}_\ell \|\boldsymbol{x}\|^4 \leq C\Lambda^2,$$
$$\mathbf{E}_\ell\big[(\boldsymbol{x}^\top \boldsymbol{v})^2\big] \leq 2\Lambda \|\boldsymbol{v}\|^2, \qquad \mathbf{E}_\ell\big[(\boldsymbol{x}^\top \boldsymbol{v})^4\big] \leq C\Lambda^2 \|\boldsymbol{v}\|^4.$$

Proof. Write $\boldsymbol{x} = \boldsymbol{\mu}_\ell^* + \boldsymbol{Z}$ with $\boldsymbol{Z} \sim N(0, I_d)$. Then $\mathbf{E}_\ell \|\boldsymbol{x}\|^2 = \|\boldsymbol{\mu}_\ell^*\|^2 + d \leq D_{\max}^2 + d = \Lambda$. For the fourth moment, $(a+b)^4 \leq 8a^4 + 8b^4$ and $\mathbf{E}\|\boldsymbol{Z}\|^4 = d(d+2) \leq 3d^2$ give $\mathbf{E}_\ell \|\boldsymbol{x}\|^4 \leq 8D_{\max}^4 + 24d^2 \leq 32\Lambda^2$. For the projections, $\boldsymbol{x}^\top \boldsymbol{v} \sim N(\langle \boldsymbol{\mu}_\ell^*, \boldsymbol{v}\rangle, \|\boldsymbol{v}\|^2)$, so its second moment is $\langle \boldsymbol{\mu}_\ell^*, \boldsymbol{v}\rangle^2 + \|\boldsymbol{v}\|^2 \leq (D_{\max}^2 + 1)\|\boldsymbol{v}\|^2 \leq 2\Lambda\|\boldsymbol{v}\|^2$, and for $W \sim N(\varrho, s^2)$ one has $\mathbf{E}W^4 = \varrho^4 + 6\varrho^2 s^2 + 3s^4$, whence $\mathbf{E}_\ell(\boldsymbol{x}^\top \boldsymbol{v})^4 \leq \big(D_{\max}^4 + 6D_{\max}^2 + 3\big)\|\boldsymbol{v}\|^4 \leq C\Lambda^2\|\boldsymbol{v}\|^4$. □

A.4 STEIN'S LEMMA AND THE GRADIENT

Lemma A.3 (Stein's lemma). *Let $\boldsymbol{w} \sim N(\boldsymbol{a}, \sigma^2 I_k)$ and let $g : \mathbb{R}^k \to \mathbb{R}$ be differentiable. Then $\mathbf{E}\big[g(\boldsymbol{w})(\boldsymbol{w}-\boldsymbol{a})\big] = \sigma^2\,\mathbf{E}\big[\nabla g(\boldsymbol{w})\big]$ when the expectations are well-defined. Equivalently, for a differentiable function $F : \mathbb{R}^k \to \mathbb{R}^k$, $\mathbf{E}\,\langle \boldsymbol{w}-\boldsymbol{a}, F(\boldsymbol{w})\rangle = \sigma^2\mathbf{E}\big[\nabla\cdot F(\boldsymbol{w})\big]$.*

We apply Lemma A.3 to the pair $(\boldsymbol{Z}, \epsilon) \sim N(0, I_{d+1})$ of (8). Every function to which it is applied below is a posterior weight, a residual, or a product of these with a polynomial; all such functions are smooth and of at most polynomial growth after the change of variables, so the integrability hypotheses hold throughout and we do not repeat this verification.

Claim A.1 (Gradient EM is gradient descent on $\mathcal{L}$; restatement of Claim 3.1). *For any fixed $(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')$, $\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} Q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta} \mid \boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')\big|_{(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')} = -\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\pi}', \boldsymbol{\mu}', \boldsymbol{\theta}')$, and*

$$\nabla_{\boldsymbol{\mu}_i}\mathcal{L} = \mathbf{E}_{p_*}\big[\psi_i(\boldsymbol{z})(\boldsymbol{\mu}_i - \boldsymbol{x})\big], \qquad \nabla_{\boldsymbol{\theta}_i}\mathcal{L} = -\mathbf{E}_{p_*}\big[\psi_i(\boldsymbol{z})\, r_i(\boldsymbol{z})\, \boldsymbol{x}\big], \qquad \nabla_{\pi_i}\mathcal{L} = -\mathbf{E}_{p_*}\Big[\frac{\varphi_i}{p}\Big]. \tag{20}$$

Proof. By (5), $\mathcal{L} = \mathbf{E}_{p_*}[-\log p] + \text{const}$, where the constant is the (parameter-free) negative entropy of p_*. Since $p = \sum_k \pi_k\varphi_k$ and only the ith summand depends on $(\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\theta}_i)$, the log-derivative identity gives

$$\nabla_{\boldsymbol{\mu}_i}\mathcal{L} = -\mathbf{E}_{p_*}\Big[\frac{\pi_i \nabla_{\boldsymbol{\mu}_i}\varphi_i}{p}\Big] = -\mathbf{E}_{p_*}\Big[\frac{\pi_i\varphi_i}{p}\,\nabla_{\boldsymbol{\mu}_i}\log\varphi_i\Big] = -\mathbf{E}_{p_*}\big[\psi_i(\boldsymbol{x}-\boldsymbol{\mu}_i)\big],$$

because $\log\varphi_i(\boldsymbol{z}) = \text{const} - \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2 - \frac{1}{2}r_i(\boldsymbol{z})^2$ has $\nabla_{\boldsymbol{\mu}_i}\log\varphi_i = \boldsymbol{x}-\boldsymbol{\mu}_i$ and $\nabla_{\boldsymbol{\theta}_i}\log\varphi_i = r_i\boldsymbol{x}$; the same computation with $\nabla_{\boldsymbol{\theta}_i}$ and with $\partial_{\pi_i}\varphi$ gives the two remaining identities in (20). In Q the posterior ψ' is frozen, so $\nabla_{\boldsymbol{\mu}_i}Q = \mathbf{E}_{p_*}[\psi_i'(\boldsymbol{x}-\boldsymbol{\mu}_i)]$ and $\nabla_{\boldsymbol{\theta}_i}Q = \mathbf{E}_{p_*}[\psi_i' r_i\boldsymbol{x}]$; evaluating at $(\boldsymbol{\mu},\boldsymbol{\theta}) = (\boldsymbol{\mu}', \boldsymbol{\theta}')$ makes $\psi' = \psi$ and yields exactly $-\nabla_{\boldsymbol{\mu},\boldsymbol{\theta}}\mathcal{L}$. □

Claim A.2 (Stein form of the gradient). *Fix a data cluster j, use the recentered parameters* (9)*, and write the posterior averages, in which the sums run over all n components with the global posterior ψ_k,*

$$\bar{\boldsymbol{h}}^{(j)} := \sum_{k\in[n]}\psi_k\boldsymbol{h}_k^{(j)}, \quad \bar{\boldsymbol{w}}^{(j)} := \sum_{k\in[n]}\psi_k r_k\beta_k^{(j)},$$
$$\bar{\beta}^{(j)} := \sum_{k\in[n]}\psi_k\beta_k^{(j)}, \quad \bar{a}^{(j)} := \sum_{k\in[n]}\psi_k a_k^{(j)} = \boldsymbol{x}^\top\bar{\beta}^{(j)}.$$

Then

$$\nabla_{\boldsymbol{\mu}_i}\mathcal{L} = \sum_j \pi_j^*\,\mathbf{E}_j\Big[\psi_i\big(\bar{\boldsymbol{h}}^{(j)} + \bar{\boldsymbol{w}}^{(j)} - r_i\beta_i^{(j)}\big)\Big], \tag{21}$$

$$\nabla_{\boldsymbol{\theta}_i}\mathcal{L} = \sum_j \pi_j^*\,\mathbf{E}_j\Big[\psi_i\,\boldsymbol{x}\,\boldsymbol{x}^\top\bar{\beta}^{(j)}\Big]. \tag{22}$$

Proof. Fix j and drop the superscript. Under cluster j we compare each $\log(\pi_k\varphi_k)$ with $\log\varphi_j^*$. Expanding $\|\boldsymbol{x}-\boldsymbol{\mu}_k\|^2 = \|\boldsymbol{x}-\boldsymbol{\mu}_j^*\|^2 - 2\langle\boldsymbol{x}-\boldsymbol{\mu}_j^*, \boldsymbol{h}_k\rangle + \|\boldsymbol{h}_k\|^2$ handles the $\boldsymbol{x}$ part, and substituting $y = \boldsymbol{x}^\top\boldsymbol{\theta}_j^* + \epsilon$ gives $r_k = \epsilon - a_k$, hence $-\frac{1}{2}r_k^2 + \frac{1}{2}\epsilon^2 = \epsilon a_k - \frac{1}{2}a_k^2$ for the y part. Therefore $\log(\pi_k\varphi_k) - \log\varphi_j^* = \log\pi_k + f_k$ with

$$f_k(\boldsymbol{z}) = \big\langle\boldsymbol{h}_k, \boldsymbol{x}-\boldsymbol{\mu}_j^*\big\rangle - \tfrac{1}{2}\|\boldsymbol{h}_k\|^2 + \epsilon a_k - \tfrac{1}{2}a_k^2, \qquad \nabla_{\boldsymbol{x}} f_k = \boldsymbol{h}_k + r_k\beta_k, \qquad \partial_\epsilon f_k = a_k, \tag{23}$$

where $\nabla_{\boldsymbol{x}}\big(\epsilon a_k - \frac{1}{2}a_k^2\big) = (\epsilon - a_k)\beta_k = r_k\beta_k$. Since $\psi_i = e^{\log\pi_i + f_i}\big/\sum_k e^{\log\pi_k + f_k}$ is a softmax in the f_k, the softmax Jacobian gives

$$\nabla_{\boldsymbol{x}}\psi_i = \psi_i\Big(\nabla_{\boldsymbol{x}} f_i - \sum_k \psi_k\nabla_{\boldsymbol{x}} f_k\Big) = \psi_i\big(\boldsymbol{h}_i + r_i\beta_i - \bar{\boldsymbol{h}} - \bar{\boldsymbol{w}}\big), \qquad \partial_\epsilon\psi_i = \psi_i(a_i - \bar{a}). \tag{24}$$

For (21), split $\mathbf{E}_j[\psi_i(\boldsymbol{\mu}_i - \boldsymbol{x})] = \mathbf{E}_j[\psi_i]\boldsymbol{h}_i - \mathbf{E}_j[\psi_i(\boldsymbol{x}-\boldsymbol{\mu}_j^*)]$ and apply Lemma A.3 in $\boldsymbol{Z} = \boldsymbol{x}-\boldsymbol{\mu}_j^*$ componentwise, so that $\mathbf{E}_j[\psi_i(\boldsymbol{x}-\boldsymbol{\mu}_j^*)] = \mathbf{E}_j[\nabla_{\boldsymbol{x}}\psi_i]$; inserting (24) cancels the terms $\psi_i\boldsymbol{h}_i$ and adds $\psi_i(\bar{\boldsymbol{h}} + \bar{\boldsymbol{w}} - r_i\beta_i)$, which is (21). For (22), write $-\mathbf{E}_j[\psi_i r_i\boldsymbol{x}] = -\mathbf{E}_j[\psi_i\epsilon\boldsymbol{x}] + \mathbf{E}_j[\psi_i a_i\boldsymbol{x}]$ and apply Lemma A.3 in ϵ with $\boldsymbol{x}$ held fixed: $\mathbf{E}_j[\psi_i\epsilon\boldsymbol{x}] = \mathbf{E}_j[\boldsymbol{x}\,\partial_\epsilon\psi_i] = \mathbf{E}_j[\psi_i(a_i - \bar{a})\boldsymbol{x}]$. The terms $\psi_i a_i\boldsymbol{x}$ cancel and $-\mathbf{E}_j[\psi_i r_i\boldsymbol{x}] = \mathbf{E}_j[\psi_i\bar{a}\,\boldsymbol{x}] = \mathbf{E}_j[\psi_i\boldsymbol{x}\boldsymbol{x}^\top\bar{\beta}]$. □

A.5 THE WEIGHT SUBPROBLEM AND THE DUAL GAP

For fixed $(\boldsymbol{\mu}, \boldsymbol{\theta})$ the map $\boldsymbol{\pi} \mapsto \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta})$ is convex on the simplex $\mathcal{P}$, since p is linear in $\boldsymbol{\pi}$ and $-\log$ is convex. Its suboptimality is measured by the simplex dual gap

$$G_\pi(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) := \max_{\tilde{\boldsymbol{\pi}} \in \mathcal{P}} \langle \nabla_{\boldsymbol{\pi}} \mathcal{L}, \boldsymbol{\pi} - \tilde{\boldsymbol{\pi}} \rangle = \max_i \big[-(\nabla_{\pi_i} \mathcal{L} + 1) \big], \tag{25}$$

where the second equality uses $\sum_i \pi_i \nabla_{\pi_i} \mathcal{L} = -\mathbf{E}_{p_*}[p/p] = -1$ and the fact that a linear function on $\mathcal{P}$ is maximized at a vertex. A weight update is ε_π-*optimal* if $G_\pi(\boldsymbol{\pi}^+, \boldsymbol{\mu}, \boldsymbol{\theta}) \le \varepsilon_\pi$ and $\mathcal{L}(\boldsymbol{\pi}^+, \boldsymbol{\mu}, \boldsymbol{\theta}) \le \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta})$. Both properties hold for the exact minimizer, and can be enforced for any monotone convex solver, e.g. the EM fixed point $\pi_i \leftarrow \mathbf{E}_{p_*} \psi_i$, which is loss non-increasing and converges to the minimizer.

Lemma A.4 (Consequences of the dual gap)**.** *Fix $(\boldsymbol{\mu}, \boldsymbol{\theta})$ and let $\boldsymbol{\pi} \in \mathcal{P}$ satisfy $G_\pi(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) \le \varepsilon_\pi$. Then*

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) \le \inf_{\tilde{\boldsymbol{\pi}} \in \mathcal{P}} \mathcal{L}(\tilde{\boldsymbol{\pi}}, \boldsymbol{\mu}, \boldsymbol{\theta}) + \varepsilon_\pi \qquad \textit{and} \qquad \sum_{i=1}^n \big| \pi_i - \mathbf{E}_{p_*} \psi_i \big| \le 2\varepsilon_\pi.$$

Proof. Write $g_i := \nabla_{\pi_i} \mathcal{L}$ and $\tilde{g}_i := g_i + 1$, so that $\tilde{g}_i \ge -\varepsilon_\pi$ for every i by (25). Convexity of $\mathcal{L}$ in $\boldsymbol{\pi}$ gives, for any $\tilde{\boldsymbol{\pi}} \in \mathcal{P}$, $\mathcal{L}(\boldsymbol{\pi}) - \mathcal{L}(\tilde{\boldsymbol{\pi}}) \le \langle \nabla_{\boldsymbol{\pi}} \mathcal{L}, \boldsymbol{\pi} - \tilde{\boldsymbol{\pi}} \rangle \le \varepsilon_\pi$, which is the first claim. For the second, $\sum_i \pi_i \tilde{g}_i = \sum_i \pi_i g_i + 1 = 0$, so the positive and negative parts of $(\tilde{g}_i)_i$ have equal $\boldsymbol{\pi}$-mass; each is at most ε_π because $\pi_i \tilde{g}_i^- \le \varepsilon_\pi \pi_i$ summed over i is at most ε_π. Hence $\sum_i \pi_i |\tilde{g}_i| \le 2\varepsilon_\pi$, and the claim follows from $\pi_i \tilde{g}_i = \pi_i + \pi_i \nabla_{\pi_i} \mathcal{L} = \pi_i - \mathbf{E}_{p_*}[\pi_i \varphi_i / p] = \pi_i - \mathbf{E}_{p_*} \psi_i$. □

B INITIALIZATION AND GLOBAL CONVERGENCE

Throughout this appendix $S_\ell, \hat{\pi}_\ell, U_\mu, U_\theta, U$ are as in (4), evaluated at the current iterate, and we freely use (17), (18) and Lemma A.1.

B.1 INITIALIZATION: PROOF OF LEMMA 4.1

Lemma B.1 (Initialization; formal version of Lemma 4.1)**.** *Let Assumption 1 hold, fix $\delta \in (0,1)$, suppose $n \ge \frac{2}{\pi_{\min}} \log \frac{2m}{\delta}$, and run Algorithm 1 with $\rho_0 = C\sqrt{d + \log(4n/\delta)}$ and $R = \rho_0 + 2\sqrt{d}$; assume $\Delta \ge 4(R + \rho_0)$, which is implied by* (15)*. Write J_i for the cluster from which $\boldsymbol{\mu}_i^{(0)}$ was drawn. Then with probability at least $1 - \delta$:*

1. *for each $\ell \in [m]$ there is an i with $J_i = \ell$ and $\left\| \boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}_\ell^* \right\| \le \sqrt{2d}$;*
2. *for every i, $\left\| \boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}_{J_i}^* \right\| \le \rho_0$; consequently $i \in S_{J_i}$ and $U_\mu(\boldsymbol{\mu}^{(0)}) \le n\rho_0^2$;*
3. (least-squares warm start) *for every i,*

$$\left\| \boldsymbol{\theta}_i^{(0)} - \boldsymbol{\theta}_{J_i}^* \right\| \le \frac{C\Lambda\, D_\theta}{\pi_{\min}}\, e^{-\Delta^2/16}, \qquad \textit{hence} \qquad U_\theta(\boldsymbol{\mu}^{(0)}, \boldsymbol{\theta}^{(0)}) \le n \Big(\frac{C\Lambda D_\theta}{\pi_{\min}} \Big)^2 e^{-\Delta^2/8};$$

4. *if in addition $\Lambda\, U_\theta(\boldsymbol{\mu}^{(0)}, \boldsymbol{\theta}^{(0)}) \le 1$, then $\mathcal{L}^{(0)} \le \log n + \frac{3}{2}(d+1) \le L_0$.*

Proof. Write the gate initialization as: draw $J_i \in [m]$ with probabilities $\boldsymbol{\pi}^*$ and set $\boldsymbol{\mu}_i^{(0)} = \boldsymbol{\mu}_{J_i}^* + \boldsymbol{Z}_i$ with $\boldsymbol{Z}_i \sim N(0, I_d)$, independently over i. This is exactly sampling $\boldsymbol{\mu}_i^{(0)}$ from the $\boldsymbol{x}$-marginal $p_*^{\boldsymbol{x}}$.

Item 1 (coupon collector). Let $p := \Pr\{\|\boldsymbol{Z}\|^2 \le 2d\}$. Since $\mathbf{E}\,\|\boldsymbol{Z}\|^2 = d$, Markov's inequality gives $\Pr\{\|\boldsymbol{Z}\|^2 > 2d\} \le \frac{1}{2}$, so $p \ge \frac{1}{2}$. The events $\{J_i = \ell\}$ and $\{\|\boldsymbol{Z}_i\|^2 \le 2d\}$ are independent, so a single draw succeeds for cluster ℓ with probability $\pi_\ell^* p$, and by independence across i the probability that no draw succeeds is

$$(1 - \pi_\ell^* p)^n \;\le\; \Big(1 - \tfrac{\pi_{\min}}{2}\Big)^n \;\le\; e^{-n\pi_{\min}/2} \;\le\; \frac{\delta}{2m},$$

using $1 - u \le e^{-u}$ and $n \ge \frac{2}{\pi_{\min}} \log \frac{2m}{\delta}$. A union bound over the m clusters bounds the failure probability by $\delta/2$. On the complementary event, $\left\| \boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}_\ell^* \right\| = \|\boldsymbol{Z}_i\| \le \sqrt{2d}$ for the selected i.

Item 2 (concentration). The map $\boldsymbol{Z} \mapsto \|\boldsymbol{Z}\|$ is 1-Lipschitz, so Gaussian concentration gives $\Pr\{\|\boldsymbol{Z}_i\| > \sqrt{d} + t\} \le e^{-t^2/2}$. Taking $t = \sqrt{2\log(4n/\delta)}$ makes each failure probability $\delta/(4n)$, and a union bound over i leaves probability at least $1 - \delta/4$ that

$$\|\boldsymbol{Z}_i\| \;\le\; \sqrt{d} + \sqrt{2\log(4n/\delta)} \;\le\; \sqrt{3}\,\sqrt{d + \log(4n/\delta)} \;\le\; \rho_0 \qquad \text{for all } i,$$

where the middle step is Cauchy–Schwarz, $\sqrt{a}+\sqrt{2b} \le \sqrt{3}\sqrt{a+b}$, and the last step holds for $C \ge \sqrt{3}$ in the definition of ρ_0. By Lemma A.1(iv) and $\Delta \ge 4(R+\rho_0) > 4\rho_0$ we get $\left\|\boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}^*_{J_i}\right\| \le \rho_0 < \Delta/4 < \Delta/2$; since distinct true centers are at distance at least Δ, $\boldsymbol{\mu}^*_{J_i}$ is the strictly nearest true center, i.e. $i \in S_{J_i}$. Summing, $U_\mu(\boldsymbol{\mu}^{(0)}) \le \sum_i \|\boldsymbol{Z}_i\|^2 \le n\rho_0^2$. Items 1 and 2 hold simultaneously with probability at least $1 - \frac{3}{4}\delta \ge 1 - \delta$.

Item 3 (least-squares warm start). Fix i, write $\ell = J_i$ and $w(\boldsymbol{x}) := \mathbb{1}\{\left\|\boldsymbol{x} - \boldsymbol{\mu}_i^{(0)}\right\| \le R\}$. Since $R = \rho_0 + 2\sqrt{d}$ and $\left\|\boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}^*_\ell\right\| \le \rho_0$, the triangle inequality gives the nesting

$$B(\boldsymbol{\mu}^*_\ell, 2\sqrt{d}) \;\subset\; B(\boldsymbol{\mu}_i^{(0)}, R) \;\subset\; B(\boldsymbol{\mu}^*_\ell, R + \rho_0),$$

and, for $j \ne \ell$, every $\boldsymbol{x}$ in the window satisfies $\left\|\boldsymbol{x} - \boldsymbol{\mu}^*_j\right\| \ge \Delta - (R + \rho_0) \ge \frac{3}{4}\Delta$ because $\Delta \ge 4(R+\rho_0)$.

The objective in (6) is the convex quadratic $\boldsymbol{\theta} \mapsto \boldsymbol{\theta}^\top M \boldsymbol{\theta} - 2\boldsymbol{b}^\top \boldsymbol{\theta} + \text{const}$ with

$$M := \mathbf{E}_{p_*}[w(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top] = \sum_j \pi_j^* M_j, \qquad M_j := \mathbf{E}_j[w(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top],$$

$$\boldsymbol{b} := \mathbf{E}_{p_*}[w(\boldsymbol{x})\boldsymbol{x}y] = \sum_j \pi_j^* M_j \boldsymbol{\theta}_j^*,$$

the last identity because under cluster j we have $y = \boldsymbol{x}^\top \boldsymbol{\theta}_j^* + \epsilon$ with $\mathbf{E}[\epsilon \mid \boldsymbol{x}] = 0$, so the noise cross term vanishes. Once $M \succ 0$ is established, $\boldsymbol{\theta}_i^{(0)} = M^{-1}\boldsymbol{b}$, and subtracting $\boldsymbol{\theta}_\ell^* = M^{-1}M\boldsymbol{\theta}_\ell^* = M^{-1}\sum_j \pi_j^* M_j \boldsymbol{\theta}_\ell^*$ gives

$$\boldsymbol{\theta}_i^{(0)} - \boldsymbol{\theta}_\ell^* = M^{-1} \sum_{j \ne \ell} \pi_j^* M_j (\boldsymbol{\theta}_j^* - \boldsymbol{\theta}_\ell^*). \tag{26}$$

Lower bound on M. By the inner inclusion, $w(\boldsymbol{x}) \ge \mathbb{1}\{\|\boldsymbol{Z}\| \le 2\sqrt{d}\}$ for $\boldsymbol{x} = \boldsymbol{\mu}^*_\ell + \boldsymbol{Z}$. For a unit vector $\boldsymbol{u}$, expanding $\langle \boldsymbol{u}, \boldsymbol{\mu}^*_\ell + \boldsymbol{Z}\rangle^2 = \langle \boldsymbol{u}, \boldsymbol{\mu}^*_\ell\rangle^2 + 2\langle \boldsymbol{u}, \boldsymbol{\mu}^*_\ell\rangle\langle \boldsymbol{u}, \boldsymbol{Z}\rangle + \langle \boldsymbol{u}, \boldsymbol{Z}\rangle^2$ and noting that the domain $\{\|\boldsymbol{Z}\| \le 2\sqrt{d}\}$ is invariant under $\boldsymbol{Z} \mapsto -\boldsymbol{Z}$ (so the odd cross term integrates to 0) yields

$$\begin{aligned}\boldsymbol{u}^\top M_\ell \boldsymbol{u} \;&\ge\; \langle \boldsymbol{u}, \boldsymbol{\mu}^*_\ell\rangle^2 \Pr\{\|\boldsymbol{Z}\| \le 2\sqrt{d}\} + \mathbf{E}\big[\mathbb{1}\{\|\boldsymbol{Z}\| \le 2\sqrt{d}\}\,\langle \boldsymbol{u}, \boldsymbol{Z}\rangle^2\big] \\ &\ge\; \mathbf{E}\big[\mathbb{1}\{\|\boldsymbol{Z}\| \le 2\sqrt{d}\} Z_1^2\big] =: c_1,\end{aligned}$$

where the last equality uses rotational invariance to take $\boldsymbol{u} = \boldsymbol{e}_1$. By the law of total expectation and Cauchy–Schwarz with $\mathbf{E}Z_1^4 = 3$,

$$c_1 = \mathbf{E}[Z_1^2] - \mathbf{E}\big[Z_1^2 \mathbb{1}\{\|\boldsymbol{Z}\| > 2\sqrt{d}\}\big] \;\ge\; 1 - \sqrt{3\Pr\{\|\boldsymbol{Z}\| > 2\sqrt{d}\}}.$$

For $d \ge 6$, $\Pr\{\|\boldsymbol{Z}\| > 2\sqrt{d}\} \le e^{-d/2}$ gives $c_1 > \frac{1}{2}$; for $d \le 5$, c_1 is the integral of a strictly positive function over a set of positive measure, hence a fixed positive number. In either case c_1 is a universal positive constant and, since all $M_j \succeq 0$, $M \succeq \pi_\ell^* M_\ell \succeq c_1 \pi_{\min} I_d$.

Upper bound on M_j, $j \ne \ell$. Since $M_j \succeq 0$, its operator norm is at most its trace, so by Cauchy–Schwarz, $w^2 = w$ and Lemma A.2,

$$\begin{aligned}\|M_j\|_{\text{op}} &\le \mathbf{E}_j\big[w(\boldsymbol{x})\,\|\boldsymbol{x}\|^2\big] \le \sqrt{\mathbf{E}_j\,\|\boldsymbol{x}\|^4}\,\sqrt{\Pr_j\{\left\|\boldsymbol{x} - \boldsymbol{\mu}_i^{(0)}\right\| \le R\}} \\ &\le C\Lambda\sqrt{\Pr\{\|\boldsymbol{Z}\| \ge \tfrac{3}{4}\Delta\}} \;\le\; C\Lambda e^{-\Delta^2/16}.\end{aligned}$$

The last step uses $\Pr\{\|\boldsymbol{Z}\| \geq \sqrt{d} + t\} \leq e^{-t^2/2}$ with $t = \frac{3}{4}\Delta - \sqrt{d} \geq \frac{1}{2}\Delta$, which holds because $\Delta \geq 4R \geq 8\sqrt{d}$.

Substituting both bounds into (26), using $\left\|\boldsymbol{\theta}_j^* - \boldsymbol{\theta}_\ell^*\right\| \leq D_\theta$ and $\sum_{j \neq \ell} \pi_j^* \leq 1$,

$$\left\|\boldsymbol{\theta}_i^{(0)} - \boldsymbol{\theta}_\ell^*\right\| \;\leq\; \frac{1}{c_1 \pi_{\min}} \cdot C\Lambda e^{-\Delta^2/16} \cdot D_\theta,$$

and squaring and summing over the n components bounds U_θ as claimed.

Item 4 (initial loss). Fix ℓ and let i be the component supplied by item 1, so that $\left\|\boldsymbol{\mu}_i^{(0)} - \boldsymbol{\mu}_\ell^*\right\|^2 \leq 2d$ and $\left\|\boldsymbol{\theta}_i^{(0)} - \boldsymbol{\theta}_\ell^*\right\|^2 \leq U_\theta^{(0)}$. Since $p_* \leq (2\pi)^{-(d+1)/2}$ pointwise (a mixture of unit-covariance Gaussian densities) and $p^{(0)} \geq \frac{1}{n}\varphi_i$, the normalizing constants cancel and

$$\log \frac{p_*(\boldsymbol{z})}{p^{(0)}(\boldsymbol{z})} \;\leq\; \log n + \tfrac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{\mu}_i^{(0)}\right\|^2 + \tfrac{1}{2} r_i(\boldsymbol{z})^2.$$

Under cluster ℓ, the bias–variance decomposition gives $\mathbf{E}_\ell \left\|\boldsymbol{x} - \boldsymbol{\mu}_i^{(0)}\right\|^2 = d + \left\|\boldsymbol{\mu}_\ell^* - \boldsymbol{\mu}_i^{(0)}\right\|^2 \leq 3d$, and $r_i = \epsilon + \boldsymbol{x}^\top(\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_i^{(0)})$ with $\mathbf{E}[\epsilon \mid \boldsymbol{x}] = 0$ gives, by Lemma A.2,

$$\mathbf{E}_\ell r_i^2 = 1 + \mathbf{E}_\ell\big[(\boldsymbol{x}^\top(\boldsymbol{\theta}_\ell^* - \boldsymbol{\theta}_i^{(0)}))^2\big] \leq 1 + 2\Lambda\, U_\theta^{(0)} \leq 3.$$

Averaging over ℓ with weights π_ℓ^* gives $\mathcal{L}^{(0)} \leq \log n + \frac{3}{2}d + \frac{3}{2}$, which is at most L_0 by (11). □

Remark 1. *Combining item 3 with* (16) *and* $\varepsilon \leq c\,\pi_{\min} \leq 1$, $B \geq 2$,

$$U_\theta(\boldsymbol{\mu}^{(0)}, \boldsymbol{\theta}^{(0)}) \;\leq\; \mathsf{A}_{\mathrm{g}} e^{-\Delta^2/8} \;\leq\; \frac{B_\theta^2}{4}, \tag{27}$$

so the hypothesis $\Lambda U_\theta^{(0)} \leq 1$ *of item 4 holds by Lemma A.1(ii), and the expert potential starts at a quarter of its budget.*

B.2 The global-convergence theorem

Theorem B.2 (Global convergence of gradient EM; formal version of Theorem 4.2)**.** *Let Assumption 1 and the separation condition* (15) *hold, work on the event of Lemma B.1, and let* $L_0, B, B_\theta, A, A_\eta, A_\pi$ *be as in* (11)–(13)*. Fix a target accuracy* ε *with* $e^{-c\,dn} \leq \varepsilon \leq c\,\pi_{\min}$*. Assume each weight update is loss non-increasing and that every post-weight point satisfies the dual-gap condition of Appendix A.5,*

$$G_\pi\big(\boldsymbol{\pi}^{(t+1)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}\big) \leq \varepsilon_\pi, \qquad \varepsilon_\pi \leq \frac{\varepsilon}{A_\pi} = \frac{c\,\varepsilon}{\mathsf{A}_{\mathrm{g}}}. \tag{28}$$

If $\eta \leq c/A_\eta$*, then*

$$\mathcal{L}\big(\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}\big) \leq \varepsilon \qquad \textit{for some } t \leq T := \frac{CB^2}{\eta\varepsilon},$$

and for every $t \leq T$ *before the target is reached,*

$$U(\boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}) \leq \frac{B^2}{2}, \qquad U_\theta(\boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)}) \leq B_\theta^2, \qquad \textit{and the groups } S_\ell \textit{ do not change.}$$

For small ε we have $A_\eta = L_0 B^2/(c_\theta^2 \varepsilon^2)$ by (12)–(13), so the largest admissible step size $\eta = c/A_\eta$ gives $T = CB^2 A_\eta/(c\varepsilon) = O(L_0 B^4/\varepsilon^3)$: the iteration count is polynomial in $(d, m, n, 1/\pi_{\min}, \log(1/\delta), D_{\max}, D_\theta)$ and of order $1/\varepsilon^3$. The lower restriction $\varepsilon \geq e^{-c\,dn}$ is exactly what makes (15) compatible with Assumption 1, whose separation is at least of order $\sqrt{dn}$.

The proof combines a one-step statement (Proposition B.4 below) with the following control on the accumulated drift of the expert potential.

Lemma B.3 (Expert-potential drift along the trajectory)**.** *Let $(\boldsymbol{\pi}^{(s)}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\theta}^{(s)})_{s\le t+1}$ be iterates of Algorithm 1 such that the hypotheses of Proposition B.4 hold at every post-weight point up to time t. Then*

$$U_\theta^{(t+1)} \;\le\; U_\theta^{(0)} + 2\eta(t+1)\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + 2\eta\big(L_0 - \mathcal{L}^{(t+1)}\big).$$

Proof. Write $\mathcal{L}^{(s+\frac{1}{2})} := \mathcal{L}(\boldsymbol{\pi}^{(s+1)}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\theta}^{(s)})$ for the post-weight loss. Since the weight step does not move $(\boldsymbol{\mu}, \boldsymbol{\theta})$ and is loss non-increasing, $\mathcal{L}^{(s+\frac{1}{2})} \le \mathcal{L}^{(s)}$. The third inequality of Proposition B.4 and the descent inequality $\mathcal{L}^{(s+1)} \le \mathcal{L}^{(s+\frac{1}{2})} - \frac{\eta}{2}\left\|\nabla\mathcal{L}^{(s+\frac{1}{2})}\right\|_F^2$ of Lemma D.1 (applicable under those hypotheses, as verified in Appendix B.3) give, for every $s \le t$,

$$\begin{aligned} U_\theta^{(s+1)} - U_\theta^{(s)} &\le 2\eta\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + \eta^2\left\|\nabla\mathcal{L}^{(s+\frac{1}{2})}\right\|_F^2 \\ &\le 2\eta\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + 2\eta\big(\mathcal{L}^{(s+\frac{1}{2})} - \mathcal{L}^{(s+1)}\big) \;\le\; 2\eta\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + 2\eta\big(\mathcal{L}^{(s)} - \mathcal{L}^{(s+1)}\big), \end{aligned}$$

where the middle step multiplies the descent inequality by 2η. Summing over $s = 0, \dots, t$ telescopes both the left-hand side and the loss differences, and $\mathcal{L}^{(0)} \le L_0$ gives the claim. The potential U_θ is evaluated with the groups S_ℓ, which do not change along the trajectory by Proposition E.2, so the telescoping is legitimate: no term in the sum is affected by a reassignment. □

Proof of Theorem B.2. Write $L_t := \mathcal{L}(\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)})$, $L_{t+\frac{1}{2}} := \mathcal{L}(\boldsymbol{\pi}^{(t+1)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)})$, $U_t := U(\boldsymbol{\mu}^{(t)}, \boldsymbol{\theta}^{(t)})$ and $U_{\theta,t}$ likewise, and let τ be the first time at which the loss is at most ε. We prove by induction on $t < \tau$, $t \le T$, that

$$L_t \le L_0, \qquad U_t \le \frac{B^2}{2}, \qquad U_{\theta,t} \le B_\theta^2, \qquad S_\ell(\boldsymbol{\mu}^{(t)}) = S_\ell(\boldsymbol{\mu}^{(0)}) \;\; \forall \ell. \tag{29}$$

Base case. At $t = 0$ the first two statements are the initialization event together with the choices (11), and the third is the sharper bound $U_{\theta,0} \le B_\theta^2/4$ of (27).

Induction step. Suppose (29) holds at time $t < \tau$. The weight step leaves $(\boldsymbol{\mu}, \boldsymbol{\theta})$ and hence U, U_θ, S_ℓ unchanged, and is loss non-increasing, so $L_{t+\frac{1}{2}} \le L_t$. If $L_{t+\frac{1}{2}} \le \varepsilon$ we are done. Otherwise $\varepsilon < L_{t+\frac{1}{2}} \le L_0$, so Proposition B.4 applies at the post-weight point and gives $U_{t+1} \le U_t \le B^2/2$ and $L_{t+1} \le L_{t+\frac{1}{2}} - \frac{c\eta}{B^2} L_{t+\frac{1}{2}}^2 \le L_0$; the groups are unchanged by Proposition E.2. For the expert potential, Lemma B.3 together with (27) gives

$$U_{\theta,t+1} \;\le\; \frac{B_\theta^2}{4} + 2\eta(t+1)\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + 2\eta L_0 \;\le\; \frac{B_\theta^2}{4} + \frac{2CB^2}{\varepsilon}\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + \frac{B_\theta^2}{4} \;\le\; B_\theta^2,$$

where the middle step uses $t + 1 \le T = CB^2/(\eta\varepsilon)$, so that the factor η cancels, together with $\eta \le c/A_\eta \le c\,B_\theta^2/L_0 \le B_\theta^2/(8L_0)$ for $c \le \frac{1}{8}$; and the last step uses (16) in the form $\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} \le \varepsilon B_\theta^2/(8CB^2)$. This completes the induction.

Recurrence. We convert the per-step decrease into a recurrence for L_t. If $L_{t+\frac{1}{2}} \ge L_t/2$, then substituting this lower bound into the quadratic term and $L_{t+\frac{1}{2}} \le L_t$ into the linear term gives $L_{t+1} \le L_t - \frac{c\eta}{4B^2} L_t^2$. If instead $L_{t+\frac{1}{2}} < L_t/2$, then $L_{t+1} \le L_{t+\frac{1}{2}} < L_t/2 \le L_t - \frac{c\eta}{B^2} L_t^2$, the last inequality being equivalent to $\frac{c\eta}{B^2} L_t \le \frac{1}{2}$, which holds because $\eta \le c/A$, $L_t \le L_0 \le A$ by (14) and $B^2 \ge 1$, so that $\frac{c\eta L_t}{B^2} \le c^2 \frac{L_0}{A} \frac{1}{B^2} \le c^2 \le \frac{1}{2}$. Absorbing the factor $\frac{1}{4}$ into c, in both cases

$$L_{t+1} \le L_t - \frac{c\eta}{B^2} L_t^2 \qquad \Longrightarrow \qquad \frac{1}{L_{t+1}} - \frac{1}{L_t} = \frac{L_t - L_{t+1}}{L_t L_{t+1}} \;\ge\; \frac{L_t - L_{t+1}}{L_t^2} \;\ge\; \frac{c\eta}{B^2},$$

using $L_{t+1} \le L_t$ (if $L_{t+1} = 0$ the target is already reached). Summing, $1/L_t \ge 1/L_0 + c\eta t/B^2$, hence $L_t \le B^2/(c\eta t) \le \varepsilon$ as soon as $t \ge B^2/(c\eta\varepsilon)$. Therefore $\tau \le T = CB^2/(\eta\varepsilon)$, which is within the range of the induction, and (29) holds up to τ. □

B.3 One global step

Proposition B.4 (One global step)**.** *Under the assumptions of Theorem B.2, consider a post-weight point* $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta})$ *with*

$$\varepsilon \le \mathcal{L} := \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) \le L_0, \qquad U(\boldsymbol{\mu}, \boldsymbol{\theta}) \le \frac{B^2}{2}, \qquad U_\theta(\boldsymbol{\mu}, \boldsymbol{\theta}) \le B_\theta^2,$$

satisfying (28)*, and let* $(\boldsymbol{\mu}^+, \boldsymbol{\theta}^+) := (\boldsymbol{\mu}, \boldsymbol{\theta}) - \eta \nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta})$*. If* $\eta \le c/A$*, then*

$$U(\boldsymbol{\mu}^+, \boldsymbol{\theta}^+) \le U(\boldsymbol{\mu}, \boldsymbol{\theta}), \qquad \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}^+, \boldsymbol{\theta}^+) \le \mathcal{L} - \frac{c\eta}{B^2}\mathcal{L}^2,$$
$$U_\theta(\boldsymbol{\mu}^+, \boldsymbol{\theta}^+) \le U_\theta(\boldsymbol{\mu}, \boldsymbol{\theta}) + 2\eta \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} + \eta^2 \left\| \nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) \right\|_F^2 .$$

Proof. Write $g := \nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) = (g^\mu, g^\theta)$ and let $P = P_\mu + P_\theta$ be the projection (31) defined below.

Verification of the smoothness hypotheses. By Lemma C.2, $\hat{\pi}_\ell \ge \pi_\ell^*/2$ for every ℓ, so Lemma D.1 applies with $c_{\mathrm{cov}} = \frac{1}{2}$. By (17), $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| \le B$ and $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*\| \le B_\theta \le 1$ for $i \in S_\ell$, hence

$$\frac{1}{\hat{\pi}_\ell} \sum_{i \in S_\ell} \pi_i \big(\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\|^2 + \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*\|^2 \big) \le B^2 + B_\theta^2 \le 2B^2,$$

and the factor 2 is absorbed into the constant C inside A. Lemma D.1 therefore gives

$$\|g\|_F^2 \le 2A\mathcal{L}, \qquad \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu} - \eta g^\mu, \boldsymbol{\theta} - \eta g^\theta) \le \mathcal{L} - \frac{\eta}{2} \|g\|_F^2 \qquad (\eta \le c/A). \tag{30}$$

Total potential. By Proposition E.2 the groups do not change, so U may be evaluated at the new point with the same assignment. Expanding each squared distance, $\|(\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*) - \eta g_i^\mu\|^2 = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\|^2 - 2\eta \langle g_i^\mu, \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^* \rangle + \eta^2 \|g_i^\mu\|^2$ and likewise for $\boldsymbol{\theta}$, and summing over $i \in S_\ell$ and ℓ,

$$U(\boldsymbol{\mu}^+, \boldsymbol{\theta}^+) = U - 2\eta P + \eta^2 \|g\|_F^2 \;\le\; U - 2c\eta\mathcal{L} + 2\eta^2 A\mathcal{L} \;=\; U - 2\eta\mathcal{L}\,(c - \eta A) \;\le\; U,$$

by Lemma B.6 ($P \ge c\mathcal{L}$), (30) and $\eta \le c/A$.

Loss decrease. The projection is the inner product of g with the displacement vector, whose squared norm is exactly U, so Cauchy–Schwarz gives $P \le \|g\|_F \sqrt{U} \le \|g\|_F B/\sqrt{2}$, i.e. $\|g\|_F \ge \sqrt{2}\, P/B \ge c\,\mathcal{L}/B$ by Lemma B.6. Inserting this into the descent inequality of (30),

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}^+, \boldsymbol{\theta}^+) \;\le\; \mathcal{L} - \frac{\eta}{2} \|g\|_F^2 \;\le\; \mathcal{L} - \frac{c\eta}{B^2}\mathcal{L}^2.$$

Expert potential. Exactly the same expansion restricted to the $\boldsymbol{\theta}$-block gives $U_\theta(\boldsymbol{\mu}^+, \boldsymbol{\theta}^+) = U_\theta - 2\eta P_\theta + \eta^2 \left\|g^\theta\right\|_F^2$. The second part of Lemma B.6 bounds $-2\eta P_\theta \le 2\eta \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$, and $\left\|g^\theta\right\|_F^2 \le \|g\|_F^2$. □

B.4 The descent direction

Recall the projection onto the "move every gate and every expert towards the truth of its group" direction,

$$P := \sum_{\ell \in [m]} \sum_{i \in S_\ell} \Big(\langle \nabla_{\boldsymbol{\mu}_i} \mathcal{L}, \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^* \rangle + \langle \nabla_{\boldsymbol{\theta}_i} \mathcal{L}, \boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^* \rangle \Big) =: P_\mu + P_\theta. \tag{31}$$

For $i \in S_\ell$ define the integrands

$$T_i(\boldsymbol{z}) := \langle \boldsymbol{\mu}_i - \boldsymbol{x}, \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^* \rangle - r_i(\boldsymbol{z})\, \boldsymbol{x}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*), \qquad T_i^\theta(\boldsymbol{z}) := -r_i(\boldsymbol{z})\, \boldsymbol{x}^\top (\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*), \tag{32}$$

so that, by (20), $P = \mathbf{E}_{p_*}\big[\sum_i \psi_i T_i\big]$ and $P_\theta = \mathbf{E}_{p_*}\big[\sum_i \psi_i T_i^\theta\big]$. The *one-cluster projection* of cluster ℓ replaces the global posterior by the within-group posterior and the mixture p_* by the single component φ_ℓ^*:

$$P_\ell := \mathbf{E}_\ell\Big[\sum_{i \in S_\ell} \psi_i^{(\ell)}(\boldsymbol{z})\, T_i(\boldsymbol{z})\Big], \qquad P_{\ell,\theta} := \mathbf{E}_\ell\Big[\sum_{i \in S_\ell} \psi_i^{(\ell)}(\boldsymbol{z})\, T_i^\theta(\boldsymbol{z})\Big]. \tag{33}$$

In the recentered coordinates of cluster ℓ,

$$\begin{aligned} \boldsymbol{x}' &:= \boldsymbol{x} - \boldsymbol{\mu}_\ell^*, \qquad y' := y - \boldsymbol{x}^\top \boldsymbol{\theta}_\ell^*, \\ \boldsymbol{h}_i &:= \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*, \qquad \beta_i := \boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*, \qquad a_i := \langle \boldsymbol{x}', \beta_i \rangle + \langle \boldsymbol{\mu}_\ell^*, \beta_i \rangle = \boldsymbol{x}^\top \beta_i, \end{aligned} \tag{34}$$

one has $r_i = y' - a_i$ and

$$T_i = \|\boldsymbol{h}_i\|^2 - \langle \boldsymbol{x}', \boldsymbol{h}_i \rangle + a_i^2 - y' a_i, \qquad T_i^\theta = a_i^2 - y' a_i, \tag{35}$$

because $\langle \boldsymbol{\mu}_i - \boldsymbol{x}, \boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^* \rangle = \langle \boldsymbol{h}_i - \boldsymbol{x}', \boldsymbol{h}_i \rangle$ and $-r_i \boldsymbol{x}^\top \beta_i = -(y' - a_i) a_i$. Thus P_ℓ is exactly the quantity P of the abstract one-cluster problem of Appendix C.3.

The following two statements are the formal versions of the group-decoupling and descent lemmas stated in the body of the paper.

Lemma B.5 (Group decoupling; formal version of Lemma 4.6)**.** *Under the assumptions of Proposition B.4,*

$$P \;\geq\; \sum_{\ell \in [m]} \pi_\ell^* P_\ell - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}, \qquad P_\theta \;\geq\; \sum_{\ell \in [m]} \pi_\ell^* P_{\ell,\theta} - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}.$$

Lemma B.6 (Descent direction; formal version of Lemma 4.7)**.** *Under the assumptions of Proposition B.4, $P \geq c\,\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta})$ and $P_\theta \geq -\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$.*

Proof of Lemma B.5. A polynomial envelope. Fix a data cluster j and any component i, with $\ell = \ell(i)$, and recall $\epsilon = y - \boldsymbol{x}^\top \boldsymbol{\theta}_j^*$. By (17) and the triangle inequality, $\|\boldsymbol{x}\| \leq \|\boldsymbol{x} - \boldsymbol{\mu}_j^*\| + D_{\max}$, $\|\boldsymbol{\mu}_i - \boldsymbol{x}\| \leq \|\boldsymbol{x} - \boldsymbol{\mu}_j^*\| + 2D_{\max} + B$ and $|r_i| \leq |\epsilon| + \|\boldsymbol{x}\|\,(D_\theta + 1)$, since $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j^*\| \leq D_\theta + B_\theta \leq D_\theta + 1$. Each of T_i, T_i^θ is a product of at most three such factors, so

$$|T_i(\boldsymbol{z})| + |T_i^\theta(\boldsymbol{z})| \;\leq\; W_j(\boldsymbol{z}) := C(1+B)^2(1+D_{\max})^2(1+D_\theta)\big(1 + \|\boldsymbol{x} - \boldsymbol{\mu}_j^*\| + |\epsilon|\big)^3, \tag{36}$$

a weight of the form covered by Lemma C.1 with $K = 3$.

Reduction to same-group terms. Split $\mathbf{E}_{p_*} = \sum_j \pi_j^* \mathbf{E}_j$ and $\sum_i = \sum_{i \in S_j} + \sum_{i \notin S_j}$. For the foreign components, (36) and Lemma C.1(2) give

$$\Big| \mathbf{E}_j \Big[\sum_{i \notin S_j} \psi_i T_i \Big] \Big| \;\leq\; \mathbf{E}_j \Big[\sum_{i \notin S_j} \psi_i W_j \Big] \;\leq\; \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}.$$

For $i \in S_j$ we factor the global posterior through the within-group one:

$$\psi_i = \frac{\pi_i \varphi_i}{p} = \frac{\hat{\pi}_j q_j}{p} \cdot \frac{\lambda_i^{(j)} \varphi_i}{q_j} = \rho_j\, \psi_i^{(j)}, \qquad \rho_j = \frac{1}{1 + \chi_j} \in (0, 1],$$

with χ_j as in Lemma C.1(3). Hence

$$\mathbf{E}_j \Big[\sum_{i \in S_j} \psi_i T_i \Big] = P_j - \mathbf{E}_j \Big[(1 - \rho_j) \sum_{i \in S_j} \psi_i^{(j)} T_i \Big] \;\geq\; P_j - \mathbf{E}_j \big[(1 - \rho_j) W_j \big] \;\geq\; P_j - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2},$$

again by Lemma C.1(3). Multiplying by π_j^*, summing over j and using $\sum_j \pi_j^* = 1$ (the factor m and the constant 2 are absorbed into A_{g}) proves the first inequality. Since T_i^θ obeys the same envelope (36), the identical argument proves the second. □

Proof of Lemma B.6. The expert part. By Lemma B.5 and Proposition E.1, which states $P_{\ell,\theta} \geq 0$ for every ℓ, we may drop the (nonnegative) sum and obtain $P_\theta \geq -\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$.

One-cluster projection comparison. Apply Lemma C.4 with $\tau := c\varepsilon / B^2$, which lies in $(0, 1]$ because $\varepsilon \leq c\pi_{\min} \leq 1 \leq B^2$. The resulting deficit is

$$\frac{\tau}{2} B^2 + 6\Big(1 + \frac{1}{2\tau}\Big) B_\theta^2 = \frac{c\varepsilon}{2} + 6B_\theta^2 + \frac{3B^2 B_\theta^2}{c\,\varepsilon} \;\leq\; \frac{c\varepsilon}{2} + \Big(6c_\theta^2 + \frac{3c_\theta^2}{c}\Big)\varepsilon \;\leq\; c\,\varepsilon,$$

where we used $B_\theta \leq c_\theta \varepsilon / B$ from (12), so that $B_\theta^2 \leq c_\theta^2 \varepsilon^2 / B^2 \leq c_\theta^2 \varepsilon$, and then chose c_θ small enough given c. Writing $\mathcal{L}_\ell = D_{\mathrm{KL}}(\varphi_\ell^* \| q_\ell)$ and using the second part of Lemma C.4,

$$\sum_\ell \pi_\ell^* P_\ell \;\geq\; 2 \sum_\ell \pi_\ell^* \mathcal{L}_\ell - 2 \sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)} \| \lambda^{(\ell)}) - c\varepsilon \;\geq\; 2 \sum_\ell \pi_\ell^* \mathcal{L}_\ell - 2\varepsilon_\pi - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} - c\varepsilon.$$

Comparison with the loss. By Lemma C.5, $\sum_\ell \pi_\ell^* \mathcal{L}_\ell \geq \mathcal{L} - C\varepsilon_\pi - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$. Combining with Lemma B.5,

$$P \;\geq\; 2\mathcal{L} - C\varepsilon_\pi - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} - c\varepsilon \;\geq\; \mathcal{L},$$

because $\varepsilon_\pi \leq c\mathsf{A}_{\mathrm{g}}^{-1}\varepsilon$ by (28), $\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} \leq c\varepsilon$ by (16), and $\mathcal{L} \geq \varepsilon$, so that the three error terms together are at most $c\varepsilon \leq c\mathcal{L} \leq \mathcal{L}$. □

C Technical Lemmas for the Global Convergence

The inputs to this appendix are the Gaussian ratio bounds of Lemma D.4, proved in Appendix D, which we restate in the form used here. Under (17), (18) and (15), for every data cluster j, every $\ell \neq j$, $i \in S_\ell$, $k \in S_j$, $r \neq j$, and every polynomial weight $W_K(\boldsymbol{z}) = (1 + \|\boldsymbol{x} - \boldsymbol{\mu}_j^*\| + |\epsilon|)^K$ of fixed degree K,

$$\begin{gathered}\mathbf{E}_j\Big[\Big(\frac{\varphi_i}{\varphi_k}\Big)^{1/2} W_K\Big] \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}, \qquad \int \sqrt{\varphi_j^*}\,\varphi_i\,(\varphi_\ell^*)^{-1/2}\,d\boldsymbol{z} \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}, \\ \mathbf{E}_j\Big[\Big(\frac{\varphi_r^*}{\varphi_j^*}\Big)^{1/2}\Big] \leq e^{-\Delta^2/8}.\end{gathered} \tag{37}$$

C.1 Cross-group estimates

Lemma C.1 (Cross-group estimates; formal version of Lemma 4.3)**.** *Under the assumptions of Proposition B.4 the following hold for every $\ell \neq j$, every $i \in S_\ell$ and every polynomial weight $W(\boldsymbol{z}) = (1 + \|\boldsymbol{x} - \boldsymbol{\mu}_j^*\| + |\epsilon|)^K$ of fixed degree K, with $\epsilon = y - \boldsymbol{x}^\top \boldsymbol{\theta}_j^*$:*

1. $\hat{\pi}_\ell \geq \exp\big(-C\frac{L_0 + \log(1/\pi_{\min})}{\pi_{\min}}\big)$ *for every ℓ;*
2. $\mathbf{E}_j[\psi_i(\boldsymbol{z}) W(\boldsymbol{z})] \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$*, and consequently*
$$\mathbf{E}_\ell\Big[\sum_{i \notin S_\ell} \psi_i W\Big] + \sum_{j \neq \ell} \mathbf{E}_j\Big[\sum_{i \in S_\ell} \psi_i W\Big] \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2};$$
3. *with $p_{-\ell} := \sum_{r \neq \ell} \sum_{i \in S_r} \pi_i \varphi_i$, $\chi_\ell := p_{-\ell}/(\hat{\pi}_\ell q_\ell)$ and $\rho_\ell := 1/(1+\chi_\ell) = \hat{\pi}_\ell q_\ell / p$,*
$$\mathbf{E}_\ell\big[\log(1+\chi_\ell)\big] + \mathbf{E}_\ell\big[(1-\rho_\ell) W\big] \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2};$$
4. *with $\chi_\ell^* := \sum_{r \neq \ell} \pi_r^* \varphi_r^* / (\pi_\ell^* \varphi_\ell^*)$, $\mathbf{E}_\ell[\log(1+\chi_\ell^*)] \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$.*

The quantities χ_ℓ and χ_ℓ^* are the interference ratios of the model and of the truth: they factor $p = \hat{\pi}_\ell q_\ell (1+\chi_\ell)$ and $p_* = \pi_\ell^* \varphi_\ell^* (1+\chi_\ell^*)$ for data from cluster ℓ.

Proof. If $m = 1$ there is nothing to prove, so assume $m \geq 2$.

Item 1 (lower bound on the group mass). Since $p_* \geq \pi_r^* \varphi_r^*$ pointwise for every r, for $i \in S_\ell$ and data from cluster ℓ,

$$\mathbf{E}_\ell\Big[\frac{\varphi_i}{p_*}\Big] \leq \frac{1}{\pi_\ell^*} \mathbf{E}_\ell\Big[\frac{\varphi_i}{\varphi_\ell^*}\Big] = \frac{1}{\pi_\ell^*} \int \varphi_i\,d\boldsymbol{z} = \frac{1}{\pi_\ell^*},$$

because $\mathbf{E}_\ell$ integrates against φ_ℓ^*, which cancels. For $i \in S_r$ with $r \neq \ell$ we instead take the geometric mean of the two pointwise bounds $p_* \geq \pi_\ell^* \varphi_\ell^*$ and $p_* \geq \pi_r^* \varphi_r^*$:

$$\mathbf{E}_\ell\Big[\frac{\varphi_i}{p_*}\Big] \leq \frac{1}{\pi_{\min}} \mathbf{E}_\ell\Big[\Big(\frac{\varphi_i}{\varphi_\ell^*}\Big)^{1/2}\Big(\frac{\varphi_i}{\varphi_r^*}\Big)^{1/2}\Big] = \frac{1}{\pi_{\min}} \int \sqrt{\varphi_\ell^*}\,\varphi_i\,(\varphi_r^*)^{-1/2}\,d\boldsymbol{z} \leq \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$$

by the second bound in (37). Multiplying by π_i and summing, and noting that the components inside S_ℓ carry total mass $\hat{\pi}_\ell$,

$$\begin{gathered}\mathbf{E}_\ell\Big[\frac{p}{p_*}\Big] \leq \frac{\hat{\pi}_\ell}{\pi_\ell^*} + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}, \qquad \text{and in particular} \\ \mathbf{E}_r\Big[\frac{p}{p_*}\Big] \leq \frac{1}{\pi_{\min}} + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} \leq \frac{2}{\pi_{\min}} \quad \text{for every } r.\end{gathered} \tag{38}$$

In the other direction, since $-\log$ is convex, Jensen's inequality gives

$$L_0 \;\geq\; \mathcal{L} = \sum_r \pi_r^* \mathbf{E}_r\Big[-\log\frac{p}{p_*}\Big] \;\geq\; -\sum_r \pi_r^* \log \mathbf{E}_r\Big[\frac{p}{p_*}\Big] \;\geq\; -\pi_\ell^* \log \mathbf{E}_\ell\Big[\frac{p}{p_*}\Big] - \log\frac{2}{\pi_{\min}},$$

where the last step bounds the $r \neq \ell$ terms using the second display in (38) and $\sum_{r\neq\ell}\pi_r^* \leq 1$. Rearranging, $\mathbf{E}_\ell[p/p_*] \geq \exp\big(-\frac{L_0+\log(2/\pi_{\min})}{\pi_\ell^*}\big)$. Combining with (38) and $\pi_\ell^* \geq \pi_{\min}$,

$$\hat\pi_\ell \;\geq\; \pi_{\min}\exp\Big(-\frac{L_0+\log(2/\pi_{\min})}{\pi_{\min}}\Big) - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2},$$

and by the second and third terms of (15) the subtracted term is at most half the first, which is absorbed by enlarging C in the exponent.

Item 2 (posterior cross terms). Fix $j \neq \ell$ and $i \in S_\ell$. Since $|S_j| \leq n$, the pigeonhole principle supplies $k \in S_j$ with $\pi_k \geq \hat\pi_j/n$. Because $p \geq \pi_k\varphi_k$ and $\psi_i \leq 1$, and since $\min\{1,u\} \leq \sqrt{u}$ for $u \geq 0$,

$$\psi_i = \frac{\pi_i\varphi_i}{p} \leq \min\Big\{1, \frac{\pi_i\varphi_i}{\pi_k\varphi_k}\Big\} \leq \Big(\frac{\pi_i}{\pi_k}\Big)^{1/2}\Big(\frac{\varphi_i}{\varphi_k}\Big)^{1/2} \leq \Big(\frac{n}{\hat\pi_j}\Big)^{1/2}\Big(\frac{\varphi_i}{\varphi_k}\Big)^{1/2}.$$

By item 1, $(n/\hat\pi_j)^{1/2} \leq \sqrt{n}\exp\big(C\frac{L_0+\log(1/\pi_{\min})}{\pi_{\min}}\big)$, a factor that (15) absorbs into $\mathsf{A}_{\mathrm{g}}e^{-c\Delta^2}$ after halving c, since $\Delta^2 \gtrsim (L_0+\log(1/\pi_{\min}))/\pi_{\min}$. The first bound in (37) then gives $\mathbf{E}_j[\psi_i W] \leq \mathsf{A}_{\mathrm{g}}e^{-c\Delta^2}$. Summing over at most n indices i and m clusters, and absorbing nm into A_{g}, proves the second statement.

Item 3 (model interference ratio). For $\boldsymbol{z}$ from cluster ℓ pick $k \in S_\ell$ with $\pi_k \geq \hat\pi_\ell/n$; since $\hat\pi_\ell q_\ell \geq \pi_k\varphi_k$,

$$\chi_\ell = \frac{p_{-\ell}}{\hat\pi_\ell q_\ell} \leq \sum_{i\notin S_\ell}\frac{\pi_i\varphi_i}{\pi_k\varphi_k}.$$

We use the elementary inequalities $\log(1+u) \leq \sqrt{u}$ and $\frac{u}{1+u} \leq \sqrt{u}$ for $u \geq 0$ (both follow from $\sqrt{u} \geq \min\{u,1\}$, and $\frac{d}{du}\big(\sqrt{u}-\log(1+u)\big) = \frac{1+u-2\sqrt{u}}{2\sqrt{u}(1+u)} \geq 0$), together with the subadditivity $\sqrt{\sum_i u_i} \leq \sum_i\sqrt{u_i}$. Since $1-\rho_\ell = \chi_\ell/(1+\chi_\ell)$,

$$\log(1+\chi_\ell) + (1-\rho_\ell)W \;\leq\; (1+W)\sqrt{\chi_\ell} \;\leq\; (1+W)\sum_{i\notin S_\ell}\Big(\frac{\pi_i}{\pi_k}\Big)^{1/2}\Big(\frac{\varphi_i}{\varphi_k}\Big)^{1/2}.$$

Taking $\mathbf{E}_\ell$ and applying the first bound in (37) (with data cluster ℓ, $i \in S_r$ for $r \neq \ell$, and $k \in S_\ell$), with the weight ratio absorbed exactly as in item 2, proves item 3.

Item 4 (true interference ratio). By the same two inequalities,

$$\log(1+\chi_\ell^*) \;\leq\; \sqrt{\chi_\ell^*} \;\leq\; \sum_{r\neq\ell}\Big(\frac{\pi_r^*}{\pi_\ell^*}\Big)^{1/2}\Big(\frac{\varphi_r^*}{\varphi_\ell^*}\Big)^{1/2},$$

and the third bound in (37) gives $\mathbf{E}_\ell\log(1+\chi_\ell^*) \leq m\,\pi_{\min}^{-1/2}e^{-\Delta^2/8} \leq \mathsf{A}_{\mathrm{g}}e^{-c\Delta^2}$. □

C.2 Group-weight control

Lemma C.2 (Group-weight control)**.** *Under the assumptions of Proposition B.4,*

$$\sum_{\ell=1}^m|\hat\pi_\ell - \pi_\ell^*| \leq 2\varepsilon_\pi + \mathsf{A}_{\mathrm{g}}e^{-c\Delta^2}, \qquad \textit{and hence} \qquad \hat\pi_\ell \geq \frac{\pi_\ell^*}{2} \quad \forall\ell.$$

Proof. By Lemma A.4, $\sum_i|\pi_i - \mathbf{E}_{p_*}\psi_i| \leq 2\varepsilon_\pi$. Writing $\Psi_\ell := \sum_{i\in S_\ell}\psi_i$ and using that the S_ℓ partition $[n]$, the triangle inequality gives $\sum_\ell|\hat\pi_\ell - \mathbf{E}_{p_*}\Psi_\ell| \leq \sum_i|\pi_i - \mathbf{E}_{p_*}\psi_i| \leq 2\varepsilon_\pi$. Next, splitting $\mathbf{E}_{p_*} = \sum_j \pi_j^*\mathbf{E}_j$ and using $\sum_{i=1}^n\psi_i = 1$,

$$\mathbf{E}_{p_*}\Psi_\ell = \pi_\ell^*\Big(1 - \mathbf{E}_\ell\Big[\sum_{i\notin S_\ell}\psi_i\Big]\Big) + \sum_{j\neq\ell}\pi_j^*\mathbf{E}_j\Psi_\ell,$$

so that $\sum_\ell |\mathbf{E}_{p_*}\Psi_\ell - \pi^*_\ell| \le \sum_\ell \mathbf{E}_\ell\big[\sum_{i\notin S_\ell}\psi_i\big] + \sum_\ell \sum_{j\neq\ell}\mathbf{E}_j\big[\sum_{i\in S_\ell}\psi_i\big] \le \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$ by Lemma C.1(2) with $K=0$. A final triangle inequality proves the first claim.

For the second, (28), (16) and $\varepsilon \le c\,\pi_{\min}$ give $2\varepsilon_\pi + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2} \le c\,\varepsilon \le \pi_{\min}/2$. Since every summand is nonnegative, $|\hat\pi_\ell - \pi^*_\ell| \le \pi_{\min}/2 \le \pi^*_\ell/2$ for each ℓ, whence $\hat\pi_\ell \ge \pi^*_\ell/2$. □

C.3 One-cluster projection comparison

We first isolate an abstract one-cluster statement; the mixture-of-experts application follows by the recentering (34).

Let ν be the standard Gaussian on $(\boldsymbol{x}', y') \in \mathbb{R}^{d+1}$. Fix weights $\lambda \in \mathcal{P}^{N-1}$, vectors $\boldsymbol{h}_i, \beta_i \in \mathbb{R}^d$ and scalars $c_i \in \mathbb{R}$, define the affine regressions $a_i(\boldsymbol{x}') := \langle \boldsymbol{x}', \beta_i\rangle + c_i$, and set

$$q(\boldsymbol{x}', y') := \sum_i \lambda_i\, \varphi(\boldsymbol{x}'; \boldsymbol{h}_i)\, \varphi_1(y'; a_i(\boldsymbol{x}')), \qquad \psi_i := \frac{\lambda_i \varphi(\boldsymbol{x}'; \boldsymbol{h}_i)\varphi_1(y'; a_i)}{q}, \qquad r_i := y' - a_i. \tag{39}$$

Write $\bar\lambda_i := \mathbf{E}_\nu \psi_i$ and introduce the posterior means, variances and cross-covariance

$$\bar{\boldsymbol{h}} := \sum_i \psi_i \boldsymbol{h}_i, \qquad \bar{\boldsymbol{w}} := \sum_i \psi_i r_i \beta_i, \qquad \bar a := \sum_i \psi_i a_i,$$
$$V_h := \sum_i \psi_i \left\|\boldsymbol{h}_i - \bar{\boldsymbol{h}}\right\|^2, \qquad V_w := \sum_i \psi_i \left\|r_i\beta_i - \bar{\boldsymbol{w}}\right\|^2, \qquad V_y := \sum_i \psi_i (a_i - \bar a)^2,$$
$$\mathcal{C} := \sum_i \psi_i \left\langle r_i\beta_i - \bar{\boldsymbol{w}},\, \boldsymbol{h}_i - \bar{\boldsymbol{h}}\right\rangle,$$

and the one-cluster projection

$$P := \mathbf{E}_\nu\Big[\sum_i \psi_i\big(\|\boldsymbol{h}_i\|^2 - \langle \boldsymbol{x}', \boldsymbol{h}_i\rangle + a_i^2 - y' a_i\big)\Big]. \tag{40}$$

Remark 2 (The recentering is measure preserving)**.** *For cluster ℓ of the mixture of experts,* (39) *is q_ℓ written in the coordinates* (34)*, with $\lambda = \lambda^{(\ell)}$, $\boldsymbol{h}_i = \boldsymbol{\mu}_i - \boldsymbol{\mu}^*_\ell$, $\beta_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}^*_\ell$ and $c_i = \langle \boldsymbol{\mu}^*_\ell, \beta_i\rangle$; indeed $y - \boldsymbol{x}^\top\boldsymbol{\theta}_i = y' - (\langle \boldsymbol{x}', \beta_i\rangle + \langle \boldsymbol{\mu}^*_\ell, \beta_i\rangle) = r_i$. The map $(\boldsymbol{x}, y) \mapsto (\boldsymbol{x}', y')$ is the shift-and-shear $(\boldsymbol{x}, y) \mapsto (\boldsymbol{x} - \boldsymbol{\mu}^*_\ell,\ y - \boldsymbol{x}^\top\boldsymbol{\theta}^*_\ell)$, whose Jacobian matrix is lower triangular with unit diagonal and hence has determinant* 1*. Consequently $D_{\mathrm{KL}}(\varphi^*_\ell \| q_\ell) = D_{\mathrm{KL}}(\nu\|q)$ and $\mathbf{E}_\ell \psi_i^{(\ell)} = \mathbf{E}_\nu \psi_i$, and by* (35) *the projection* (40) *coincides with P_ℓ of* (33)*.*

Lemma C.3 (One-cluster projection comparison)**.** *In the setting* (39)*, for every $\tau \in (0, 1]$,*

$$P \;\ge\; 2D_{\mathrm{KL}}(\nu\|q) - 2D_{\mathrm{KL}}(\bar\lambda\|\lambda) - \frac{\tau}{2}\,\mathbf{E}_\nu V_h - \Big(1 + \frac{1}{2\tau}\Big)\mathbf{E}_\nu V_w.$$

More precisely, one has the exact identity and the entropy bound

$$P = 2D_{\mathrm{KL}}(\nu\|q) - 2\,\mathbf{E}_\nu\Big[\sum_i \psi_i \log\frac{\psi_i}{\lambda_i}\Big] + \mathbf{E}_\nu\big[V_h + V_y + \mathcal{C}\big],$$
$$\mathbf{E}_\nu\Big[\sum_i \psi_i \log\frac{\psi_i}{\bar\lambda_i}\Big] \le \tfrac{1}{2}\mathbf{E}_\nu\big[V_h + 2\mathcal{C} + V_w + V_y\big]. \tag{41}$$

Proof. Indices with $\lambda_i = 0$ are dropped. Write

$$f_i := \log\frac{\varphi(\boldsymbol{x}'; \boldsymbol{h}_i)\varphi_1(y'; a_i)}{\varphi(\boldsymbol{x}'; 0)\varphi_1(y'; 0)} = \langle \boldsymbol{h}_i, \boldsymbol{x}'\rangle - \tfrac{1}{2}\|\boldsymbol{h}_i\|^2 + y' a_i - \tfrac{1}{2}a_i^2,$$

obtained by cancelling the normalizing constants and expanding the two squares, so that $q/\nu = \sum_i \lambda_i e^{f_i}$ and $\psi_i = \lambda_i e^{f_i}/\sum_k \lambda_k e^{f_k}$. As in (23)–(24),

$$\nabla_{\boldsymbol{x}'} f_i = \boldsymbol{h}_i + r_i\beta_i, \quad \partial_{y'} f_i = a_i, \quad \nabla_{\boldsymbol{x}'}\psi_i = \psi_i\big(\boldsymbol{h}_i - \bar{\boldsymbol{h}} + r_i\beta_i - \bar{\boldsymbol{w}}\big), \quad \partial_{y'}\psi_i = \psi_i(a_i - \bar a). \tag{42}$$

(a) Posterior identity. For every i with $\psi_i > 0$ we have $\log \frac{q}{\nu} = \log \frac{\lambda_i}{\psi_i} + f_i$. Averaging this identity with the weights ψ_i (which sum to one) and taking $\mathbf{E}_\nu$,

$$D_{\mathrm{KL}}(\nu\|q) = -\mathbf{E}_\nu \log \frac{q}{\nu} = \mathbf{E}_\nu\Big[\sum_i \psi_i \log \frac{\psi_i}{\lambda_i}\Big] - \mathbf{E}_\nu\Big[\sum_i \psi_i f_i\Big]. \tag{43}$$

(b) Stein computations and an exact identity. Two centring identities are used repeatedly: since $\sum_i \psi_i(\boldsymbol{h}_i - \bar{\boldsymbol{h}}) = 0$ and $\sum_i \psi_i(r_i\beta_i - \bar{\boldsymbol{w}}) = 0$,

$$\sum_i \psi_i \left\langle \boldsymbol{h}_i - \bar{\boldsymbol{h}}, \boldsymbol{h}_i \right\rangle = V_h, \qquad \sum_i \psi_i \left\langle r_i\beta_i - \bar{\boldsymbol{w}}, \boldsymbol{h}_i \right\rangle = \mathcal{C}, \qquad \sum_i \psi_i (a_i - \bar{a}) a_i = V_y,$$

and likewise $\sum_i \psi_i \|\boldsymbol{h}_i\|^2 = V_h + \|\bar{\boldsymbol{h}}\|^2$ and $\sum_i \psi_i a_i^2 = V_y + \bar{a}^2$. By Lemma A.3 applied in $\boldsymbol{x}'$ to the field $F = \bar{\boldsymbol{h}} = \sum_i \psi_i \boldsymbol{h}_i$, and by (42),

$$\mathbf{E}_\nu \left\langle \boldsymbol{x}', \bar{\boldsymbol{h}} \right\rangle = \mathbf{E}_\nu \sum_i \left\langle \nabla_{\boldsymbol{x}'} \psi_i, \boldsymbol{h}_i \right\rangle = \mathbf{E}_\nu \sum_i \psi_i \left\langle \boldsymbol{h}_i - \bar{\boldsymbol{h}} + r_i\beta_i - \bar{\boldsymbol{w}}, \boldsymbol{h}_i \right\rangle = \mathbf{E}_\nu\big[V_h + \mathcal{C}\big],$$

and by Lemma A.3 in y' with $\boldsymbol{x}'$ fixed, $\mathbf{E}_\nu\big[y' \sum_i \psi_i a_i\big] = \mathbf{E}_\nu \sum_i \partial_{y'} \psi_i \, a_i = \mathbf{E}_\nu V_y$. Substituting these into f_i and into (40),

$$\begin{aligned}
\mathbf{E}_\nu \sum_i \psi_i f_i &= \mathbf{E}_\nu\big[V_h + \mathcal{C}\big] - \tfrac{1}{2}\mathbf{E}_\nu\big[V_h + \|\bar{\boldsymbol{h}}\|^2\big] + \mathbf{E}_\nu V_y - \tfrac{1}{2}\mathbf{E}_\nu\big[V_y + \bar{a}^2\big] \\
&= \mathbf{E}_\nu\Big[\tfrac{1}{2}V_h + \mathcal{C} - \tfrac{1}{2}\|\bar{\boldsymbol{h}}\|^2 + \tfrac{1}{2}V_y - \tfrac{1}{2}\bar{a}^2\Big], \qquad (44) \\
P &= \mathbf{E}_\nu\big[V_h + \|\bar{\boldsymbol{h}}\|^2\big] - \mathbf{E}_\nu\big[V_h + \mathcal{C}\big] + \mathbf{E}_\nu\big[V_y + \bar{a}^2\big] - \mathbf{E}_\nu V_y \\
&= \mathbf{E}_\nu\big[\,\|\bar{\boldsymbol{h}}\|^2 + \bar{a}^2 - \mathcal{C}\big]. \qquad (45)
\end{aligned}$$

Inserting (44) into (43), multiplying by 2 and using (45) to substitute P for $\mathbf{E}_\nu[\|\bar{\boldsymbol{h}}\|^2 + \bar{a}^2 - \mathcal{C}]$ gives the exact identity in (41).

(c) Log-Sobolev. Gross's Gaussian logarithmic Sobolev inequality, in the form

$$\mathbf{E}_\nu\big[f^2 \log f^2\big] - \mathbf{E}_\nu[f^2]\,\log \mathbf{E}_\nu[f^2] \;\le\; 2\,\mathbf{E}_\nu \|\nabla f\|^2,$$

becomes, after the substitution $F = f^2$ (so that $\|\nabla f\|^2 = \|\nabla F\|^2/(4F)$),

$$\mathbf{E}_\nu[F \log F] - \mathbf{E}_\nu F \log \mathbf{E}_\nu F \;\le\; \tfrac{1}{2}\,\mathbf{E}_\nu\big[\,\|\nabla F\|^2/F\big] \qquad \text{for positive smooth } F.$$

Apply this to $F = \psi_i$, which is smooth and strictly positive, and for which $\mathbf{E}_\nu \psi_i = \bar{\lambda}_i$, so that the left-hand side equals $\mathbf{E}_\nu[\psi_i \log(\psi_i/\bar{\lambda}_i)]$. Summing over i and using (42),

$$\sum_i \frac{\|\nabla \psi_i\|^2}{\psi_i} = \sum_i \psi_i \Big(\big\|(\boldsymbol{h}_i - \bar{\boldsymbol{h}}) + (r_i\beta_i - \bar{\boldsymbol{w}})\big\|^2 + (a_i - \bar{a})^2 \Big) = V_h + 2\mathcal{C} + V_w + V_y,$$

which is the second part of (41). Since $\sum_i \psi_i \log \frac{\psi_i}{\lambda_i} = \sum_i \psi_i \log \frac{\psi_i}{\bar{\lambda}_i} + \sum_i \psi_i \log \frac{\bar{\lambda}_i}{\lambda_i}$ and $\mathbf{E}_\nu \sum_i \psi_i \log \frac{\bar{\lambda}_i}{\lambda_i} = D_{\mathrm{KL}}(\bar{\lambda}\|\lambda)$, combining the two parts of (41) makes V_h and V_y cancel:

$$P \;\ge\; 2D_{\mathrm{KL}}(\nu\|q) - 2D_{\mathrm{KL}}(\bar{\lambda}\|\lambda) - \mathbf{E}_\nu\big[\mathcal{C} + V_w\big].$$

(d) The deficit. By Cauchy–Schwarz in the posterior and Young's inequality $ab \le \frac{\tau}{2}a^2 + \frac{1}{2\tau}b^2$,

$$|\mathcal{C}| \le \sum_i \psi_i \, \|r_i\beta_i - \bar{\boldsymbol{w}}\| \, \|\boldsymbol{h}_i - \bar{\boldsymbol{h}}\| \le \frac{\tau}{2} V_h + \frac{1}{2\tau} V_w,$$

so $\mathbf{E}_\nu[\mathcal{C} + V_w] \le \frac{\tau}{2}\mathbf{E}_\nu V_h + (1 + \frac{1}{2\tau})\mathbf{E}_\nu V_w$. □

Lemma C.4 (Projection comparison and weight KL; formal version of Lemma 4.5)**.** *Under the assumptions of Proposition B.4, for every $\ell \in [m]$ and every $\tau \in (0, 1]$,*

$$P_\ell \;\ge\; 2D_{\mathrm{KL}}(\varphi_\ell^* \| q_\ell) - 2D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)} \| \lambda^{(\ell)}) - \tfrac{\tau}{2}B^2 - 6\Big(1 + \tfrac{1}{2\tau}\Big)B_\theta^2,$$

$$\sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)} \| \lambda^{(\ell)}) \le \varepsilon_\pi + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}.$$

Proof. First inequality. Apply Lemma C.3 to cluster ℓ in the recentered coordinates; by Remark 2 its P is P_ℓ and its $D_{\mathrm{KL}}(\nu\|q)$ is $D_{\mathrm{KL}}(\varphi_\ell^*\|q_\ell)$. It remains to bound the two variances. A variance is at most the corresponding second moment, so by (17), $V_h \le \sum_i \psi_i \|\boldsymbol{h}_i\|^2 \le B^2$. Similarly $V_w \le \sum_i \psi_i r_i^2 \|\beta_i\|^2 \le B_\theta^2 \sum_i \psi_i r_i^2$, and, using $(u-v)^2 \le 2u^2+2v^2$ twice together with $\|\beta_i\| \le B_\theta$ and $|c_i| = |\langle \boldsymbol{\mu}_\ell^*, \beta_i\rangle| \le D_{\max} B_\theta$,

$$r_i^2 \le 2y'^2 + 2a_i^2 \le 2y'^2 + 4B_\theta^2\big(\|\boldsymbol{x}'\|^2 + D_{\max}^2\big).$$

Taking $\mathbf{E}_\nu$, so that $\mathbf{E}_\nu y'^2 = 1$ and $\mathbf{E}_\nu \|\boldsymbol{x}'\|^2 = d$, and using $B_\theta^2\Lambda \le 1$ from Lemma A.1(ii),

$$\mathbf{E}_\nu V_w \;\le\; B_\theta^2\big(2 + 4B_\theta^2\Lambda\big) \;\le\; 6B_\theta^2.$$

Second inequality. Let $\bar{q}_\ell := \sum_{i\in S_\ell} \bar{\lambda}_i^{(\ell)} \varphi_i$ and $\bar{p} := \sum_\ell \hat{\pi}_\ell \bar{q}_\ell$, i.e. the model with weights $\bar{\pi}_i := \hat{\pi}_\ell \bar{\lambda}_i^{(\ell)}$ for $i \in S_\ell$; since $\sum_{i\in S_\ell} \bar{\lambda}_i^{(\ell)} = 1$, we have $\bar{\boldsymbol{\pi}} \in \mathcal{P}$. By Lemma A.4, $\mathbf{E}_{p_*} \log \frac{\bar{p}}{p} = \mathcal{L}(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\theta}) - \mathcal{L}(\bar{\boldsymbol{\pi}},\boldsymbol{\mu},\boldsymbol{\theta}) \le \varepsilon_\pi$. Conversely, for $\boldsymbol{z}$ from cluster ℓ we have $\bar{p} \ge \hat{\pi}_\ell \bar{q}_\ell$ and $p = \hat{\pi}_\ell q_\ell(1+\chi_\ell)$, so, using $\varphi_i/q_\ell = \psi_i^{(\ell)}/\lambda_i^{(\ell)}$ and then Jensen's inequality for the concave logarithm against the probability vector $\psi^{(\ell)}$,

$$\log\frac{\bar{p}}{p} \;\ge\; \log\frac{\bar{q}_\ell}{q_\ell} - \log(1+\chi_\ell) = \log\sum_{i\in S_\ell} \psi_i^{(\ell)} \frac{\bar{\lambda}_i^{(\ell)}}{\lambda_i^{(\ell)}} - \log(1+\chi_\ell) \;\ge\; \sum_{i\in S_\ell} \psi_i^{(\ell)} \log\frac{\bar{\lambda}_i^{(\ell)}}{\lambda_i^{(\ell)}} - \log(1+\chi_\ell).$$

Taking $\mathbf{E}_\ell$ turns $\psi_i^{(\ell)}$ into $\bar{\lambda}_i^{(\ell)}$ and hence the first sum into $D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)}\|\lambda^{(\ell)})$; multiplying by π_ℓ^*, summing over ℓ and using Lemma C.1(3) for the interference term gives $\varepsilon_\pi \ge \sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\bar{\lambda}^{(\ell)}\|\lambda^{(\ell)}) - \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$. □

C.4 Loss versus within-group KL divergence

Lemma C.5 (Loss vs. within-group KL; formal version of Lemma 4.4)**.** *Under the assumptions of Proposition B.4,*

$$\mathcal{L}(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\theta}) \;\le\; \sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\varphi_\ell^*\|q_\ell) + C\varepsilon_\pi + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}.$$

Proof. For $\boldsymbol{z}$ from cluster ℓ, factoring the local component out of each mixture gives $p_* = \pi_\ell^*\varphi_\ell^*(1+\chi_\ell^*)$, while dropping all but the ℓth group of the model gives $p \ge \hat{\pi}_\ell q_\ell$. Hence

$$\log\frac{p_*}{p} \;\le\; \log\frac{\varphi_\ell^*}{q_\ell} + \log\frac{\pi_\ell^*}{\hat{\pi}_\ell} + \log(1+\chi_\ell^*).$$

Taking $\mathbf{E}_\ell$, multiplying by π_ℓ^* and summing over ℓ, and noting that the middle term is deterministic,

$$\mathcal{L} \;\le\; \sum_\ell \pi_\ell^* D_{\mathrm{KL}}(\varphi_\ell^*\|q_\ell) + \sum_\ell \pi_\ell^* \log\frac{\pi_\ell^*}{\hat{\pi}_\ell} + \sum_\ell \pi_\ell^* \mathbf{E}_\ell \log(1+\chi_\ell^*).$$

By Lemma C.2, $\hat{\pi}_\ell \ge \pi_\ell^*/2$, so $\log u \le u-1$ with $u = \pi_\ell^*/\hat{\pi}_\ell$ gives $\log\frac{\pi_\ell^*}{\hat{\pi}_\ell} \le \frac{\pi_\ell^*-\hat{\pi}_\ell}{\hat{\pi}_\ell} \le \frac{2|\pi_\ell^*-\hat{\pi}_\ell|}{\pi_\ell^*}$; the factors π_ℓ^* cancel and the middle sum is at most $2\sum_\ell|\hat{\pi}_\ell - \pi_\ell^*| \le 4\varepsilon_\pi + \mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$. The last sum is at most $\mathsf{A}_{\mathrm{g}} e^{-c\Delta^2}$ by Lemma C.1(4), since $\sum_\ell \pi_\ell^* = 1$. □

D Regularity and Gaussian Ratio Estimates

D.1 Local smoothness and self-bounding of the gradient

Lemma D.1 (Local smoothness)**.** *Fix $\boldsymbol{\pi}$ and $(\boldsymbol{\mu},\boldsymbol{\theta})$ and suppose that, for every $\ell \in [m]$, $\hat{\pi}_\ell \ge c_{\mathrm{cov}}\pi_\ell^*$ and*

$$\frac{1}{\hat{\pi}_\ell}\sum_{i\in S_\ell} \pi_i\big(\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\|^2 + \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*\|^2\big) \le B^2, \qquad \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_\ell^*\| \le 1 \;\; (i \in S_\ell).$$

Let $\boldsymbol{\delta} = (\boldsymbol{\delta}_i^{\mu}, \boldsymbol{\delta}_i^{\theta})_{i \le n}$ *satisfy* $\frac{1}{\hat{\pi}_\ell} \sum_{i \in S_\ell} \pi_i \|\boldsymbol{\delta}_i\|^2 \le D^2$ *for every* ℓ*, and set*

$$A(D) := C(1+\Lambda)^2 \Big(1 + d + B^2 + D^2 + \log \tfrac{1}{c_{\mathrm{cov}} \pi_{\min}}\Big),$$
$$A := C(1+\Lambda)^2 \Big(1 + d + B^2 + \log \tfrac{1}{c_{\mathrm{cov}} \pi_{\min}} + \tfrac{1}{c_{\mathrm{cov}} \pi_{\min}}\Big)$$

with C *sufficiently large; for* $c_{\mathrm{cov}} = \frac{1}{2}$ *this* A *is the constant* (13)*. Then, writing* $g := \nabla_{\boldsymbol{\mu},\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta})$*,*

$$\mathcal{L}(\boldsymbol{\pi}, (\boldsymbol{\mu}, \boldsymbol{\theta}) + \boldsymbol{\delta}) \le \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) + \langle g, \boldsymbol{\delta} \rangle + \tfrac{1}{2} A(D) \|\boldsymbol{\delta}\|_F^2, \tag{46}$$
$$\|g\|_F^2 \le A, \tag{47}$$
$$\mathcal{L}(\boldsymbol{\pi}, (\boldsymbol{\mu}, \boldsymbol{\theta}) - \eta g) \le \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}) - \tfrac{\eta}{2} \|g\|_F^2 \qquad (\eta \le c/A), \tag{48}$$
$$\|g\|_F^2 \le 2A\, \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\theta}). \tag{49}$$

Proof. Hessian bound. For a single sample $\boldsymbol{z}$ let $f_{\boldsymbol{z}}(\boldsymbol{\mu}, \boldsymbol{\theta}) := -\log \sum_i \pi_i \varphi_i(\boldsymbol{z})$, so that $\mathcal{L} = \mathbf{E}_{p_*} f_{\boldsymbol{z}} + \text{const}$. Fix a direction $\boldsymbol{v}$, put $F(t) := f_{\boldsymbol{z}}\big((\boldsymbol{\mu}, \boldsymbol{\theta}) + t\boldsymbol{v}\big)$ and $\ell_i(t) := \log \pi_i + \log \varphi_i(\boldsymbol{z})$ evaluated at $(\boldsymbol{\mu}, \boldsymbol{\theta}) + t\boldsymbol{v}$. Differentiating $\log \varphi_i = \text{const} - \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2 - \frac{1}{2} r_i^2$ along $\boldsymbol{v}$,

$$\ell_i' = \langle \boldsymbol{x} - \boldsymbol{\mu}_i(t), \boldsymbol{v}_i^{\mu} \rangle + r_i(t) \langle \boldsymbol{x}, \boldsymbol{v}_i^{\theta} \rangle, \qquad \ell_i'' = -\|\boldsymbol{v}_i^{\mu}\|^2 - \langle \boldsymbol{x}, \boldsymbol{v}_i^{\theta} \rangle^2.$$

Since $F = -\log \sum_i e^{\ell_i}$, we have $F' = -\sum_i \psi_i \ell_i'$ and, using $\frac{d}{dt}\psi_i = \psi_i(\ell_i' - \sum_k \psi_k \ell_k')$,

$$F'' = -\sum_i \psi_i \ell_i'' - \Big[\sum_i \psi_i (\ell_i')^2 - \Big(\sum_i \psi_i \ell_i'\Big)^2\Big].$$

The bracket is the ψ-variance of ℓ', hence nonnegative, and $\ell_i'' \le 0$, so $-\sum_i \psi_i (\ell_i')^2 \le F'' \le \sum_i \psi_i \big(\|\boldsymbol{v}_i^{\mu}\|^2 + \|\boldsymbol{x}\|^2 \|\boldsymbol{v}_i^{\theta}\|^2\big)$. With $(\ell_i')^2 \le 2\big(\|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2 \|\boldsymbol{x}\|^2\big) \|\boldsymbol{v}_i\|^2$ by $(a+b)^2 \le 2a^2 + 2b^2$ and Cauchy–Schwarz, taking the supremum over $\|\boldsymbol{v}\|_F \le 1$ gives

$$\begin{aligned} \big\|\nabla^2 f_{\boldsymbol{z}}\big\|_{\mathrm{op}} &\le 1 + \|\boldsymbol{x}\|^2 + 2\sum_i \psi_i \big(\|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2 \|\boldsymbol{x}\|^2\big) \\ &\le (1 + \|\boldsymbol{x}\|^2)\Big(1 + 2\sum_i \psi_i \big(\|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2\big)\Big), \end{aligned} \tag{50}$$

all posteriors being evaluated at $(\boldsymbol{\mu}, \boldsymbol{\theta}) + t\boldsymbol{v}$; the last step uses $a + b\|\boldsymbol{x}\|^2 \le (a+b)(1 + \|\boldsymbol{x}\|^2)$ for $a, b \ge 0$.

To control the remaining sum we use the Gibbs variational principle $-\log \sum_i \pi_i e^{-\alpha_i} = \min_{\varrho \in \mathcal{P}} \big\{\sum_i \varrho_i \alpha_i + \sum_i \varrho_i \log(\varrho_i/\pi_i)\big\}$ with $\alpha_i := \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + \frac{1}{2} r_i(t)^2$, whose minimizer is $\varrho = \psi$. Evaluating the objective at the feasible point $\varrho_i = \pi_i/\hat{\pi}_\ell$ on S_ℓ and $\varrho_i = 0$ elsewhere, whose entropy term is exactly $\log(1/\hat{\pi}_\ell)$, and dropping the nonnegative term $\sum_i \psi_i \log(\psi_i/\pi_i)$ on the left,

$$\sum_i \psi_i \big(\|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2\big) \;\le\; \frac{1}{\hat{\pi}_\ell} \sum_{i \in S_\ell} \pi_i \big(\|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2\big) + 2\log \frac{1}{\hat{\pi}_\ell}. \tag{51}$$

Now take $\boldsymbol{z}$ from cluster ℓ, $\boldsymbol{v} = \boldsymbol{\delta}$ and $t \in [0,1]$, and write $\boldsymbol{h}_i(t) := \boldsymbol{\mu}_i(t) - \boldsymbol{\mu}_\ell^*$, $\beta_i(t) := \boldsymbol{\theta}_i(t) - \boldsymbol{\theta}_\ell^*$, so that $\|\boldsymbol{h}_i(t)\|^2 \le 2\|\boldsymbol{h}_i\|^2 + 2\|\boldsymbol{\delta}_i^{\mu}\|^2$ and $\|\beta_i(t)\|^2 \le 2\|\beta_i\|^2 + 2\|\boldsymbol{\delta}_i^{\theta}\|^2$. Apply Cauchy–Schwarz to (50) in the form $\mathbf{E}_\ell[XY] \le \|X\|_{L^2} \|Y\|_{L^2}$ with $X = 1 + \|\boldsymbol{x}\|^2$ and Y the second factor. By Lemma A.2, $\|X\|_{L^2} \le C(1+\Lambda)$. For Y, insert (51) and apply Minkowski's inequality, which moves the L^2 norm inside the convex combination:

$$\|Y\|_{L^2} \;\le\; 1 + 4\log \frac{1}{\hat{\pi}_\ell} + \frac{2}{\hat{\pi}_\ell} \sum_{i \in S_\ell} \pi_i \big\| \|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2 \big\|_{L^2}.$$

Since $\mathbf{E}_\ell \|\boldsymbol{x} - \boldsymbol{\mu}_i(t)\|^4 \le C\big(d^2 + \|\boldsymbol{h}_i(t)\|^4\big)$ and, because $r_i(t) = \epsilon - \boldsymbol{x}^\top \beta_i(t)$ with $\boldsymbol{x}^\top \beta_i(t) \sim N(\langle \boldsymbol{\mu}_\ell^*, \beta_i(t) \rangle, \|\beta_i(t)\|^2)$, $\mathbf{E}_\ell r_i(t)^4 \le C\big(1 + \Lambda^2 \|\beta_i(t)\|^4\big)$ by Lemma A.2, subadditivity of the

square root gives $\big\| \|\boldsymbol{x}-\boldsymbol{\mu}_i(t)\|^2 + r_i(t)^2 \big\|_{L^2} \le C\big(1+d+\|\boldsymbol{h}_i(t)\|^2 + \Lambda\,\|\beta_i(t)\|^2\big)$. Averaging over $i \in S_\ell$ with weights $\pi_i/\hat\pi_\ell$, using $\Lambda \ge 1$ and the two hypotheses on B^2 and D^2,

$$\frac{1}{\hat\pi_\ell}\sum_{i\in S_\ell}\pi_i\big(1+d+\|\boldsymbol{h}_i(t)\|^2+\Lambda\,\|\beta_i(t)\|^2\big) \;\le\; 1+d+2\Lambda\big(B^2+D^2\big) \;\le\; 2\Lambda\big(1+d+B^2+D^2\big).$$

Finally $\log(1/\hat\pi_\ell) \le \log\big(1/(c_{\mathrm{cov}}\pi_{\min})\big)$, so $\|Y\|_{L^2} \le C(1+\Lambda)\big(1+d+B^2+D^2+\log\frac{1}{c_{\mathrm{cov}}\pi_{\min}}\big)$ and therefore $\mathbf{E}_\ell\,\big\|\nabla^2 f_{\boldsymbol{z}}\big\|_{\mathrm{op}} \le A(D)$. Averaging over ℓ with weights π^*_ℓ gives $\big\|\nabla^2\mathcal{L}\big\|_{\mathrm{op}} \le A(D)$ along the whole segment (the groups S_ℓ are those of the base point and are used only to choose the comparison ϱ; no reassignment is needed). Taylor's theorem with integral remainder, $\mathcal{L}(1) = \mathcal{L}(0) + \langle g, \boldsymbol{\delta}\rangle + \int_0^1 (1-t)\,\frac{d^2}{dt^2}\mathcal{L}\,dt$, and $\int_0^1 (1-t)dt = \frac12$ give (46).

Gradient bound. By (20), Jensen's inequality $\|\mathbf{E}X\|^2 \le \mathbf{E}\,\|X\|^2$ and $\psi_i^2 \le \psi_i$,

$$\|g\|_F^2 = \sum_i \Big(\big\|\mathbf{E}_{p_*}[\psi_i(\boldsymbol{\mu}_i-\boldsymbol{x})]\big\|^2+\big\|\mathbf{E}_{p_*}[\psi_i r_i \boldsymbol{x}]\big\|^2\Big) \le \mathbf{E}_{p_*}\Big[\sum_i \psi_i\big(\,\|\boldsymbol{x}-\boldsymbol{\mu}_i\|^2+r_i^2\,\|\boldsymbol{x}\|^2\big)\Big] \le A,$$

the last step being the estimate just performed with $t=0$ and $\boldsymbol{\delta}=0$, which yields $A(0) \le A$.

Descent and self-bound. For $\boldsymbol{\delta} = -\eta g$ with $\eta \le c/A$,

$$\frac{1}{\hat\pi_\ell}\sum_{i\in S_\ell}\pi_i\,\|\boldsymbol{\delta}_i\|^2 \le \frac{\eta^2\,\|g\|_F^2}{c_{\mathrm{cov}}\pi_{\min}} \le \frac{c^2}{A\,c_{\mathrm{cov}}\pi_{\min}} \le 1,$$

because A contains the term $1/(c_{\mathrm{cov}}\pi_{\min})$ with a large constant. Hence (46) applies with $D=1$, and $A(1) \le A$ by the same observation, giving $\mathcal{L}((\boldsymbol{\mu},\boldsymbol{\theta})-\eta g) \le \mathcal{L} - \eta\,\|g\|_F^2 + \frac{A\eta^2}{2}\,\|g\|_F^2 \le \mathcal{L} - \frac{\eta}{2}\,\|g\|_F^2$ for $\eta \le 1/A$, which is (48). Taking $\eta = 1/A$ and using $\mathcal{L} \ge 0$ on the left-hand side gives $0 \le \mathcal{L} - \frac{1}{2A}\,\|g\|_F^2$, i.e. (49). □

D.2 Gaussian products and the regression factor

Lemma D.2 (Gaussian products)**.** *Let $p \ge 1$, let $s_1,\dots,s_K \in \mathbb{R}$ with $\sum_k s_k = 1$, and let $\boldsymbol{m}_1,\dots,\boldsymbol{m}_K \in \mathbb{R}^p$. Then, for all $\boldsymbol{x} \in \mathbb{R}^p$,*

$$\prod_k \varphi(\boldsymbol{x};\boldsymbol{m}_k)^{s_k} = \exp\Big(-\tfrac12\sum_{k<l} s_k s_l\,\|\boldsymbol{m}_k-\boldsymbol{m}_l\|^2\Big)\,\varphi\Big(\boldsymbol{x};\bar{\boldsymbol{m}}\Big), \qquad \bar{\boldsymbol{m}} := \sum_k s_k\boldsymbol{m}_k.$$

Consequently, for any weight W,

$$\int \prod_k \varphi(\boldsymbol{x};\boldsymbol{m}_k)^{s_k} W(\boldsymbol{x})\,d\boldsymbol{x} = \exp\Big(-\tfrac12\sum_{k<l} s_k s_l\,\|\boldsymbol{m}_k-\boldsymbol{m}_l\|^2\Big)\,\mathbf{E}_{\boldsymbol{x}\sim N(\bar{\boldsymbol{m}},I_p)}W(\boldsymbol{x}).$$

Proof. Using $\sum_k s_k = 1$ and completing the square,

$$\sum_k s_k\,\|\boldsymbol{x}-\boldsymbol{m}_k\|^2 = \|\boldsymbol{x}\|^2 - 2\,\langle \boldsymbol{x},\bar{\boldsymbol{m}}\rangle + \sum_k s_k\,\|\boldsymbol{m}_k\|^2 = \|\boldsymbol{x}-\bar{\boldsymbol{m}}\|^2 + \Big(\sum_k s_k\,\|\boldsymbol{m}_k\|^2 - \|\bar{\boldsymbol{m}}\|^2\Big).$$

Expanding the double sum and using $\sum_k s_k = 1$ twice,

$$\tfrac12\sum_{k,l} s_k s_l\,\|\boldsymbol{m}_k-\boldsymbol{m}_l\|^2 = \tfrac12\sum_k s_k\,\|\boldsymbol{m}_k\|^2 + \tfrac12\sum_l s_l\,\|\boldsymbol{m}_l\|^2 - \langle\bar{\boldsymbol{m}},\bar{\boldsymbol{m}}\rangle = \sum_k s_k\,\|\boldsymbol{m}_k\|^2 - \|\bar{\boldsymbol{m}}\|^2,$$

and the diagonal terms vanish while off-diagonal pairs are counted twice, so $\frac12\sum_{k,l} = \sum_{k<l}$. The normalizing constants multiply to $\big((2\pi)^{-p/2}\big)^{\sum_k s_k} = (2\pi)^{-p/2}$, which together with $e^{-\|\boldsymbol{x}-\bar{\boldsymbol{m}}\|^2/2}$ reconstitutes $\varphi(\boldsymbol{x};\bar{\boldsymbol{m}})$. The second statement follows by pulling the constant out of the integral and recognizing the remaining integral as an expectation. □

Applied in the y variable ($p=1$), Lemma D.2 is the only place where the regression factor enters the cross-group analysis. For fixed $\boldsymbol{x}$ we write $u_\bullet := \boldsymbol{x}^\top\boldsymbol{\theta}_\bullet$ for the conditional mean of y under the component $\bullet$.

Lemma D.3 (Regression factors)**.** *Let j be the data cluster, $\ell \neq j$, $i \in S_\ell$, $k \in S_j$ and $r \neq j$, and assume* (17)*. For fixed $\boldsymbol{x}$, with κ as in* (18)*:*

1. $\int \varphi_1(y;u_j^*)\,\varphi_1(y;u_i)^{1/2}\,\varphi_1(y;u_k)^{-1/2}\,dy \;\le\; e^{\kappa\|\boldsymbol{x}\|^2/2}$*, and the Gaussian produced by Lemma D.2 is centered at $\bar{u}$ with $|\bar{u}-u_j^*| \le (1+D_\theta)\,\|\boldsymbol{x}\|$;*
2. $\int \varphi_1(y;u_j^*)^{1/2}\,\varphi_1(y;u_i)\,\varphi_1(y;u_\ell^*)^{-1/2}\,dy \le e^{\kappa\|\boldsymbol{x}\|^2/2}$*;*
3. $\int \varphi_1(y;u_j^*)^{1/2}\varphi_1(y;u_r^*)^{1/2}\,dy = e^{-(u_j^*-u_r^*)^2/8} \le 1.$

Proof. All three are instances of Lemma D.2 with $p=1$ and W being constant; the resulting Gaussian integrates to one, so each integral equals the exponential prefactor $\exp\big(-\frac{1}{2}\sum_{k<l} s_k s_l (u_k-u_l)^2\big)$.

(1) Here $s=(1,\frac{1}{2},-\frac{1}{2})$, which sums to one, and the exponent is $-\frac{1}{2}\big[\frac{1}{2}(u_j^*-u_i)^2-\frac{1}{2}(u_j^*-u_k)^2-\frac{1}{4}(u_i-u_k)^2\big]$. Put $a:=u_j^*-u_k$ and $b:=u_j^*-u_i$, so $u_i-u_k=a-b$; the exponent equals

$$-\tfrac{1}{4}b^2+\tfrac{1}{4}a^2+\tfrac{1}{8}(a-b)^2=\tfrac{3}{8}a^2-\tfrac{1}{4}ab-\tfrac{1}{8}b^2 \;\le\; |a|\big(|a|+|b|\big).$$

By (17) and Cauchy–Schwarz, $|a|\le B_\theta\,\|\boldsymbol{x}\|$ (as $k\in S_j$) and $|b|\le (D_\theta+B_\theta)\,\|\boldsymbol{x}\|$ (as $i\in S_\ell$, via $\|\boldsymbol{\theta}_j^*-\boldsymbol{\theta}_i\|\le D_\theta+B_\theta$), whence $|a|(|a|+|b|)\le B_\theta(D_\theta+2B_\theta)\,\|\boldsymbol{x}\|^2=\kappa\,\|\boldsymbol{x}\|^2/2$. The center is $\bar{u}=u_j^*+\frac{1}{2}(u_i-u_k)$ with $|u_i-u_k|\le(D_\theta+2B_\theta)\,\|\boldsymbol{x}\|$, so $|\bar{u}-u_j^*|\le\frac{1}{2}(D_\theta+2)\,\|\boldsymbol{x}\|\le(1+D_\theta)\,\|\boldsymbol{x}\|$ using $B_\theta\le 1$.

(2) Here $s=(\frac{1}{2},1,-\frac{1}{2})$. With $e:=u_j^*-u_\ell^*$ and $\delta:=u_i-u_\ell^*$, so that $|\delta|\le B_\theta\,\|\boldsymbol{x}\|$ and $|e|\le D_\theta\,\|\boldsymbol{x}\|$, the exponent is $-\frac{1}{4}(e-\delta)^2+\frac{1}{8}e^2+\frac{1}{4}\delta^2=-\frac{1}{8}e^2+\frac{1}{2}e\delta\le\frac{1}{2}D_\theta B_\theta\,\|\boldsymbol{x}\|^2\le\kappa\,\|\boldsymbol{x}\|^2/2$.

(3) Here $s=(\frac{1}{2},\frac{1}{2})$ and the prefactor is the coefficient $e^{-(u_j^*-u_r^*)^2/8}\le 1$. □

D.3 The cell ratio bounds

Lemma D.4 (Gaussian ratio bounds)**.** *Assume* (17)*, $\kappa(\Lambda+B^2)\le c_0$ with c_0 small, and $\Delta\ge 32B$. Fix a degree K. Then for every data cluster j, every $\ell\neq j$, $i\in S_\ell$, $k\in S_j$, $r\neq j$, and every $W_K(\boldsymbol{z})=(1+\|\boldsymbol{x}-\boldsymbol{\mu}_j^*\|+|\epsilon|)^K$ with $\epsilon=y-\boldsymbol{x}^\top\boldsymbol{\theta}_j^*$, the three bounds* (37) *hold with* $\mathsf{A}_{\mathrm{g}}=C_K\big((1+B)(1+D_{\max})(1+D_\theta)(1+d)\big)^{CK}$ *and $c=1/16$.*

Proof. Write $D:=\|\boldsymbol{\mu}_\ell^*-\boldsymbol{\mu}_j^*\|\ge\Delta$ and $\boldsymbol{Z}:=\boldsymbol{x}-\boldsymbol{\mu}_j^*$; the hypotheses give $D\ge\Delta\ge 32B$ and, by Lemma A.1(iii), $2\kappa\le\frac{1}{2}$.

First bound. Integrate y first. By Lemma D.3(1) and Lemma D.2, for fixed $\boldsymbol{x}$,

$$\int \varphi_1(y;u_j^*)\varphi_1(y;u_i)^{1/2}\varphi_1(y;u_k)^{-1/2}\big(1+\|\boldsymbol{Z}\|+|y-u_j^*|\big)^K dy \\ \le\; e^{\kappa\|\boldsymbol{x}\|^2/2}\,\mathbf{E}_{y\sim N(\bar{u},1)}\big(1+\|\boldsymbol{Z}\|+|y-u_j^*|\big)^K.$$

Inside the expectation, $|y-u_j^*|\le|y-\bar{u}|+(1+D_\theta)\,\|\boldsymbol{x}\|$ by Lemma D.3(1), and the elementary inequality $A+B\le A(1+B)$ for $A\ge 1$, $B\ge 0$, applied twice, gives

$$1+\|\boldsymbol{Z}\|+|y-u_j^*| \;\le\; (1+\|\boldsymbol{Z}\|)\big(1+(1+D_\theta)\,\|\boldsymbol{x}\|\,\big)\big(1+|y-\bar{u}|\big).$$

Raising to the Kth power, factoring out the two deterministic factors, and using that $y-\bar{u}$ is standard normal, so that $\mathbf{E}(1+|y-\bar{u}|)^K=:C_K$ is an absolute constant, the y integral is at most $e^{\kappa\|\boldsymbol{x}\|^2/2}W'(\boldsymbol{x})$ with

$$W'(\boldsymbol{x}):=C_K(1+\|\boldsymbol{Z}\|)^K\big(1+(1+D_\theta)\,\|\boldsymbol{x}\|\,\big)^K.$$

It remains to bound $\int \varphi(\boldsymbol{x};\boldsymbol{\mu}_j^*)\varphi(\boldsymbol{x};\boldsymbol{\mu}_i)^{1/2}\varphi(\boldsymbol{x};\boldsymbol{\mu}_k)^{-1/2}e^{\kappa\|\boldsymbol{x}\|^2/2}W'(\boldsymbol{x})\,d\boldsymbol{x}$. By Lemma D.2 with $s=(1,\frac{1}{2},-\frac{1}{2})$ this equals $M\cdot\mathbf{E}_{\boldsymbol{x}\sim N(\boldsymbol{c},I)}\big[e^{\kappa\|\boldsymbol{x}\|^2/2}W'(\boldsymbol{x})\big]$ with $\boldsymbol{c}=\boldsymbol{\mu}_j^*+\frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_k)$ and

$$\begin{aligned}M&=\exp\Big(-\tfrac{1}{4}\left\|\boldsymbol{\mu}_j^*-\boldsymbol{\mu}_i\right\|^2+\tfrac{1}{4}\left\|\boldsymbol{\mu}_j^*-\boldsymbol{\mu}_k\right\|^2+\tfrac{1}{8}\left\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_k\right\|^2\Big)\\&\le\exp\Big(-\tfrac{1}{4}(D-B)^2+\tfrac{1}{4}B^2+\tfrac{1}{8}(D+2B)^2\Big)=e^{-D^2/8+DB+B^2/2},\end{aligned}$$

using $\left\|\boldsymbol{\mu}_j^*-\boldsymbol{\mu}_i\right\|\ge D-B$, $\left\|\boldsymbol{\mu}_j^*-\boldsymbol{\mu}_k\right\|\le B$ and $\left\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_k\right\|\le D+2B$, all consequences of (17) and the triangle inequality.

For the expectation, $\|\boldsymbol{c}\|\le 2D_{\max}+B$ (since $\|\boldsymbol{\mu}_i\|,\|\boldsymbol{\mu}_k\|\le D_{\max}+B$), and $\|\boldsymbol{x}\|^2$ is noncentral chi-square under $N(\boldsymbol{c},I)$, so for $2\kappa\le\frac{1}{2}$,

$$\mathbf{E}e^{\kappa\|\boldsymbol{x}\|^2}=(1-2\kappa)^{-d/2}\exp\Big(\frac{\kappa\|\boldsymbol{c}\|^2}{1-2\kappa}\Big)\le e^{2\kappa d+2\kappa(2D_{\max}+B)^2}\le e^{C\kappa(\Lambda+B^2)}\le e^{Cc_0}\le C,$$

where $-\log(1-u)\le 2u$ for $u\le\frac{1}{2}$ was used. Moreover, since $\left\|\boldsymbol{x}-\boldsymbol{\mu}_j^*\right\|\le\left\|\boldsymbol{c}-\boldsymbol{\mu}_j^*\right\|+\|\boldsymbol{x}-\boldsymbol{c}\|\le D_{\max}+B+\|\boldsymbol{x}-\boldsymbol{c}\|$ and $\|\boldsymbol{x}\|\le 2D_{\max}+B+\|\boldsymbol{x}-\boldsymbol{c}\|$, factoring as above and using that $\|\boldsymbol{x}-\boldsymbol{c}\|$ has moments of every order $p\le 4K$ bounded by $C_K(1+d)^{2K}$ gives $\mathbf{E}W'(\boldsymbol{x})^2\le C_K\big((1+D_\theta)(1+D_{\max})(1+B)(1+d)\big)^{CK}$. Cauchy–Schwarz therefore yields $\mathbf{E}\big[e^{\kappa\|\boldsymbol{x}\|^2/2}W'\big]\le\mathsf{A}_{\mathrm{g}}$, so the first bound is at most $\mathsf{A}_{\mathrm{g}}e^{-D^2/8+DB+B^2/2}\le\mathsf{A}_{\mathrm{g}}e^{-D^2/16}$: indeed, for $B\le D/32$,

$$-\tfrac{1}{8}D^2+DB+\tfrac{1}{2}B^2\ \le\ -\tfrac{1}{16}D^2.$$

Second bound. Identical, with Lemma D.3(2) for the y integral and Lemma D.2 with $s=(\frac{1}{2},1,-\frac{1}{2})$ for the $\boldsymbol{x}$ integral, whose constant is

$$\begin{aligned}M&=\exp\Big(-\tfrac{1}{4}\left\|\boldsymbol{\mu}_j^*-\boldsymbol{\mu}_i\right\|^2+\tfrac{1}{8}D^2+\tfrac{1}{4}\left\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_\ell^*\right\|^2\Big)\\&\le\exp\Big(-\tfrac{1}{4}(D-B)^2+\tfrac{1}{8}D^2+\tfrac{1}{4}B^2\Big)=e^{-D^2/8+DB/2},\end{aligned}$$

and whose center $\boldsymbol{c}=\frac{1}{2}\boldsymbol{\mu}_j^*+\boldsymbol{\mu}_i-\frac{1}{2}\boldsymbol{\mu}_\ell^*$ again satisfies $\|\boldsymbol{c}\|\le 2D_{\max}+B$; the factor $\mathbf{E}e^{\kappa\|\boldsymbol{x}\|^2/2}\le C$ is bounded as before.

Third bound. As a density in $\boldsymbol{z}$, φ_r^* is Gaussian with a covariance that depends on $\boldsymbol{\theta}_r^*$, so Lemma D.2 does not apply in $\boldsymbol{z}$ directly. Integrating y first, Lemma D.3(3) bounds the inner integral by one, and Lemma D.2 in $\boldsymbol{x}$ then gives $\mathbf{E}_j\big[(\varphi_r^*/\varphi_j^*)^{1/2}\big]=\int\sqrt{\varphi_r^*\varphi_j^*}\,d\boldsymbol{z}\le\int\sqrt{\varphi(\boldsymbol{x};\boldsymbol{\mu}_r^*)\varphi(\boldsymbol{x};\boldsymbol{\mu}_j^*)}\,d\boldsymbol{x}=e^{-\left\|\boldsymbol{\mu}_r^*-\boldsymbol{\mu}_j^*\right\|^2/8}\le e^{-\Delta^2/8}$. □

E Additional Results

E.1 The expert part of the projection is a quadratic form

Proposition E.1. *Fix ℓ and write $\bar\beta^{(\ell)}(\boldsymbol{z}):=\sum_{i\in S_\ell}\psi_i^{(\ell)}(\boldsymbol{z})(\boldsymbol{\theta}_i-\boldsymbol{\theta}_\ell^*)$. Then the expert part of the one-cluster projection* (33) *satisfies*

$$P_{\ell,\theta}=-\mathbf{E}_\ell\Big[\sum_{i\in S_\ell}\psi_i^{(\ell)}(\boldsymbol{z})\,r_i(\boldsymbol{z})\,\boldsymbol{x}^\top(\boldsymbol{\theta}_i-\boldsymbol{\theta}_\ell^*)\Big]=\mathbf{E}_\ell\Big[\big(\boldsymbol{x}^\top\bar\beta^{(\ell)}(\boldsymbol{z})\big)^2\Big]\ \ge\ 0.$$

In particular, up to the cross-group errors of Lemma B.5, a gradient step never moves the experts of a group away from the true expert of that group on average.

Proof. Work in the recentered coordinates of Remark 2. By (35), $P_{\ell,\theta}=\mathbf{E}_\nu\big[\sum_i\psi_i(a_i^2-y'a_i)\big]$ with $a_i=\boldsymbol{x}^\top(\boldsymbol{\theta}_i-\boldsymbol{\theta}_\ell^*)$. The Stein computation in part (b) of the proof of Lemma C.3 gives $\mathbf{E}_\nu\big[y'\sum_i\psi_i a_i\big]=\mathbf{E}_\nu V_y$, while $\sum_i\psi_i a_i^2=V_y+\bar a^2$. The variances cancel and $P_{\ell,\theta}=\mathbf{E}_\nu\bar a^2=\mathbf{E}_\ell\big[(\boldsymbol{x}^\top\bar\beta^{(\ell)})^2\big]\ge 0$. □

E.2 Invariance of the groups during the global phase

Proposition E.2. *Under the assumptions of Proposition B.4 and $\eta \le c/A$, every $i \in S_\ell$ satisfies $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| \le B < \Delta/4$ and $\|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_\ell^*\| \le B + 1 < \Delta/2$; consequently $S_\ell(\boldsymbol{\mu}^+) = S_\ell(\boldsymbol{\mu})$ for every ℓ.*

Proof. By (17), $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| \le B$ for $i \in S_\ell$, and $B < \Delta/4$ by Lemma A.1(iv). The single-component gradient is a block of the full gradient, so $\|\nabla_{\boldsymbol{\mu}_i}\mathcal{L}\| \le \|g\|_F \le \sqrt{2AL_0}$ by (49) and $\mathcal{L} \le L_0$. Hence, by the triangle inequality and $\eta \le c/A$,

$$\|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_\ell^*\| \;\le\; \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_\ell^*\| + \eta\,\|\nabla_{\boldsymbol{\mu}_i}\mathcal{L}\| \;\le\; B + \frac{c}{A}\sqrt{2AL_0} = B + c\sqrt{\frac{2L_0}{A}} \;\le\; B + 1,$$

because $A \ge L_0$ by (14) and c is small. For $j \ne \ell$, the reverse triangle inequality gives $\|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_j^*\| \ge \Delta - (B+1) > B + 1 \ge \|\boldsymbol{\mu}_i^+ - \boldsymbol{\mu}_\ell^*\|$, where the strict inequality is $\Delta > 2(B+1)$ from Lemma A.1(iv). Thus $\boldsymbol{\mu}_\ell^*$ remains the strictly nearest true center of $\boldsymbol{\mu}_i^+$, and no component changes group. □

F Experimental details

All experiments sample data from (2) with $d = 10$ and run Algorithm 1 with the population expectations replaced by averages over $N = 10{,}000$ training points, for at most $T = 10{,}000$ iterations. Table 1 lists the settings of each experiment.

Mixing-weight update. Each outer iteration performs four EM fixed-point updates of the fitted weights, holding the centers and experts fixed: $\pi_i \leftarrow N^{-1}\sum_{k=1}^N \psi_i(\boldsymbol{z}_k)$, where ψ_i is the current joint responsibility. These four updates are not an exact solve of the weight subproblem. The balancing and pruning runs penalize $\lambda D_{\mathrm{KL}}(\bar{\boldsymbol{\rho}}\|\mathrm{Unif}_n)$, where $\bar\rho_i = N^{-1}\sum_k P(i \mid \boldsymbol{x}_k)$ is the mean input-only router probability on training data. Their four inner weight updates use safeguarded generalized EM: an ordinary EM proposal and a backtracked gradient proposal are evaluated against the penalized objective, and an improving proposal is retained. The penalty used here is distinct from $\lambda D_{\mathrm{KL}}(\mathrm{Unif}_n\|\boldsymbol{\pi})$; the latter prior-weight penalty was not used in these runs. Centers and experts take one gradient step per outer iteration. The heatmap uses a fixed step size; the other experiments start each step at 0.02 and backtrack as needed.

Ground truth. In the heatmap, centers are proposed sequentially from $\mathcal{N}(0, (8/\sqrt{2})^2 I_d)$ and accepted only if they are at least 6.4 from every previously accepted center; the accepted centers are then recentered. Thus 8 is the placement scale, while the realized minimum separation is at least 6.4, not exactly 8. This rejection construction also applies when $m = 50$. Heatmap experts have independent $\mathcal{N}(0, 0.5^2 I_d)$ coefficients, and its true mixing weights follow $\mathrm{Dirichlet}(3, \ldots, 3)$.

For the other experiments, the centers are a randomly rotated regular m-gon, with independent Gaussian coordinate jitter of standard deviation 0.08, recentered and rescaled so that their minimum pairwise distance equals Δ exactly. True expert coefficients are independent standard Gaussians, centered across experts and rescaled to coordinate root-mean-square s_θ, as specified in Table 1. The warm-start experiment uses uniform true mixing weights. The balancing and pruning experiments use either uniform weights or the skewed prior

$$\boldsymbol{\pi}^* = (0.000001,\ 0.07367372,\ 0.11166848,\ 0.16925777,\ 0.25654680,\ 0.38885223),$$

with $\sum_\ell (\pi_\ell^*)^2 = 0.2636$, i.e. $n_{\mathrm{eff}}^* = 3.794$. Its smallest component has expected training count 0.01; the skewed setting does not measure reliable recovery of that rare expert. The local least-squares start fits $\boldsymbol{\theta}_i^{(0)}$ on the observations within radius R of $\boldsymbol{\mu}_i^{(0)}$, with no intercept and a $10^{-8} I_d$ numerical ridge. Initial fitted weights are uniform. Initial centers are sampled from the training inputs without replacement for the heatmap and independently from the true input marginal for the other experiments.

Evaluation. Since p_* is known, joint KL is estimated by $N_{\mathrm{test}}^{-1}\sum_k \log[p_*(\boldsymbol{z}_k)/p(\boldsymbol{z}_k)]$ and the conditional KL of Figure 4 by $N_{\mathrm{test}}^{-1}\sum_k \log[p_*(y_k \mid \boldsymbol{x}_k)/p(y_k \mid \boldsymbol{x}_k)]$, over $N_{\mathrm{test}} = 20{,}000$ held-out points in the warm-start, balancing, and pruning experiments. The heatmap instead estimates joint

Table 1: Settings of the experiments. The dimension is $d = 10$, the sample size $N = 10{,}000$, and the iteration cap $T = 10{,}000$ throughout.

	Figure 1	Figure 2	Figures 3, 4
true experts m	$2, \dots, 50$	6	6
fitted experts n	$2, 4, \dots, 100$	$4, 8, \dots, 24$	16
separation Δ	minimum ≥ 6.4 (scale 8)	2, 12 (exact)	8 (exact)
expert scale s_θ	0.5	1	0.5
mixing weights $\boldsymbol{\pi}^*$	$\mathrm{Dirichlet}(3, \dots, 3)$	uniform	uniform; skewed
expert initialization	local least squares	local least squares; $\mathcal{N}(0, I_d)$	local least squares
window radius R	$2.5\sqrt{d + \log(2n)}$	$2.5\sqrt{d + \log(2n)}$	$2.5\sqrt{d + \log(2n)}$
step size η	fixed 0.02	initial 0.02; backtracking	initial 0.02; backtracking
balancing strength λ	–	–	$0, 0.1, 0.2, 0.4, 0.6, 0.8, 1$
trials / seeds	35 per cell	20	20
success criterion	final joint KL < 0.02 (train)	last three joint KL ≤ 0.02 (test)	last three joint KL ≤ 0.02 (test)

KL on its training points. The other experiments evaluate every 200 updates; success requires held-out joint KL at most 0.02 at each of the last three available checkpoints. Balancing models use the full budget. Warm-start and pruning models stop early after three consecutive *validation* joint-KL estimates at most 0.02, using another independent sample of $20{,}000$ points. Pruning retains the k experts with largest fitted π_i, renormalizes their weights, and evaluates conditional KL without retraining. Figure 4 plots arithmetic means across all 20 seeds, including unsuccessful fits, without confidence bands. Success-probability intervals in Figures 2 and 3 are pointwise 95% Wilson intervals; effective-count intervals in Figure 3 are pointwise 95% Student-t intervals for the seed mean.

Compute and code. The warm-start, balancing, and pruning runs used one NVIDIA A40 GPU, float32 training, and batches of at most 32 independent trials; initialization and evaluation used float64. Across their full saved grids, recorded amortized GPU training times were 0.17, 2.39, and 2.05 hours, respectively, and recorded evaluation times were 0.25, 1.11, and 1.27 hours. These are measured components rather than end-to-end job wall times, which were not retained. The heatmap runtime was not retained.