ADAPTIVE MEAN ESTIMATION BY IN-CONTEXT LEARNING: A GRADIENT-FLOW ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior Fitted Networks (PFNs) such as TabPFN (Zhang et al., 2025) now rival established statistical procedures across prediction and estimation tasks. A natural explanation is that PFNs have the property of *statistical adaptivity*, that is they perform nearly as well as a method tailored to the true data-generating model for a heterogeneous set of models, while not being told which model the data comes from. Prior work does not give a conclusive explanation, how gradient-based pretraining learns an adaptive estimator. We study how such adaptivity is learned in a controlled location-estimation problem. Each task is an unlabeled sample whose family is hidden: Gaussian data call for averaging, with error of order n^{-1}, whereas uniform data are best estimated from their extremes, at the faster rate n^{-2}. We also provide the example of a symmetric Gaussian mixture, for which a rate of σ_n^2/n can be attained. On scalar inputs, softmax attention computes the derivative of the empirical cumulant-generating function. A single primitive therefore both supplies features that distinguish the families and forms estimators interpolating between the sample mean and the mid-range. Building on this, we combine attention experts through either a softmax mixture of experts or a gated linear unit (GLU), and analyze stagewise gradient flow. With $\widetilde{\Omega}(n^{1+\epsilon})$ pretraining tasks, the learned estimator is asymptotically efficient on Gaussian tasks, within a factor n^{ϵ} of the minimax rate on uniform tasks, and order-optimal on mixtures in a shrinking-variance regime. These guarantees extend to new locations and longer contexts. A risk decomposition separates expert error, routing error and normalization error, which clarifies the architectural contrast. Softmax gating enforces normalization and exact translation equivariance, whereas the GLU must learn it: its dynamics separate into fast bias removal followed by slow expert selection. End-to-end experiments recover the predicted specialization.

1 INTRODUCTION

In a landmark paper, Brown et al. (2020) showed that large language models can learn to perform new tasks from examples supplied in a prompt (also called a context), without updating their parameters at inference time, a capability they called in-context learning (ICL). Garg et al. (2023) formulated ICL as a new paradigm for statistical learning: whereas the classical approach fits a *predictor* to each new dataset through parameter optimization, ICL pretrains a reusable prediction *procedure* across datasets and applies it to a new dataset through context alone. Tabular foundation models (TFMs) have made this paradigm practical for tabular prediction, with evaluations on the broad TabArena and TALENT benchmarks showing that TFMs can even outperform previously state-of-the-art methods such as boosted tree ensembles (Erickson et al., 2025; Ye et al., 2025). Other recent studies show how the ICL paradigm can outperform well-established classical methods in estimating various statistical quantities, including heterogeneous treatment effects (Balazadeh et al., 2025), survival curves (Qi et al., 2026), mutual information (Peyrard & Cho, 2025), and Poisson means in empirical Bayes problems (Teh et al., 2025). These works suggest that the central theoretical question for ICL should not be whether it can learn to approximate an existing algorithm or statistical procedure, but *how it is able to outperform established procedures across diverse statistical problems*.

Several theoretical explanations have been developed by analyzing empirical risk minimization (ERM) of transformers. Bai et al. (2023) show that transformers can implement in-context algorithm selection,

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.

including choosing among ridge penalties using a train-validation split. Wakayama & Suzuki (2025) and Ma et al. (2025) establish approximation of the posterior predictive mean under hierarchical priors over regression-function classes. In particular, Ma et al. (2025) use posterior contraction to show that this predictor achieves optimal rates corresponding to the unknown smoothness and effective dimension of the task. These results suggest that part of ICL's strength may lie in its ability to adjust the statistical procedure to the structure revealed by the context. This is a form of *statistical adaptivity*, which is when a single estimator attains a close to optimal accuracy appropriate to the underlying model for multiple models, without knowing that model in advance. Such adaptation is a longstanding objective in statistics, often achieved through procedures tailored to particular models and inferential targets (Bickel et al., 1993). Pretraining offers a potentially general recipe to acquiring this behavior, but guarantees for ERM do not establish that gradient-based training learns the required mechanisms.

Understanding how such adaptation is learned requires an analysis of pretraining itself. A complementary literature studies pretraining dynamics directly. Results include gradient-flow convergence for linear regression with linear attention (Zhang et al., 2024a;b), task allocation in multi-head softmax attention for multi-task linear regression (Chen et al., 2024a), and gradient-based learning of shared low-dimensional structure for nonlinear single-index regression (Oko et al., 2024). These results explain important forms of learned statistical behavior, including specialization and adaptation to shared structure.

In this paper, *we study how adaptation can emerge from gradient-based pretraining, focusing on a simple one-dimensional location-estimation problem*. Here, each task draws a context comprising n i.i.d. observations from either a Gaussian or a uniform distribution, and the estimation target is the location parameter of the underlying distribution. The appropriate estimator depends on the distributional family. For Gaussian observations, a sample average is optimal, with mean-squared error of order n^{-1}, whereas uniform observations permit the faster n^{-2} rate through the sample mid-range. Neither estimator attains both rates. Moreover, the family label is hidden and has to be inferred from the context. During pretraining, the learner observes the target mean but receives neither the family label nor a prescribed estimator choice. We also consider a symmetric two-component Gaussian mixture as the alternative to the Gaussian family. Its shrinking component variance σ_n^2 permits risk of order σ_n^2/n, providing a second example of adaptation to distributional structure.

Multi-head attention can be used to identify the distribution of the samples in context as well as to estimate their location. With scalar observations, the query and key matrices of an attention head collapse to a single scalar which we denote by λ. Thus, attention effectively computes the exponentially tilted average $h(\mathbf{x}, \lambda) = \sum_i x_i e^{\lambda x_i} / \sum_i e^{\lambda x_i}$. This is the derivative of the empirical cumulant-generating function (CGF). Since the population CGF can be used to characterize distributions, evaluating this function at several locations supplies features for distinguishing between the two families. The same operation also constructs the relevant estimators: setting $\lambda = 0$ gives the sample mean, while sending $\lambda \to \pm\infty$ approaches the sample maximum and minimum, which are required to compute the mid-range.

We use these two roles of multi-head attention to construct a bank of estimators, or *experts*, as well as a feature embedding, that allows distinguishing different distributions. A learned gating parameter θ_i is used within a gating layer, either a Softmax MoE or a Gated Linear Unit, to output a weighted combination of the outputs of the different experts. This weighted sum of the expert estimates forms the location estimate. Both of these architectures are minor modifications of commonly used layers within standard Transformer architectures (Shazeer, 2020). We analyze stagewise gradient flow that first learns the gating parameters θ_i with the expert parameters λ_i fixed, then learns the expert parameters λ_i with the gating parameters θ_i fixed. This is also done to simplify the proof structure, while our experiments demonstrate that this is not strictly necessary. Notably, our results are stronger than an expressivity statement, as they demonstrate that gradient flow on the empirical loss is able to learn how to do in context location estimation with close to optimal rates.

Our contributions are fourfold:

- **Adaptive estimation and transfer.** We establish family-specific risk guarantees for the output of stagewise pretraining under both the Softmax MoE and the GLU architecture. Both architectures achieve Gaussian asymptotic efficiency and nearly (up to a n^{ϵ} factor) optimal uniform estimation rates. For a Gaussian mixture distribution, we also obtain

order-optimal rates. This implies that our architecture performs nearly as well as the sample mean on Gaussian data and the sample mid-range on uniform data. We also establish out-of-distribution risk bounds for longer contexts as well as for samples with a fixed location.

- **Statistical mechanism and architectural comparison.** Our results show how the risk can be decomposed into the variance of the experts, the bias of the final estimator, and the error accumulated through the gating mechanism (Lemma 3.5). They explain how distribution-sensitive weighting preserves the accuracy of the appropriate expert, and how the architectures differ: softmax enforces normalization of the weights and thus exact translation equivariance, whereas the GLU must learn approximate normalization of the weights (Theorem 3.1).
- **Learning dynamics.** We show how training learns the gating parameters and expert parameters required by the statistical analysis (Lemma 3.3 and Lemma 3.4). Since the loss admits a bias-variance decomposition, our analysis tracks the movement of the bias and variance separately. The major terms of, the bias are created by the sum of the weights, while variance is accumulated through the variance of the separate experts. Additionally, we observe that the bias is scaled at a different order, proportional to the prior variance σ_μ^2, than the variance which is of order $\frac{1}{n}$ where n is the sample size. Thus, bias and variance move at different timescales. We exploit this fact to characterize the dynamics of the parameters (Lemma B.8 and Lemma B.9). Subsequent training of the attention layer lets the experts specialize their predictions to their assigned distribution (Lemma 3.4). Here, our analysis exploits that the second layer has already converged to a near-perfect gating of the different distributions.
- **Finite Sample Guarantees.** Our analysis demonstrates that for $m = \widetilde{\Omega}(n^{1+\epsilon})$ pretraining tasks, the gradient flow on the empirical loss $\mathcal{L}_m$ converges to the correct solution.

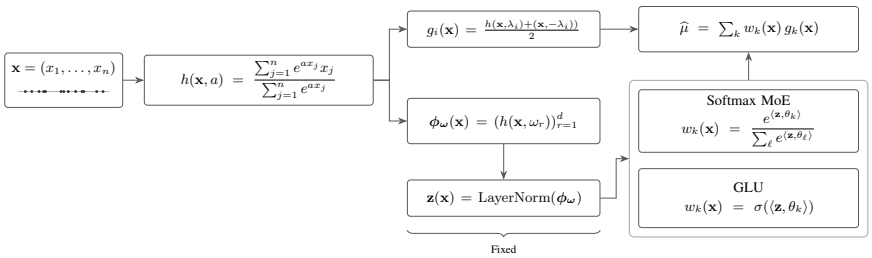

Figure 1: Illustration of our studied architecture. We also highlight the parts which were fixed for simplicity.

Notation. We write $[n] = \{1, \ldots, n\}$. The norms $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ denote the Euclidean and Frobenius norms, respectively. Parameter collections and their gradients are identified with stacked vectors. Let e_i be the i-th coordinate vector. The notation $\mathcal{O}$, Θ, Ω, and o has its usual meaning, $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ suppress logarithmic factors, and $\lesssim$ denotes inequality up to an absolute multiplicative constant. Constants may change from line to line. Uppercase X_i and $\mathbf{X}$ denote random observations and contexts, while lowercase x_i and $\mathbf{x} = (x_1, \ldots, x_n)$ denote their realizations. We define estimators as functions of $\mathbf{x}$ and evaluate their risks at the random context $\mathbf{X}$. We use $\mathbb{E}_j[\,\cdot\,] = \mathbb{E}[\,\cdot \mid \tau = j]$ for expectation under family j, including both the location prior and the conditional sample. We write $\mathbb{P}_j$ for this family-conditional law. For $1 \le p < \infty$, $\|Y\|_{L^p(\mathbb{P}_j)} = (\mathbb{E}_j[|Y|^p])^{1/p}$, using the Euclidean norm for random vectors and the unconditional law when the family subscript is omitted.

2 Problem and Architectures

We consider a location-estimation task. During training and inference, the learner is presented with observations from a distribution in context and is tasked with estimating their location. The learner is not explicitly told which distribution the observations are from. To allow a fair comparison, the distributions differ only in shape not in variance.

Definition 2.1 (Adaptive mean-estimation tasks)**.** Sample a location μ and a family label τ independently:

$$\mu \sim \mathcal{N}(0, \sigma_\mu^2), \qquad \mathbb{P}(\tau = \mathrm{G}) = \mathbb{P}(\tau = \mathrm{A}) = \tfrac{1}{2}$$

Conditional on (μ, τ), draw $\xi_i \overset{\text{i.i.d.}}{\sim} Q_\tau$ and set

$$X_i = \mu + \xi_i, \qquad i \in [n], \qquad \mathbf{X} = (X_1, \ldots, X_n).$$

The centered Gaussian law is $Q_{\mathrm{G}} = \mathcal{N}(0, 1)$. The alternative centered law Q_{A} is either

$$\text{Uniform:} \quad Q_{\mathrm{A}} = Q_{\mathrm{U}} = \mathrm{Unif}[-\sqrt{3}, \sqrt{3}],$$

$$\text{Gaussian mixture:} \quad Q_{\mathrm{A}} = Q_{\mathrm{M}} = \tfrac{1}{2}\mathcal{N}(-\sqrt{1-\sigma_n^2}, \sigma_n^2) + \tfrac{1}{2}\mathcal{N}(\sqrt{1-\sigma_n^2}, \sigma_n^2),$$

where $\sigma_n^2 \in (0, 1)$. In statements specific to one alternative, we write U or M in place of A. Write $P_{\mu,j}$ for the law of $\mu + \xi$ with $\xi \sim Q_j$. Each observed task (X, μ) is the tuple of the context and the location and thus, the family label τ is never observed.

Risk and Lower Bounds. For $j \in \{\mathrm{G}, \mathrm{A}\}$, define the risk at location μ and its average over the location prior by

$$R_{j,n}(\widehat{\mu}; \mu) := \mathbb{E}_{\mathbf{X} \sim P_{\mu,j}^{\otimes n}}[(\widehat{\mu}(\mathbf{X}) - \mu)^2], \qquad R_{j,n}(\widehat{\mu}) := \mathbb{E}_j[(\widehat{\mu}(\mathbf{X}) - \mu)^2].$$

For the trained estimator, these expectations are over an independent test task with the trained parameters held fixed. For a fixed family j, define the location-minimax risk

$$R_{j,n}^\star := \inf_{\widetilde{\mu}} \sup_{\mu \in \mathbb{R}} \mathbb{E}_{\mathbf{X} \sim P_{\mu,j}^{\otimes n}}[(\widetilde{\mu}(\mathbf{X}) - \mu)^2].$$

The sample mean has risk $1/n$ under every family. For Gaussian data, the sample mean attains the exact location-minimax risk n^{-1}. For uniform data, the sample mid-range attains the faster location-minimax rate n^{-2}. For completeness, we give the proofs of these benchmarks in Section H.1. For the Gaussian mixture, revealing component labels reduces the problem to Gaussian location estimation with variance σ_n^2, yielding the lower bound σ_n^2/n (see Lemma H.2).

2.1 Architecture

We consider a two-layer transformer architecture that consists of one multi-head attention layer followed by a second layer, which is either a Softmax Mixture of Experts (MoE) layer or a Gated Linear Unit. In the following, we will fix some of the parameters of the transformer architecture to decide before training which attention heads will be used as estimators and which will be used to identify the distribution.

Definition 2.2 (Transformer architecture)**.** Let K be the number of experts and $d \geq 3$ be the number of heads dedicated to identifying the distribution family. Let $\mathbf{x}$ be the context of size n, containing the input dataset. For $\lambda \in \mathbb{R}$, define

$$h(\mathbf{x}, \lambda) = \frac{\sum_{i=1}^n e^{\lambda x_i} x_i}{\sum_{i=1}^n e^{\lambda x_i}}. \tag{1}$$

Expert $k \in [K]$ is $g_k(\mathbf{x}) = \frac{1}{2}\left(h(\mathbf{x}, \lambda_k) + h(\mathbf{x}, -\lambda_k)\right)$ with $\lambda_k(0) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[-1, 1]$. For gating, sample $\omega_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, $r \in [d]$, and hold $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ fixed. We define $\boldsymbol{\phi}_{\boldsymbol{\omega}}(\mathbf{x}) = (h(\mathbf{x}, \omega_1), \ldots, h(\mathbf{x}, \omega_d))^\top$ and then apply a LayerNorm with fixed parameters to obtain $\mathbf{z}$

$$\mathbf{z}(\mathbf{x}) = \mathrm{LayerNorm}(\boldsymbol{\phi}_{\boldsymbol{\omega}}(\mathbf{x})) = \frac{P_d \boldsymbol{\phi}_{\boldsymbol{\omega}}(\mathbf{x})}{\|P_d \boldsymbol{\phi}_{\boldsymbol{\omega}}(\mathbf{x})\|_2}, \tag{2}$$

where $P_d = I_d - \mathbf{1}\mathbf{1}^\top/d$. For $k \in [K]$, initialize $\theta_k(0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\theta^2 I_d)$. We will consider two different choices for w, the Softmax Mixture of Experts (3) and the Gated Linear Unit (4):

$$w_k(\mathbf{x}; \theta) = \frac{\exp(\langle \mathbf{z}(\mathbf{x}), \theta_k \rangle)}{\sum_{\ell=1}^K \exp(\langle \mathbf{z}(\mathbf{x}), \theta_\ell \rangle)}, \tag{3}$$

$$w_k(\mathbf{x}; \theta) = \sigma(\langle \mathbf{z}(\mathbf{x}), \theta_k \rangle) = \frac{1}{1 + e^{-\langle \mathbf{z}(\mathbf{x}), \theta_k \rangle}}. \tag{4}$$

Finally, we define the final network output as

$$\widehat{\mu}_{\theta,\lambda}(\mathbf{x}) = \sum_{k=1}^K w_k(\mathbf{x}; \theta) g_k(\mathbf{x}). \tag{5}$$

The parameters $\theta = (\theta_1, \ldots, \theta_K)$ and $\lambda = (\lambda_1, \ldots, \lambda_K)$ are trained, while $\boldsymbol{\omega}$ remains fixed.

The architecture in Definition 2.2 fixes some parameters to remove permutation invariances. The attention heads are split into the experts g_k, which implement the candidate estimators, and the heads forming the feature $\mathbf{z}$, which identifies the distribution family. When the routing feature is replaced by its population value, the unbiased experts g_k give an unbiased Softmax MoE output, whereas the GLU must also learn normalization. For sample-dependent gates, softmax guarantees exact translation equivariance but need not be exactly unbiased. The feature does not depend on the location, $\mathbf{z}(\mathbf{X} + \mu\mathbf{1}) = \mathbf{z}(\mathbf{X})$, so its distribution depends only on the family. Under family j, $\mathbf{z}(\mathbf{X}) \to \bar{\mathbf{z}}_j$ in probability as $n \to \infty$, where $\bar{\mathbf{z}}_j$ is the centered and normalized vector of derivatives of the CGF of Q_j at $\omega_1, \ldots, \omega_d$. Since $\bar{\mathbf{z}}_{\mathrm{G}} \neq \bar{\mathbf{z}}_{\mathrm{A}}$ almost surely (Lemma F.5), $\mathbf{z}$ distinguishes the families for large n.

Our architecture introduces a LayerNorm for the gating input to the second layer, as this simplifies calculations. We believe this is not strictly necessary. For details, see Appendix J.

2.2 Training protocol

For our training dataset, we sample m i.i.d. tasks of size n from Definition 2.1 and denote our dataset by $\{(\mathbf{X}_b, \mu_b)\}_{b=1}^m$.

Definition 2.3 (Stagewise gradient flow training)**.** For either architecture, (3) or (4), define the population loss $\mathcal{L}(\theta, \lambda) = \frac{1}{2}\mathbb{E}[(\widehat{\mu}(\mathbf{X}) - \mu)^2]$ and the empirical loss $\hat{\mathcal{L}}_m(\theta, \lambda) = \frac{1}{2m}\sum_{b=1}^m (\widehat{\mu}(\mathbf{X}_b) - \mu_b)^2$. Training proceeds in two stages:

1. For $0 \le t < t_\theta$, run $\dot{\theta} = -\nabla_\theta \hat{\mathcal{L}}_m$ with λ fixed.
2. For $t_\theta \le t < t_\lambda$, run $\dot{\lambda} = -\nabla_\lambda \hat{\mathcal{L}}_m$ with θ fixed.

The full population flow is the same procedure with $\hat{\mathcal{L}}_m$ replaced by $\mathcal{L}$ in both stages.

3 Main results

Theorem 3.1 (Main result; proof in Section B.5)**.** *Let $K = 2$ be the number of experts. Fix any $\eta, \epsilon \in (0, 1)$, a gating dimension $d \ge 3$, and sufficiently small $\sigma_\mu^2 > 0$, all independent of n. Sample $m = \widetilde{\Omega}(n^{1+\epsilon})$ independent tasks according to Definition 2.1, with $m \in \mathrm{poly}(n)$. In the Gaussian-mixture case, assume $n^{-1+\epsilon} \le \sigma_n^2 \le o(1)$. Initialize as in Definition 2.2, with fixed $\sigma_\theta^2 > 0$, sufficiently small depending on η. Relabel the experts and permute θ accordingly so that $|\lambda_1(0)| < |\lambda_2(0)|$. Let $\widehat{\mu}_{\mathrm{trained}} = \widehat{\mu}_{\theta(t_\theta),\lambda(t_\lambda)}$, with t_λ as in Lemma 3.4. For sufficiently large n, with probability at least $1 - \eta$ over the initialization and training tasks,*

$$R_{\mathrm{G},n}(\widehat{\mu}_{\mathrm{trained}}) = \frac{1 + o(1)}{n}, \qquad R_{\mathrm{A},n}(\widehat{\mu}_{\mathrm{trained}}) = \begin{cases} \mathcal{O}(n^{-2+\epsilon}), & \textit{uniform alternative}, \\ \mathcal{O}(\sigma_n^2/n), & \textit{Gaussian-mixture alternative}. \end{cases}$$

For the GLU, we have, for every $\mu \in \mathbb{R}$ and $j \in \{\mathrm{G}, \mathrm{A}\}$,

$$R_{j,n}(\widehat{\mu}_{\mathrm{trained}}; \mu) \le 2R_{j,n}(\widehat{\mu}_{\mathrm{trained}}) + \frac{C\mu^2}{n^2}.$$

For the softmax MoE, for every $\mu \in \mathbb{R}$ and $j \in \{\mathrm{G}, \mathrm{A}\}$,

$$R_{j,n}(\widehat{\mu}_{\mathrm{trained}}; \mu) = R_{j,n}(\widehat{\mu}_{\mathrm{trained}}).$$

The Gaussian and Gaussian-mixture rates match the lower bounds of Proposition H.1 and Lemma H.2 up to constants, and the uniform rate is within a factor of n^ϵ of the minimax rate n^{-2}. Table 1 summarizes the results. The Softmax MoE architecture has a stronger guarantee when $|\mu|$ is large. This follows from its exact translation equivariance, while the GLU has to learn approximate normalization.

Corollary 3.2 (Length generalization; proof in Section D.3.7)**.** *Under the assumptions of Theorem 3.1, for sufficiently large n, with probability at least $1 - \eta$ over the initialization and pretraining tasks,*

Table 1: The terminal expert parameters and fresh-context risks certified by the main theorem.

Alternative family	terminal $\lvert\lambda_2\rvert$	Gaussian risk	Alternative risk	$R^\star_{\mathrm{A},n}$
Uniform	$n^{1-\epsilon}$	$(1+o(1))/n$	$\mathcal{O}(n^{-2+\epsilon})$	$\asymp n^{-2}$
Gaussian mixture	$\frac{\log(1/\sigma_n^2)}{4\sqrt{1-\sigma_n^2}}$	$(1+o(1))/n$	$\mathcal{O}(\sigma_n^2/n)$	$\geq \sigma_n^2/n$

for every $n' \geq n$,

$$R_{\mathrm{G},n'}(\widehat{\mu}_{\mathrm{trained}}) = \mathcal{O}\left(\frac{1}{n'} + \frac{\log n'}{n^2}\right),$$

$$R_{\mathrm{A},n'}(\widehat{\mu}_{\mathrm{trained}}) = \begin{cases} \mathcal{O}(n^{-1+\epsilon}/n' + n^{-2}), & \textit{Uniform}, \\ \mathcal{O}(\sigma_n^2/n' + \log n'/n^2), & \textit{Gaussian-mixture}. \end{cases}$$

3.1 Pretraining dynamics

We establish the gate bound of Lemma 3.3 and the expert-parameter bounds of Lemma 3.4: first train the gating parameters with fixed experts, then specialize the expert parameters with fixed gates. Equation 3.4 supplies the resulting risk guarantees.

We specialize to two experts and two families. Define the population attention responses and feature direction by

$$m_j(a) = \frac{\mathbb{E}_j\left[e^{a(X-\mu)}(X-\mu)\right]}{\mathbb{E}_j\left[e^{a(X-\mu)}\right]}, \qquad \bar{\boldsymbol{\phi}}_j = \left(m_j(\omega_1), \ldots, m_j(\omega_d)\right)^\top, \qquad \bar{\mathbf{z}}_j = \frac{P_d \bar{\boldsymbol{\phi}}_j}{\|P_d \bar{\boldsymbol{\phi}}_j\|_2}.$$

Let $\bar{w}_{kj}$ denote w_k with $\mathbf{z}(\mathbf{X})$ replaced by $\bar{\mathbf{z}}_j$, and $\bar{w}_{\cdot j} = (\bar{w}_{1j}, \bar{w}_{2j})^\top$. Throughout, bars indicate evaluation at population moments rather than expectation of the corresponding nonlinear statistic. In particular, $\bar{\mathbf{z}}_j \neq \mathbb{E}_j[\mathbf{z}(\mathbf{X})]$ in general. In the proofs, Δ denotes deviation from the corresponding population-moment value. In particular, under family j, $\Delta\mathbf{z} = \mathbf{z} - \bar{\mathbf{z}}_j$.

3.2 Stage 1: gating dynamics

Lemma 3.3 (Gating dynamics; proof in Appendix E for the softmax MoE and Section B.3 for the GLU)**.** *Under the assumptions of Theorem 3.1, run the first stage of Definition 2.3. For sufficiently large n, with probability at least $1-\eta$ over the initialization and training tasks, the first stage ends at a finite time t_θ with*

$$\max\{\|\bar{w}_{\cdot\mathrm{G}}(t_\theta) - e_1\|_1, \|\bar{w}_{\cdot\mathrm{A}}(t_\theta) - e_2\|_1\} = \mathcal{O}(n^{-1}), \qquad \max_q \|\theta_q(t_\theta)\|_2 = \mathcal{O}(\log n).$$

Proof sketch The population features $\bar{\mathbf{z}}_j$ distinguish the families. Let $\boldsymbol{\Sigma}^{(j)}$ be the asymptotic covariance matrix of g given data from family j, that is,

$$\mathbb{E}_j[(g_i - \mu)(g_k - \mu)] = \frac{\Sigma^{(j)}_{ik}}{n} + o(n^{-1}).$$

We show that, for the families we study, there must exist an expert i such that for all $k \neq i$ we have almost surely

$$\Sigma^{(j)}_{ii} < \Sigma^{(j)}_{ik} < \Sigma^{(j)}_{kk}. \tag{6}$$

This implies that the optimal gating must exclusively choose the "best" expert. Finally, we observe that the main component of the loss can be identified as

$$\mathcal{L}^{\mathrm{truncated}}(\theta) = \frac{1}{4}\sum_j \left[\sigma_\mu^2\left(\sum_{k=1}^2 \bar{w}_{kj} - 1\right)^2 + \frac{1}{n}\bar{w}_{\cdot j}^\top \boldsymbol{\Sigma}^{(j)} \bar{w}_{\cdot j}\right]. \tag{7}$$

For softmax, the normalization term vanishes as $\sum_{k=1}^2 \bar{w}_{kj} = 1$.

Softmax MoE First, note that this problem is highly nonconvex due to w being generated by a softmax. Since $\sum_i w_i = 1$ and the routing feature concentrates at $\bar{\mathbf{z}}_j$, the loss expansion gives $\mathcal{L}(\theta) = \mathcal{L}^{\mathrm{sm}}(\theta) + o(1/n)$, where

$$\mathcal{L}^{\mathrm{sm}}(\theta) = \frac{1}{4n} \sum_j \bar{w}_{\cdot j}^\top \mathbf{\Sigma}^{(j)} \bar{w}_{\cdot j}. \tag{8}$$

We can analyze the gradient flow on $\mathcal{L}^{\mathrm{sm}}$ and show global convergence. Correcting for the $o(1/n)$ remainder finishes the proof.

Gated linear unit For the GLU, unbiasedness is not automatic even when the gates are evaluated at $\bar{\mathbf{z}}_j$. Our approach will be to analyze the gradient flow on $\mathcal{L}^{\mathrm{truncated}}$ and correct for the perturbation $\mathcal{L}(\theta) - \mathcal{L}^{\mathrm{truncated}}(\theta)$. The loss $\mathcal{L}^{\mathrm{truncated}}(\theta)$ has two distinct scales, where one scales with σ_μ^2 and one scales with $1/n$. The term $\sigma_\mu^2 \left(\sum_{k=1}^2 \bar{w}_{kj} - 1\right)^2$ corresponds (roughly) to the bias of the predictor. In the truncated loss, there exists a whole manifold of unbiased predictors, and we show that the gradient flow finds a solution near this manifold, while optimizing for the variance of the predictor. Choosing n sufficiently large lets us attack this problem using Tikhonov's theorem and determine where on the manifold the gradient flow converges. Notably, $\mathcal{L}(\theta) - \mathcal{L}^{\mathrm{truncated}}(\theta)$ comes from the variance of w and thus scales with σ_μ^2. Hence, the proof depends on σ_μ^2 being below some threshold. This is natural: a larger prior variance σ_μ^2 increases the cost of normalization error, amplifying inaccuracies in the gating and potentially leading to a collapse of the gating layer towards a single expert for some initializations.

3.3 Stage 2: expert dynamics

For $t \geq t_\theta$, θ is frozen. In this second stage, we analyze the dynamics of λ. We establish that the expert parameter assigned to samples from the Gaussian family moves toward zero, while uniform expert parameters grow toward endpoint estimation, and mixture expert parameters approach a certain limit. In the respective regimes, the leading expert-parameter flows satisfy $\dot{\lambda}_1 \asymp -\lambda_1^3/n$ and $\dot{\lambda}_2 \asymp 1/(n\lambda_2^2)$ assuming without loss of generality $\lambda_2 > 0$. Controlling the residuals yields the following result.

Lemma 3.4 (Expert dynamics; proof in Section F.1)**.** *Under the assumptions of Theorem 3.1, assume the first stage of Definition 2.3 ends at a finite time t_θ satisfying*

$$\max\{\|\bar{w}_{\cdot \mathrm{G}} - e_1\|_1, \|\bar{w}_{\cdot \mathrm{A}} - e_2\|_1\} \leq C/n, \qquad \max_q \|\theta_q\|_2 \leq C \log n,$$

and $|\lambda_1(t_\theta)| \leq 1$, $C^{-1} \leq |\lambda_2(t_\theta)| \leq 1$, for a constant $C < \infty$ independent of n. Use the training procedure in Definition 2.3. For sufficiently large n, with probability at least $1 - \eta$, uniformly over the possible (θ, λ) at t_θ satisfying these bounds, $|\lambda_1| \leq 2$ until $|\lambda_2|$ first reaches the value below, and this first time t_λ satisfies

$$\begin{cases} |\lambda_2(t_\lambda)| = n^{1-\epsilon}, & t_\lambda - t_\theta \leq C n^{4-3\epsilon}, & \textit{Uniform}, \\ |\lambda_2(t_\lambda)| = \dfrac{\log(1/\sigma_n^2)}{4\sqrt{1-\sigma_n^2}}, & t_\lambda - t_\theta \leq C n/\sigma_n^2, & \textit{Gaussian mixture}. \end{cases}$$

Moreover, $|\lambda_1(t_\lambda)| = o(1)$. The tasks used for training λ may also have been used for training θ.

3.4 Efficiency of the final estimator

For every expert i, the error of the final estimator decomposes as

$$\widehat{\mu} - \mu = \underbrace{g_i - \mu}_{\text{expert error}} + \underbrace{\sum_{k \neq i} w_k (g_k - g_i)}_{\text{routing error}} + \underbrace{\left(\sum_k w_k - 1\right) g_i}_{\text{normalization error}}. \tag{9}$$

The normalization error vanishes for the softmax MoE, since $\sum_k w_k = 1$. Lemma 3.5 bounds the routing and normalization errors in $L^2(\mathbb{P}_j)$ through $\|\bar{w}_{\cdot j} - e_i\|_1$.

Lemma 3.5 (Accuracy of the final estimator; proof in Appendix C)**.** *Under the data model of Definition 2.1, fix $j \in \{\mathrm{G}, \mathrm{A}\}$. Condition on the trained (θ, λ) and evaluate on an independent pair $(\mu, \mathbf{X})$, with $\max_k \|\theta_k\|_2 \le C \log n$. Then, for sufficiently large n, every expert $i \in [2]$, and arbitrary λ,*

$$\sqrt{\mathbb{E}_j\left[(\widehat{\mu}(\mathbf{X}) - \mu)^2\right]} \le \sqrt{\mathbb{E}_j\left[(g_i(\mathbf{X}) - \mu)^2\right]} + C\|\bar{w}_{\cdot j} - e_i\|_1 \left(\sigma_\mu^2 + \sqrt{\mathbb{E}_j\left[\max_{\ell \in [n]} |X_\ell - \mu|^4\right]}\right)^{1/2}.$$

Proof sketch of Theorem 3.1. Under the conclusions of Lemma 3.4, the moment estimates in Lemma F.1, Lemma F.2, and Lemma F.3 give expert risks $(1 + o(1))/n$, $\mathcal{O}(n^{-2+\epsilon})$, and $\mathcal{O}(\sigma_n^2/n)$, respectively. Applying Lemma 3.5 and its proof with $i = 1$ for $j = \mathrm{G}$ and $i = 2$ for $j = \mathrm{A}$, the gate bound of Lemma 3.3, and the sample-maximum moment bounds gives

$$\|\widehat{\mu} - g_i\|_{L^2(\mathbb{P}_j)} = \mathcal{O}\left(\frac{\sqrt{\sigma_\mu^2 + \log n}}{n}\right). \tag{10}$$

The triangle inequality then gives the stated risk bounds and preserves the Gaussian leading constant. The fixed-location claims follow from the translation identities and normalization estimate in Section D.3.6.

4 Experiments

Figures 2 and 3 show joint training of the softmax MoE and GLU with the uniform alternative. We use $m = 8000$, $n = 150$, $K = 2$, $d = 10$, and $\sigma_\mu^2 = 0.5$, with 600,000 gradient-descent steps and constant step sizes 20 for θ and 100 for λ. Both learn family-specific selection as $|\lambda_1|$ decreases toward zero and $|\lambda_2|$ grows, with greater variation in the onset of specialization across GLU runs. Gaussian error stays near the sample-mean error, while uniform error falls below it toward the sample mid-range error. Training details, Gaussian-mixture results, and stagewise experiments are in Appendix I.

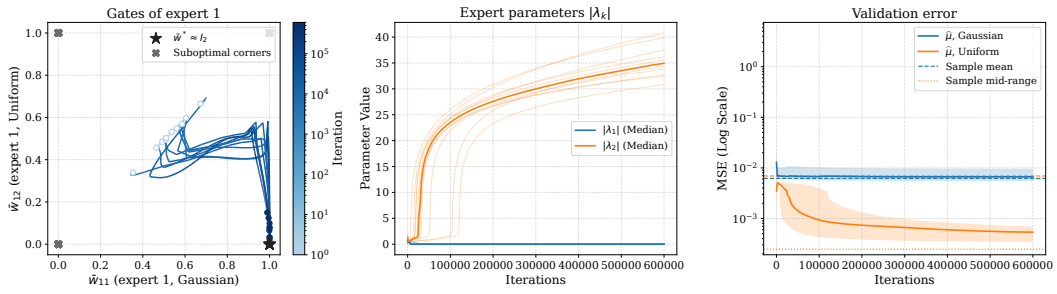

Figure 2: Softmax MoE, joint training with the uniform alternative. Left: $(\bar{w}_{1\mathrm{G}}, \bar{w}_{1\mathrm{A}})$ across runs, colored by iteration. Middle: median $|\lambda_1|$ and $|\lambda_2|$, where thin lines are individual runs. Right: median validation error of $\widehat{\mu}$ on Gaussian and uniform contexts, with the sample mean and sample mid-range as references. Shading spans all ten runs.

5 Related work

In-context learning of statistical procedures. Prior-data fitted networks and tabular foundation models learn prediction procedures by pretraining on synthetic tasks (Müller et al., 2024; Hollmann et al., 2025; Erickson et al., 2025), and related pretrained models estimate treatment effects, survival curves, mutual information, and Poisson means (Balazadeh et al., 2025; Qi et al., 2026; Peyrard & Cho, 2025; Teh et al., 2025). Their focus is prediction or posterior approximation, not how pretraining attains estimation rates that depend on an unknown family.

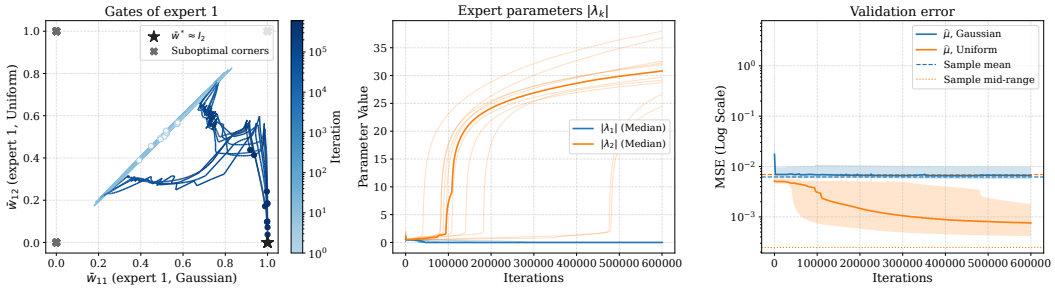

Figure 3: GLU, joint training with the uniform alternative, with panels and conventions as in Figure 2.

Statistical theory of pretrained predictors. Transformers can implement in-context algorithm selection (Bai et al., 2023), attain minimax rates over a known function class (Kim et al., 2024), and approximate Bayesian posterior predictors whose rates adapt to unknown structure (Wakayama & Suzuki, 2025; Ma et al., 2025); idealized pretrained Bayes estimators are similarly adaptive in empirical Bayes problems (Cannella et al., 2026). These guarantees concern the empirical risk minimizer or an idealized predictor; we follow the training trajectory and identify the internal parameter that carries the adaptation.

Training dynamics of attention. Gradient-based analyses of attention for in-context learning (Zhang et al., 2024a; Huang et al., 2023; Kim & Suzuki, 2024; Oko et al., 2024) include task allocation across softmax heads (Chen et al., 2024a), where the task handled by each head is fixed by the task structure. Here the family changes across contexts, so specialization must be paired with learned routing.

Adaptive estimation and architecture. Adaptivity is a classical objective in statistics (Lepskii, 1991; Donoho & Johnstone, 1995), including adaptive estimation of a symmetric location (Stone, 1975; Bickel et al., 1993); the rate gap in our problem stems from the irregularity of the uniform family (Ibragimov & Has'minskii, 1981). Our model is a small Set Transformer (Lee et al., 2019) whose second layer is a softmax mixture of experts (Jacobs et al., 1991) or a GLU (Shazeer, 2020). Appendix K discusses these connections in detail.

6 Discussion

Our current analysis is limited by making some simplifying assumptions on the training procedure as well as the architecture. Without fixing parameters, we observe that the problem becomes much easier to learn in practice, but admits many possible optimal solutions making it much harder to study. We also observe that separating training into two stages slows down training while making the problem much easier to study. Future work should focus on addressing these issues. Additionally, we believe transformers are able to do parameter estimation for much larger families of distributions. Extending our analysis towards these would also be a natural next step. Finally, we observe that there is an upper bound to the prior variance σ_μ^2, which seems to be unavoidable in a simple architecture. Still, finding an architecture that is able to handle arbitrary σ_μ^2 would be a clear next step.

7 Conclusion

We show that stage wise gradient flow on the empirical loss learns an adaptive location estimator without observing family labels. The estimator achieves Gaussian asymptotic efficiency, nearly optimal uniform rates, and order-optimal rates for Gaussian mixtures with shrinking component variance. Attention provides both distribution-sensitive features and candidate estimators, while the gating layer learns which expert to select. Softmax enforces normalization and exact translation equivariance, whereas the GLU learns approximate normalization before selecting experts.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proof, formulate mathematical claims and implement the experiment code. Generating synthetic data sets beyond our own simulation code, translation, dataset cleaning and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to polish the writing, find typesetting and grammar errors, suggest the structure of the main paper, and identify and summarize related work. We have reviewed all AI-assisted work. All proofs were verified by the authors. The experiment code was reviewed and tested by the authors We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. URL https://arxiv.org/abs/2211.15661.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in Neural Information Processing Systems*, 2023.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning, 2025. URL https://arxiv.org/abs/2506.07918.

Rudolf Beran. Asymptotically efficient adaptive rank estimates in location models. *The Annals of Statistics*, 2(1):63–74, 1974. doi: 10.1214/aos/1176342613.

Nils Berglund. Perturbation theory of dynamical systems, 2001. URL https://arxiv.org/abs/math/0111178.

P. J. Bickel. On Adaptive Estimation. *The Annals of Statistics*, 10(3):647 – 671, 1982. doi: 10.1214/aos/1176345863. URL https://doi.org/10.1214/aos/1176345863.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Qingfu Zhang, Hau-San Wong, and Taiji Suzuki. Provable in-context vector arithmetic via retrieving task concepts, 2025. URL https://arxiv.org/abs/2508.09820.

Nick Cannella, Anzo Teh, Yanjun Han, and Yury Polyanskiy. Universal priors: solving empirical Bayes via Bayesian inference and pretraining, 2026. URL https://arxiv.org/abs/2602.15136.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024a.

Xingwu Chen, Lei Zhao, and Difan Zou. How transformers utilize multi-head attention in in-context learning? a case study on sparse linear regression. *arXiv preprint arXiv:2408.04532*, 2024b.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. ISSN 1091-6490. doi: 10.1073/pnas.1912789117. URL http://dx.doi.org/10.1073/pnas.1912789117.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks, 2017. URL https://arxiv.org/abs/1612.08083.

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. doi: 10.1080/01621459.1995.10476626.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. TabArena: A living benchmark for machine learning on tabular data. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.16791.

Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):53–98, 1979. ISSN 0022-0396. doi: 10.1016/0022-0396(79)90152-9. URL https://www.sciencedirect.com/science/article/pii/0022039679901529.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023. URL https://arxiv.org/abs/2208.01066.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. *International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers, 2023. URL https://arxiv.org/abs/2310.05249.

I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer New York, New York, NY, 1981. ISBN 978-1-4899-0027-2. doi: 10.1007/978-1-4899-0027-2.

Nikolaos Ignatiadis and Sid Kankanala. Compound decisions and empirical Bayes via Bayesian nonparametrics, 2026. URL https://arxiv.org/abs/2602.20115.

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. doi: 10.1162/neco.1991.3.1.79.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape, 2024. URL https://arxiv.org/abs/2402.01258.

Juno Kim, Tai Nakamaki, and Taiji Suzuki. Transformers are minimax optimal nonparametric in-context learners, 2024. URL https://arxiv.org/abs/2408.12186.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer New York, NY, 2 edition, 1998. ISBN 978-0-387-98502-2. doi: 10.1007/b98854.

Oleg V. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, 35(3):454–466, 1991. doi: 10.1137/1135065.

Zihao Li, Yuan Cao, Cheng Gao, Yihan He, Han Liu, Jason M. Klusowski, Jianqing Fan, and Mengdi Wang. One-layer transformer provably learns one-nearest neighbor in context. *arXiv preprint arXiv:2411.10830*, 2024.

Tianyi Ma, Tengyao Wang, and Richard J. Samworth. Optimal in-context adaptivity and distributional robustness of transformers, 2025. URL https://arxiv.org/abs/2510.23254.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference, 2024. URL https://arxiv.org/abs/2112.10510.

Thomas Nagler. Statistical foundations of prior-data fitted networks, 2023. URL https://arxiv.org/abs/2305.11097.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2411.02544.

Maxime Peyrard and Kyunghyun Cho. Meta-statistical learning: Supervised learning of statistical estimators. *arXiv preprint arXiv:2502.12088*, 2025.

Shi-ang Qi, Vahid Balazadeh, Michael Cooper, Russell Greiner, and Rahul G Krishnan. SurvivalPFN: Amortizing survival prediction via in-context Bayesian inference. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=PDik7bpFhE.

J. Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer, 2003. ISBN 9780387953823.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

C. J. Stone. Adaptive maximum likelihood estimators of a location parameter. *Annals of Statistics*, 3 (2):267–284, 1975. doi: 10.1214/aos/1176343056.

Anzo Teh, Mark Jabbour, and Yury Polyanskiy. Solving empirical Bayes via transformers, 2025. URL https://arxiv.org/abs/2502.09844.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009. doi: 10.1007/b13794.

A. W. van der Vaart and J. A. Wellner. Stein 1956: Efficient nonparametric testing and estimation. *The Annals of Statistics*, 49(4), 2021. doi: 10.1214/21-AOS2056.

Tomoya Wakayama and Taiji Suzuki. In-context learning is provably Bayesian inference: A generalization theory for meta-learning, 2025. URL https://arxiv.org/abs/2510.10981.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets, 2025. URL https://arxiv.org/abs/2407.00956.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Qiong Zhang, Yan Shuo Tan, Qinglong Tian, and Pengfei Li. TabPFN: One model to rule them all? *arXiv preprint arXiv:2505.20003*, 2025. URL https://arxiv.org/abs/2505.20003.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024a.

Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. In-context learning of a linear transformer block: Benefits of the MLP component and one-step GD initialization. *Advances in Neural Information Processing Systems*, 2024b.

A Outline

Proof map. Lemma B.5 gives the reparameterization used in the dynamics, and Lemma F.5 establishes the geometry of $\mathbf{z}$. Appendix B assembles the main theorem from the fresh-context risk bound in Appendix C, the gating analysis in Appendix D and Appendix E, and the expert-parameter analysis in Appendix F. Appendix G collects auxiliary estimates, and Appendix H gives the Gaussian-mixture benchmark. Appendix I reports additional experiments, and Appendix J discusses the role of LayerNorm. Appendix K gives an extended discussion of related work. See Figure 1 for notation.

B Proof of Theorem 3.1

B.1 Expert hierarchy

Definition B.1 (Expert covariance and the better expert)**.** Let $\mathbf{\Sigma}^{(j)}$ be the matrix with entries $\Sigma_{ik}^{(j)} = \Sigma_g(\lambda_i, \lambda_k)$ evaluated under the centered law Q_j. The symmetrized covariance formula is given in Definition F.9, with family-specific calculations in the expert-covariance appendix. Expert $i \in \{1, 2\}$ is *better* for family $j \in \{\mathrm{G}, \mathrm{A}\}$ if, for the other expert $k \neq i$,

$$\Sigma_{ii}^{(j)} < \Sigma_{ik}^{(j)} < \Sigma_{kk}^{(j)}. \tag{11}$$

The experts satisfy the *expert hierarchy* if each family has a better expert. Its strict covariance gaps are the two positive differences in the ordering.

Lemma B.2 (Gaussian distribution satisfies the expert hierarchy)**.** *Let $Y \sim \mathcal{N}(0, 1)$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ with $0 < |\lambda_1| < |\lambda_2|$. Then*

$$\Sigma_g(\lambda_1, \lambda_1) < \Sigma_g(\lambda_1, \lambda_2) < \Sigma_g(\lambda_2, \lambda_2). \tag{12}$$

Proof. The statement follows from $\Sigma_g(\lambda_1, \lambda_2) = \cosh(\lambda_1\lambda_2) + \lambda_1\lambda_2 \sinh(\lambda_1\lambda_2)$, which is strictly increasing in $|\lambda_1\lambda_2|$. □

Lemma B.3 (Gaussian mixture distribution satisfies the expert hierarchy)**.** *For every fixed $0 < \delta < 1$ there exists $\sigma_\delta > 0$ such that for all $0 < \sigma_n < \sigma_\delta < 1/\sqrt{2}$ and $0 \leq |\lambda_1| < |\lambda_2| \leq L_{\max} = (1 - \delta)\log(1/\sigma_n)$,*

$$\Sigma_g(\lambda_1, \lambda_1) > \Sigma_g(\lambda_1, \lambda_2) > \Sigma_g(\lambda_2, \lambda_2). \tag{13}$$

Proof. See Lemma F.16. □

Lemma B.4 (Uniform distribution satisfies the expert hierarchy)**.** *Let $Y \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$ and $\lambda_1, \lambda_2 \in [-1, 1]$ with $|\lambda_1| < |\lambda_2|$. Then*

$$\Sigma_g(\lambda_1, \lambda_1) > \Sigma_g(\lambda_1, \lambda_2) > \Sigma_g(\lambda_2, \lambda_2). \tag{14}$$

Proof. See Corollary F.13. □

In the parameter ranges above, $0 < |\lambda_1| < |\lambda_2|$ makes expert 1 better on Gaussian contexts and expert 2 better on alternative contexts.

B.2 Gating-layer dynamics

Lemma B.5 (Decomposition of the gating parameters)**.** *For the GLU, let $Z_{jk} = \langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k \rangle$ and suppose $Z \succ 0$. Each gating parameter has the decomposition*

$$\theta_q = \theta_q^{\parallel} + \theta_q^{\perp}, \qquad \theta_q^{\parallel} = \sum_{\ell,j} (Z^{-1})_{\ell j} \sigma^{-1}(\bar{w}_{qj}) \bar{\mathbf{z}}_\ell, \qquad \theta_q^{\perp} \perp \mathrm{span}\{\bar{\mathbf{z}}_j\}_{j \in \{\mathrm{G},\mathrm{A}\}}.$$

Thus θ is equivalently parameterized by $(\bar{w}, \theta^{\perp})$. Moreover,

$$\|\theta_q\|_2 \leq \|\theta_q^{\perp}\|_2 + \sqrt{\frac{2}{\lambda_{\min}(Z)}} \max_j \log \frac{1}{\bar{w}'_{qj}}.$$

Proof. Taking the inner product of $\theta_q^{\parallel}$ with $\bar{\mathbf{z}}_j$ gives $\sigma^{-1}(\bar{w}_{qj}) = \langle \theta_q, \bar{\mathbf{z}}_j \rangle$, so the stated decomposition is orthogonal. Moreover,

$$\|\theta_q^{\parallel}\|_2^2 = \sum_{i,j} \sigma^{-1}(\bar{w}_{qi})(Z^{-1})_{ij} \sigma^{-1}(\bar{w}_{qj}) \leq \frac{1}{\lambda_{\min}(Z)} \sum_j |\sigma^{-1}(\bar{w}_{qj})|^2.$$

The conclusion follows from the triangle inequality and $|\log(p/(1-p))| \leq \log(1/(p(1-p)))$ for $0 < p < 1$. □

For the truncated flow, $\dot{\theta}_q$ lies in $\operatorname{span}\{\bar{\mathbf{z}}_j\}_{j\in\{\mathrm{G,A}\}}$, so $\theta_q^{\perp}(t) = \theta_q^{\perp}(0)$. Under the assumptions of Lemma D.10, the full flow also has $\|\theta_q^{\perp}(t)\|_2 = \mathcal{O}(1)$ for $t_{\text{crit}} \le t \le t_\theta$.

Definition B.6 (Truncated loss)**.** For fixed λ, define the bias loss

$$\mathcal{L}^{\text{bias}}(\theta) = \frac{\sigma_\mu^2}{4}\sum_j (\bar{w}_{1j} + \bar{w}_{2j} - 1)^2$$

and recall the truncated loss from (7)

$$\mathcal{L}^{\text{truncated}}(\theta) = \frac{1}{4}\sum_j \left[\sigma_\mu^2\left(\sum_{k=1}^{2}\bar{w}_{kj} - 1\right)^2 + \frac{1}{n}\bar{w}_{\cdot j}^{\top}\mathbf{\Sigma}^{(j)}\bar{w}_{\cdot j}\right].$$

Lemma B.7 (Stationary points of the truncated flow)**.** *Fix λ and $\sigma_\mu^2 > 0$, suppose the expert hierarchy of Definition B.1 holds, and let k_j denote the better expert in family j. Let $e_1 = (1,0)^\top$, $e_2 = (0,1)^\top$, and suppose $Z \succ 0$. The truncated flow is*

$$\dot{\bar{w}}_{qk} = -\frac{1}{2}\bar{w}'_{qk}\sum_j Z_{jk}\bar{w}'_{qj}S_{qj}(\bar{w}),$$

$$S_{qj}(\bar{w}) = \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right) + \frac{1}{n}\sum_{i=1}^{2}\Sigma_{qi}^{(j)}\bar{w}_{ij}.$$

Its stationary points form the set

$$\mathcal{S} = \left\{\bar{w}^\circ \in [0,1]^{2\times 2} : \bar{w}_{qj}^\circ(1-\bar{w}_{qj}^\circ)S_{qj}(\bar{w}^\circ) = 0 \text{ for all } q,j\right\},$$

with

$$\bar{w}^\circ \in \mathcal{S} \iff \bar{w}_{\cdot j}^\circ \in \Big\{0, e_1, e_2, e_1+e_2, \frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \Sigma_{11}^{(j)}}e_1, \frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \Sigma_{22}^{(j)}}e_2\Big\} \quad (j \in \{\mathrm{G,A}\}).$$

The unique optimum is

$$\bar{w}_{\cdot j}^{*} = \frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \Sigma_{k_j k_j}^{(j)}}e_{k_j},$$

$$\operatorname*{argmin}_{\bar{w}\in[0,1]^{2\times 2}} \mathcal{L}^{\text{truncated}}(\bar{w}) = \{\bar{w}^*\}.$$

Proof. See Section D.3.1. □

Lemma B.8 (Convergence to the critical manifold)**.** *Consider the unperturbed flow*

$$\dot{\theta} = -\nabla_\theta \mathcal{L}^{\text{bias}}.$$

Assume $Z \succ 0$, fixed $\sigma_\mu^2 > 0$, and initialization $\theta(0)$ such that $\max_{q,j}|\bar{w}_{qj}(0) - 1/2| \le \delta$ for sufficiently small δ. Then all gates remain in $[1/4, 3/4]$, and

$$\max_j |\bar{w}_{1j}(t) + \bar{w}_{2j}(t) - 1| \le C\delta e^{-\alpha t}, \qquad \alpha = \frac{9\sigma_\mu^2\lambda_{\min}(Z)}{256}.$$

For the truncated flow at fixed λ, there is also a time $t_{crit} = \mathcal{O}(\log n)$ such that all gates remain in $[1/4, 3/4]$ up to t_{crit} and

$$\max_j |\bar{w}_{1j}(t_{crit}) + \bar{w}_{2j}(t_{crit}) - 1| = \mathcal{O}(n^{-1}).$$

The same conclusion at t_{crit} holds for the full population flow at fixed λ if $\max_q \|\theta_q(0)\|_2$ is bounded independently of n. Under this initialization bound and the data conditions of Lemma D.4, fix $\epsilon > 0$ and sample $m = \widetilde{\Omega}(n^{1+\epsilon})$ independent tasks, with $m \in \operatorname{poly}(n)$. With probability at least $1 - 1/n$, the empirical flow $\dot{\theta} = -\nabla_\theta \hat{\mathcal{L}}_m$ has the same conclusion at $t_{crit} = \mathcal{O}(\log n)$, with all gates in $[1/4, 3/4]$ and $\sup_{0\le t\le t_{crit}}\|\theta(t)\|_2 = \mathcal{O}(1)$. Constants are independent of n, but may depend on σ_μ^2, Z, and the fixed $\lambda, \boldsymbol{\omega}$.

Proof. See Section D.3.2. □

Lemma B.9 (Convergence below loss barrier)**.** *Fix λ independently of n and suppose the strict expert hierarchy holds. Assume $Z \succ 0$ and $Z_{\mathrm{GA}} < \min\{Z_{\mathrm{GG}}, Z_{\mathrm{AA}}\}$. Suppose*

$$\bar{w}(t_{crit}) \in [1/4, 3/4]^{2\times 2}, \qquad n\sigma_\mu^2 \max_j |\bar{w}_{1j}(t_{crit}) + \bar{w}_{2j}(t_{crit}) - 1| \le C_0,$$

with C_0 independent of n. For the truncated flow and fixed $\sigma_\mu^2 > 0$, there exist $S, \kappa > 0$, independent of n, such that, for sufficiently large n,

$$\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit} + nS)) < \mathcal{L}^{\mathrm{truncated}}(\bar{w}^\circ) - \frac{\kappa}{n}$$

for every

$$\bar{w}^\circ \in \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}. \tag{15}$$

For the full population flow, assume also that $\max_q \|\theta_q^\perp(t_{crit})\|_2$ is bounded independently of n and σ_μ^2. For sufficiently small fixed $\sigma_\mu^2 > 0$ and then sufficiently large n, both $\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit} + nS))$ and $\mathcal{L}(\theta(t_{crit}+nS))$ satisfy the same upper bound. Under these assumptions and the data and sample-size conditions of the empirical part of Lemma B.8, the same conclusions hold for the empirical flow with probability at least $1-1/n$, uniformly over possibly data-dependent initial parameters satisfying the stated bounds. For the full population and empirical flows, $\sup_{t_{crit}\le t\le t_{crit}+nS} \|\theta(t)\|_2 = \mathcal{O}(1)$, with a bound independent of n and sufficiently small σ_μ^2.

Proof. See Section D.3.3. □

Lemma B.10 (Flow at fixed λ)**.** *Fix λ independently of n and suppose the strict inequalities in Definition B.1 hold. Assume $Z \succ 0$, $Z_{\mathrm{GA}} \ge 0$, and fixed $\sigma_\mu^2 > 0$. Use Lemma B.7 with $k_{\mathrm{G}} = 1$ and $k_{\mathrm{A}} = 2$. Suppose $\bar{w}(t_{crit}) \in (0,1)^{2\times 2}$, $\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit})) = \mathcal{O}(1/n)$, and, for some fixed $\kappa > 0$,*

$$\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit})) \le \mathcal{L}^{\mathrm{truncated}}(\bar{w}^\circ) - \frac{\kappa}{n}$$

for every $\bar{w}^\circ$ in (15)*. For sufficiently large n, the truncated flow reaches a finite time $t_\theta \ge t_{crit}$ with*

$$\|\bar{w}_{\cdot j}(t_\theta) - e_{k_j}\|_2 \le \frac{1}{n} + \frac{\Sigma^{(j)}_{k_j k_j}}{n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j}} = \mathcal{O}(n^{-1}).$$

For the full population flow, assume that $\|\theta(t_{crit})\|_2$ and $n\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit}))$ are bounded independently of n and σ_μ^2. For sufficiently small fixed $\sigma_\mu^2 > 0$ and then sufficiently large n, let C_0 be the constant from Lemma D.10 and set

$$\rho = \frac{C_0}{n\sigma_\mu^2}.$$

There is a time $t_\theta \le t_{crit} + 2C_0^2 n^2 \sigma_\mu^2$ with

$$\|\bar{w}_{\cdot j}(t_\theta) - \bar{w}^*_{\cdot j}\|_2 \le \rho + \frac{\Sigma^{(j)}_{k_j k_j}}{n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j}}, \qquad \mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_\theta)) \le \mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit})).$$

In particular, the gate error is $\mathcal{O}(n^{-1})$, and $\sup_{t_{crit}\le t\le t_\theta} \|\theta(t)\|_2 = \mathcal{O}(\log n)$.

Proof. See Section D.3.4. □

Lemma B.11 (Empirical flow at fixed λ)**.** *Assume the data conditions of Lemma D.4 and the full-flow assumptions of Lemma B.10, with $\|\theta(t_{crit})\|_2 \le B$ and $n\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit})) \le c$. Let C_0 and ρ be as there. Fix $\epsilon > 0$ and sample m independent tasks according to Definition 2.1, where $m \in \mathrm{poly}(n)$ and*

$$m \ge n^{1+\epsilon}.$$

Consider $\dot{\theta} = -\nabla_\theta \hat{\mathcal{L}}_m$ from possibly data-dependent parameters $\theta(t_{crit})$ satisfying these assumptions. For sufficiently large n, with probability at least $1 - 1/n$, there is a time $t_\theta \le t_{crit} + 2C_0^2 n^2 \sigma_\mu^2$ with

$$\|\bar{w}_{\cdot j}(t_\theta) - \bar{w}^*_{\cdot j}\|_2 \le \rho + \frac{\Sigma^{(j)}_{k_j k_j}}{n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j}}, \qquad \mathcal{L}^{\text{truncated}}(\bar{w}(t_\theta)) \le \mathcal{L}^{\text{truncated}}(\bar{w}(t_{crit})),$$

$$\sup_{t_{crit} \le t \le t_\theta} \|\theta(t)\|_2 \le B + 3C_0 \log(1 + cn\sigma_\mu^2) = \mathcal{O}(\log n), \qquad \sup_{t_{crit} \le t \le t_\theta} \|\theta_q^\perp(t)\|_2 = \mathcal{O}(1),$$

and

$$\mathcal{L}(\theta(t_\theta)) = \mathcal{O}(n^{-1}), \qquad \hat{\mathcal{L}}_m(\theta(t_\theta)) = \mathcal{O}(n^{-1}).$$

Proof. See Section D.3.5. □

B.3 Proof of the GLU part of Lemma 3.3

Proof of the GLU part of Lemma 3.3. Since λ is fixed in the first stage, $\lambda(t_\theta) = \lambda(0)$, and $|\lambda_k(0)| \le 1$ lies in the ranges of Lemma B.2, Lemma B.3 and Lemma B.4 for sufficiently large n. By those lemmas, $0 < |\lambda_1(0)| < |\lambda_2(0)|$ makes expert 1 better in family G and expert 2 better in family A. Under this hierarchy and $\bar{w}_{1j} + \bar{w}_{2j} = 1$, selecting the better expert minimizes the variance in family j. Lemma F.5 gives $Z \succ 0$, $Z_{\mathrm{GA}} \ge 0$ and $Z_{\mathrm{GA}} < \min\{Z_{\mathrm{GG}}, Z_{\mathrm{AA}}\}$ for $d \ge 3$, almost surely over $\boldsymbol{\omega}$.

Apply Lemma B.8 to obtain the gate and parameter bounds at t_{crit}. Then Lemma B.9 gives, at $t_{\text{crit}} + nS$, the loss bound below every suboptimal corner and $\|\theta(t_{\text{crit}} + nS)\|_2 = \mathcal{O}(1)$. Since the loss at each such corner is $\mathcal{O}(n^{-1})$, these bounds supply the entry assumptions of Lemma B.11, with its initial time shifted to $t_{\text{crit}} + nS$. Since $m = \widetilde{\Omega}(n^{1+\epsilon}) \ge n^{1+\epsilon/2}$ for sufficiently large n, apply that lemma with $\epsilon/2$ in place of ϵ to obtain a finite time t_θ with $\|\bar{w}_{\cdot j}(t_\theta) - \bar{w}^*_{\cdot j}\|_2 = \mathcal{O}(n^{-1})$ and $\max_q \|\theta_q(t_\theta)\|_2 = \mathcal{O}(\log n)$. Lemma B.7 gives $\|\bar{w}^*_{\cdot j} - e_{k_j}\|_1 = \mathcal{O}(n^{-1})$, so

$$\max\{\|\bar{w}_{\cdot \mathrm{G}}(t_\theta) - e_1\|_1, \|\bar{w}_{\cdot \mathrm{A}}(t_\theta) - e_2\|_1\} = \mathcal{O}(n^{-1}).$$

The probability that any of these first-stage conclusions fails is $\mathcal{O}(n^{-1})$. □

B.4 Attention-expert dynamics

Corollary B.12 (Accuracy with fixed gating parameters)**.** *Under the assumptions of Lemma 3.4, stop training λ at t_λ. For sufficiently large n, with probability at least $1 - \eta$, the mean-squared errors on independent test samples satisfy*

$$\mathbb{E}_{\mathrm{G}}[(\widehat{\mu} - \mu)^2] = \frac{1 + o(1)}{n}, \qquad \mathbb{E}_{\mathrm{A}}[(\widehat{\mu} - \mu)^2] = \begin{cases} \mathcal{O}(n^{-2+\epsilon}), & \textit{Uniform}, \\ \mathcal{O}(\sigma_n^2/n), & \textit{Gaussian mixture}. \end{cases}$$

Proof. See Section D.3.6. □

B.5 Theorem assembly

Proof of Theorem 3.1. Apply Lemma 3.3 for the respective architecture and freeze the gating parameters at t_θ. For the softmax MoE, represent the frozen weights by the sigmoid parameters $(\theta_1 - \theta_2, \theta_2 - \theta_1)$, which retain the $\mathcal{O}(\log n)$ norm bound and give the same loss for every λ. Since $\lambda(t_\theta) = \lambda(0)$, the expert parameters retain their initialization bounds, and $m = \widetilde{\Omega}(n^{1+\epsilon})$ satisfies the sample-size requirement of Lemma 3.4. Its uniform conclusion applies to the data-dependent parameters at t_θ using the same training tasks and gives a finite target time t_λ. Finally, Corollary B.12 gives the stated test risks at t_λ, with joint probability at least $1 - \eta$ by the failure bounds below. The bounds at fixed μ follow from the translation identities in Section D.3.6.

Under Definition 2.2, after relabeling, $0 < |\lambda_1(0)| < |\lambda_2(0)|$ almost surely and $\mathbb{P}(|\lambda_2(0)| < C^{-1}) = C^{-2}$. For sufficiently large $C > 1$, independent of n, both $C^{-2} \le \eta/4$ and $\mathbb{P}(\max_q \|\theta_q(0)\|_2 > C) \le \eta/4$. For the GLU, choose σ_θ^2 sufficiently small that $\mathbb{P}(\max_q \|\theta_q(0)\|_2 > \min\{C, 4\delta\}) \le$

$\eta/4$, where $\delta > 0$ is the sufficiently small constant in Lemma B.8. On the complementary event, $\max_{q,j} |\bar{w}_{qj}(0) - 1/2| \le \delta$, since $\|\bar{\mathbf{z}}_j\|_2 = 1$ and $|\sigma(u) - 1/2| \le |u|/4$. Thus the required initialization bounds hold jointly with probability at least $1 - \eta/2$. For sufficiently large n, the remaining failure probabilities are at most $\eta/4$ per training stage, giving probability at least $1 - \eta$ in both Lemma 3.3 and Theorem 3.1. □

C Accuracy of the final estimator

Proof of Lemma 3.5. It suffices to consider sigmoid gates, since for $K = 2$ the softmax weights equal the sigmoid weights with parameters $(\theta_1 - \theta_2, \theta_2 - \theta_1)$, whose norms are still $\mathcal{O}(\log n)$. Since $h(\mathbf{X} + \mu\mathbf{1}, a) = h(\mathbf{X}, a) + \mu$, centering the feature coordinates removes the location exactly:

$$\mathbf{z}(\mathbf{X} + \mu\mathbf{1}) = \mathbf{z}(\mathbf{X}).$$

Thus $\mathbf{z}$ and the gates depend only on the centered observations and are independent of μ. Each centered feature satisfies

$$h(\mathbf{X}, \omega_r) - \mu = \frac{n^{-1}\sum_{\ell=1}^n e^{\omega_r(X_\ell - \mu)}(X_\ell - \mu)}{n^{-1}\sum_{\ell=1}^n e^{\omega_r(X_\ell - \mu)}}.$$

The summands have finite moments of every order, uniformly in n also for the Gaussian mixture, since its laws are uniformly 1-sub-Gaussian and $\boldsymbol{\omega}$ is fixed. The population denominators are positive, and by Lemma F.5 the centered population feature norms are bounded below, so the normalization map is locally Lipschitz there. Moment bounds for the empirical averages therefore give

$$\mathbb{E}_j\left[\|\Delta\mathbf{z}\|_2^{2p}\right] = \mathcal{O}(n^{-p})$$

for every fixed integer $p \ge 1$: on a fixed neighborhood of the population averages this follows from local Lipschitz continuity, and outside that neighborhood it follows from $\|\Delta\mathbf{z}\|_2 \le 2$ and Markov's inequality applied to the averages. Using $\max_k \|\theta_k\|_2 \le C \log n$, another application of Markov's inequality gives

$$\mathbb{P}\left[\max_k |\langle \theta_k, \Delta\mathbf{z}\rangle| > 1 \mid \tau = j\right] = \mathcal{O}\left(n^{-p}(\log n)^{2p}\right).$$

Since $\max_k |\langle \theta_k, \Delta\mathbf{z}\rangle| \le 2C \log n$, splitting at the event above and choosing $p > 8C$ gives

$$\mathbb{E}_j\left[\exp\left(4\max_k |\langle \theta_k, \Delta\mathbf{z}\rangle|\right)\right] \le e^4 + \mathcal{O}\left(n^{8C-p}(\log n)^{2p}\right) \le C.$$

The sigmoid satisfies

$$\sigma(u+v) \le e^{|v|}\sigma(u), \qquad 1 - \sigma(u+v) \le e^{|v|}(1 - \sigma(u)).$$

Applying these inequalities to each gate yields

$$|1 - w_i(\mathbf{X})| + \sum_{k \ne i} w_k(\mathbf{X}) \le \exp\left(\max_k |\langle \theta_k, \Delta\mathbf{z}\rangle|\right)\|\bar{w}_{\cdot j} - e_i\|_1.$$

The difference from the selected expert is

$$\widehat{\mu} - g_i = \mu\left(\sum_k w_k - 1\right) + \sum_k (w_k - \mathbf{1}_{\{k=i\}})(g_k - \mu).$$

The gates and centered experts depend only on the centered observations and are independent of μ. Since $\mathbb{E}[\mu] = 0$, averaging over μ eliminates the cross term and gives

$$\begin{aligned}
\mathbb{E}_j\left[(\widehat{\mu} - g_i)^2\right] &= \sigma_\mu^2 \mathbb{E}_j\left[\left(\sum_k w_k - 1\right)^2\right] + \mathbb{E}_j\left[\left(\sum_k (w_k - \mathbf{1}_{\{k=i\}})(g_k - \mu)\right)^2\right] \\
&\le \|\bar{w}_{\cdot j} - e_i\|_1^2 \mathbb{E}_j\left[\exp\left(2\max_k |\langle \theta_k, \Delta\mathbf{z}\rangle|\right)\left(\sigma_\mu^2 + \max_{\ell \in [n]} |X_\ell - \mu|^2\right)\right] \\
&\le C\|\bar{w}_{\cdot j} - e_i\|_1^2 \left(\sigma_\mu^2 + \sqrt{\mathbb{E}_j\left[\max_{\ell \in [n]} |X_\ell - \mu|^4\right]}\right).
\end{aligned}$$

The first inequality uses $|g_k - \mu| \le \max_\ell |X_\ell - \mu|$, which holds for every λ_k, and the second uses Cauchy–Schwarz and the exponential moment bound above. The triangle inequality in L^2 proves the claim. □

The bound on $\mathbb{E}_j\left[(\widehat{\mu} - g_i)^2\right]$ is quadratic in $\|\bar{w}_{\cdot j} - e_i\|_1$ and uniform over λ. It therefore applies throughout training of λ after θ is fixed.

D GATING-LAYER DYNAMICS

D.1 DERIVATION OF THE POPULATION GRADIENT FLOW

Remark D.1 (Layer normalization). Since the centered population feature norms are bounded below by Lemma F.5, layer normalization has bounded derivatives on a fixed neighborhood of each such vector. Composing the moment ratios with this map therefore preserves the orders in n of the Taylor expansions and their remainder bounds in the following gradient and loss calculations, although their coefficients change. Outside this neighborhood, the moment bounds for the empirical averages and $\|\mathbf{z}\|_2 = 1$ control the contribution.

Lemma D.2 (Population gradient expansion)**.** *Let θ_q be the gating parameter vector for expert $q \in \{1, 2\}$, and let the loss function $\mathcal{L}(\theta)$ be defined as:*

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}\left[\left(\sum_{i=1}^{2} w_i(\mathbf{X})g_i - \mu\right)^2\right]$$

Then

$$\begin{aligned}
\nabla_{\theta_q}\mathcal{L}(\theta) = \frac{1}{2}\sum_j \Bigg\{ & \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\bar{w}'_{qj}\bar{\mathbf{z}}_j + \frac{\bar{w}'_{qj}}{n}\sum_{i=1}^{2}\bar{w}_{ij}\Sigma_{qi}^{(j)}\bar{\mathbf{z}}_j \\
& + \sigma_\mu^2\bar{w}'_{qj}\bar{\mathbf{z}}_j\sum_{i=1}^{2}\bar{w}'_{ij}\theta_i^\top\mathbb{E}_j[\Delta\mathbf{z}] \\
& + \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\left[\bar{w}''_{qj}\bar{\mathbf{z}}_j\theta_q^\top\mathbb{E}_j[\Delta\mathbf{z}] + \bar{w}'_{qj}\mathbb{E}_j[\Delta\mathbf{z}]\right] \\
& + \frac{\sigma_\mu^2}{n}\sum_{i=1}^{2}\bar{w}'_{ij}\left[\bar{w}''_{qj}\left(\theta_q^\top\mathbf{\Sigma}_Z^{(j)}\theta_i\right)\bar{\mathbf{z}}_j + \bar{w}'_{qj}\mathbf{\Sigma}_Z^{(j)}\theta_i\right] + \frac{\sigma_\mu^2}{n}\bar{w}''_{qj}\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\mathbf{\Sigma}_Z^{(j)}\theta_q \\
& + \frac{\sigma_\mu^2}{2n}\bar{w}'_{qj}\left[\sum_{i=1}^{2}\bar{w}''_{ij}\left(\theta_i^\top\mathbf{\Sigma}_Z^{(j)}\theta_i\right)\right]\bar{\mathbf{z}}_j \\
& + \frac{\sigma_\mu^2}{2n}\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\bar{w}'''_{qj}\left(\theta_q^\top\mathbf{\Sigma}_Z^{(j)}\theta_q\right)\bar{\mathbf{z}}_j + \bar{w}'_{qj}\mathcal{O}(n^{-3/2})\Bigg\}
\end{aligned}$$

where $\bar{w}_{ij} = \sigma(\langle\theta_i, \bar{\mathbf{z}}_j\rangle)$, $\bar{\mathbf{z}}_j$ is obtained by applying layer normalization to the population moment ratios, $\Delta w_i = w_i(\mathbf{z}) - \bar{w}_{ij}$, and $\Delta w'_q = w'_q(\mathbf{z}) - \bar{w}'_{qj}$. Primes on gates denote derivatives of the sigmoid with respect to its scalar logit, so $w'_q = w_q(1 - w_q)$, with the same convention for population gates and higher derivatives. The matrix $\mathbf{\Sigma}^{(j)}$ is defined in Definition B.1, and $\mathbf{\Sigma}_Z^{(j)}$ is the asymptotic covariance matrix of $\sqrt{n}(\mathbf{z} - \bar{\mathbf{z}}_j)$ under family j. The decomposition of θ_i is given in Lemma B.5.

Proof.

$$
\begin{aligned}
\nabla_{\theta_q}\mathcal{L}(\theta) &= \mathbb{E}\left[\left(\sum_{i=1}^{2} w_i(\mathbf{X})g_i - \mu\right) g_q w_q'(\mathbf{X})\mathbf{z}\right] \\
&= \frac{1}{2}\sum_j \mathbb{E}_j\left[\left(\sum_{i=1}^{2} w_i(\mathbf{X})\left(\mu^2 + (g_i-\mu)(g_q-\mu)\right) - \mu^2\right) w_q'(\mathbf{z})\mathbf{z}\right] \\
&= \frac{1}{2}\sum_j \mathbb{E}_j\Bigg[\left\{\sum_{i=1}^{2}(\bar{w}_{ij}+\Delta w_i)\big(\mu^2 + (g_i-\mu)(g_q-\mu)\big) - \mu^2\right\} \\
&\qquad\qquad \times (\bar{w}_{qj}' + \Delta w_q')(\bar{\mathbf{z}}_j + \Delta\mathbf{z})\Bigg] \\
&= \frac{1}{2}\sum_j \Bigg(\sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\bar{w}_{qj}'\bar{\mathbf{z}}_j + \left(\sum_{i=1}^{2}\bar{w}_{ij}\mathbb{E}_j\left[(g_i-\mu)(g_q-\mu)\right]\right)\bar{w}_{qj}'\bar{\mathbf{z}}_j \\
&\quad + \sigma_\mu^2\bar{w}_{qj}'\bar{\mathbf{z}}_j\sum_{i=1}^{2}\mathbb{E}_j[\Delta w_i] + \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\left[\bar{\mathbf{z}}_j\mathbb{E}_j[\Delta w_q'] + \bar{w}_{qj}'\mathbb{E}_j[\Delta\mathbf{z}]\right] \\
&\quad + \sigma_\mu^2\sum_{i=1}^{2}\mathbb{E}_j\left[\Delta w_i\Delta w_q'\right]\bar{\mathbf{z}}_j + \sigma_\mu^2\sum_{i=1}^{2}\mathbb{E}_j\left[\Delta w_i\Delta\mathbf{z}\right]\bar{w}_{qj}' \\
&\quad + \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\mathbb{E}_j\left[\Delta w_q'\Delta\mathbf{z}\right] \\
&\quad + \left(\sum_{i=1}^{2}\mathbb{E}_j\left[\Delta w_i(g_i-\mu)(g_q-\mu)\right]\right)\bar{w}_{qj}'\bar{\mathbf{z}}_j \\
&\quad + \sum_{i=1}^{2}\mathbb{E}_j\left[(\bar{w}_{ij}+\Delta w_i)(g_i-\mu)(g_q-\mu)\Delta\mathbf{z}\right]\bar{w}_{qj}' \\
&\quad + \sum_{i=1}^{2}\mathbb{E}_j\left[(\bar{w}_{ij}+\Delta w_i)(g_i-\mu)(g_q-\mu)\Delta w_q'\right]\bar{\mathbf{z}}_j + \sigma_\mu^2\sum_{i=1}^{2}\mathbb{E}_j\left[\Delta w_i\Delta w_q'\Delta\mathbf{z}\right] \\
&\quad + \sum_{i=1}^{2}\mathbb{E}_j\left[(\bar{w}_{ij}+\Delta w_i)(g_i-\mu)(g_q-\mu)\Delta w_q'\Delta\mathbf{z}\right]\Bigg) \\
&= \frac{1}{2}\sum_j \Bigg(\sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\bar{w}_{qj}'\bar{\mathbf{z}}_j + \left(\sum_{i=1}^{2}\bar{w}_{ij}\mathbb{E}_j\left[(g_i-\mu)(g_q-\mu)\right]\right)\bar{w}_{qj}'\bar{\mathbf{z}}_j \\
&\quad + \sigma_\mu^2\bar{w}_{qj}'\bar{\mathbf{z}}_j\sum_{i=1}^{2}\mathbb{E}_j[\Delta w_i] \\
&\quad + \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\left[\bar{\mathbf{z}}_j\mathbb{E}_j[\Delta w_q'] + \bar{w}_{qj}'\mathbb{E}_j[\Delta\mathbf{z}]\right] \\
&\quad + \sigma_\mu^2\sum_{i=1}^{2}\mathbb{E}_j\left[\Delta w_i\Delta w_q'\right]\bar{\mathbf{z}}_j + \sigma_\mu^2\sum_{i=1}^{2}\mathbb{E}_j\left[\Delta w_i\Delta\mathbf{z}\right]\bar{w}_{qj}' \\
&\quad + \sigma_\mu^2\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\mathbb{E}_j\left[\Delta w_q'\Delta\mathbf{z}\right] + \bar{w}_{qj}'\mathcal{O}\left(n^{-3/2}\right)\Bigg)
\end{aligned}
$$

Taylor expansion of the gates and Lemma F.6 give

$$\begin{aligned}\nabla_{\theta_q}\mathcal{L}(\theta) = \frac{1}{2}\sum_j \Bigg\{ & \sigma_\mu^2 \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right) \bar{w}'_{qj}\bar{\mathbf{z}}_j + \frac{\bar{w}'_{qj}}{n}\sum_{i=1}^{2} \bar{w}_{ij}\Sigma_{qi}^{(j)}\bar{\mathbf{z}}_j \\ & + \sigma_\mu^2 \bar{w}'_{qj}\bar{\mathbf{z}}_j \sum_{i=1}^{2} \bar{w}'_{ij}\theta_i^\top \mathbb{E}_j[\Delta\mathbf{z}] \\ & + \sigma_\mu^2 \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right) \left[\bar{w}''_{qj}\bar{\mathbf{z}}_j\theta_q^\top \mathbb{E}_j[\Delta\mathbf{z}] + \bar{w}'_{qj}\mathbb{E}_j[\Delta\mathbf{z}]\right] \\ & + \frac{\sigma_\mu^2}{n}\sum_{i=1}^{2} \bar{w}'_{ij}\left[\bar{w}''_{qj}\left(\theta_q^\top \mathbf{\Sigma}_Z^{(j)}\theta_i\right)\bar{\mathbf{z}}_j + \bar{w}'_{qj}\mathbf{\Sigma}_Z^{(j)}\theta_i\right] + \frac{\sigma_\mu^2}{n}\bar{w}''_{qj}\left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right)\mathbf{\Sigma}_Z^{(j)}\theta_q \\ & + \frac{\sigma_\mu^2}{2n}\bar{w}'_{qj}\left[\sum_{i=1}^{2} \bar{w}''_{ij}\left(\theta_i^\top \mathbf{\Sigma}_Z^{(j)}\theta_i\right)\right]\bar{\mathbf{z}}_j \\ & + \frac{\sigma_\mu^2}{2n}\left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right)\bar{w}'''_{qj}\left(\theta_q^\top \mathbf{\Sigma}_Z^{(j)}\theta_q\right)\bar{\mathbf{z}}_j + \bar{w}'_{qj}\mathcal{O}(n^{-3/2})\Bigg\}\end{aligned}$$

□

Lemma D.3 (Population loss expansion)**.** *Let θ_q be the gating parameter vector for expert $q \in \{1, 2\}$, and let the loss function $\mathcal{L}(\theta)$ be defined as:*

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}\left[\left(\sum_{i=1}^{2} w_i(\mathbf{X})g_i - \mu\right)^2\right]$$

Then the expected loss can be expanded as:

$$\begin{aligned}\mathcal{L}(\theta) = \frac{1}{4}\sum_j \Bigg\{ & \sigma_\mu^2 \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right)^2 + \frac{1}{n}\sum_{i=1}^{2}\sum_{q=1}^{2} \bar{w}_{ij}\bar{w}_{qj}\Sigma_{iq}^{(j)} \\ & + \frac{\sigma_\mu^2}{n}\sum_{i=1}^{2}\sum_{q=1}^{2} \bar{w}'_{ij}\bar{w}'_{qj}\left(\theta_i^\top \mathbf{\Sigma}_Z^{(j)}\theta_q\right) \\ & + 2\sigma_\mu^2 \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right)\sum_{i=1}^{2} \bar{w}'_{ij}\theta_i^\top \mathbb{E}_j[\Delta\mathbf{z}] \\ & + \frac{\sigma_\mu^2}{n}\left(\sum_{i=1}^{2} \bar{w}_{ij} - 1\right)\sum_{i=1}^{2} \bar{w}''_{ij}\left(\theta_i^\top \mathbf{\Sigma}_Z^{(j)}\theta_i\right)\Bigg\} + \mathcal{O}(n^{-3/2})\end{aligned}$$

Proof. We can directly compute

$$
\begin{aligned}
\mathcal{L}(\theta) &= \frac{1}{4}\sum_j \mathbb{E}_j\left[\mu^2\left(\sum_{i=1}^2 w_i(\mathbf{X})-1\right)^2+\left(\sum_{i=1}^2 w_i(\mathbf{X})(g_i-\mu)\right)^2\right] \\
&= \frac{1}{4}\sum_j \mathbb{E}_j\left[\sigma_\mu^2\left(\sum_{i=1}^2 w_i(\mathbf{X})-1\right)^2+\frac{1}{n}\sum_{i=1}^2\sum_{q=1}^2 w_i(\mathbf{X})w_q(\mathbf{X})\Sigma_{iq}^{(j)}\right]+\mathcal{O}(n^{-3/2}) \\
&= \frac{1}{4}\sum_j \mathbb{E}_j\Bigg[\sigma_\mu^2\left(\sum_{i=1}^2(\bar{w}_{ij}+\Delta w_i)-1\right)^2 \\
&\quad +\frac{1}{n}\sum_{i=1}^2\sum_{q=1}^2(\bar{w}_{ij}+\Delta w_i)(\bar{w}_{qj}+\Delta w_q)\Sigma_{iq}^{(j)}\Bigg]+\mathcal{O}(n^{-3/2}) \\
&= \frac{1}{4}\sum_j\Bigg\{\sigma_\mu^2\left(\sum_{i=1}^2\bar{w}_{ij}-1\right)^2+\frac{1}{n}\sum_{i=1}^2\sum_{q=1}^2\bar{w}_{ij}\bar{w}_{qj}\Sigma_{iq}^{(j)}+2\sigma_\mu^2\left(\sum_{i=1}^2\bar{w}_{ij}-1\right)\sum_{i=1}^2\mathbb{E}_j\left[\Delta w_i\right] \\
&\quad +\sigma_\mu^2\sum_{i=1}^2\sum_{q=1}^2\mathbb{E}_j\left[\Delta w_i\Delta w_q\right]+\frac{1}{n}\sum_{i=1}^2\sum_{q=1}^2\Sigma_{iq}^{(j)}\mathbb{E}_j\left[\Delta w_i\Delta w_q\right]\Bigg\}+\mathcal{O}(n^{-3/2}) \\
&= \frac{1}{4}\sum_j\Bigg\{\sigma_\mu^2\left(\sum_{i=1}^2\bar{w}_{ij}-1\right)^2+\frac{1}{n}\sum_{i=1}^2\sum_{q=1}^2\bar{w}_{ij}\bar{w}_{qj}\Sigma_{iq}^{(j)} \\
&\quad +\frac{\sigma_\mu^2}{n}\sum_{i=1}^2\sum_{q=1}^2\bar{w}'_{ij}\bar{w}'_{qj}\left(\theta_i^\top\mathbf{\Sigma}_Z^{(j)}\theta_q\right) \\
&\quad +2\sigma_\mu^2\left(\sum_{i=1}^2\bar{w}_{ij}-1\right)\sum_{i=1}^2\bar{w}'_{ij}\theta_i^\top\mathbb{E}_j[\Delta\mathbf{z}] \\
&\quad +\frac{\sigma_\mu^2}{n}\left(\sum_{i=1}^2\bar{w}_{ij}-1\right)\sum_{i=1}^2\bar{w}''_{ij}\left(\theta_i^\top\mathbf{\Sigma}_Z^{(j)}\theta_i\right)\Bigg\}+\mathcal{O}(n^{-3/2})
\end{aligned}
$$

□

D.1.1 FLOW REPARAMETERIZATION

With some abuse of notation, we write $\mathcal{L}(\bar{w}, \theta)$ as $\mathcal{L}(\theta)$, where θ is such that $\bar{w}_{ij} = \sigma(\langle \theta_i, \bar{\mathbf{z}}_j \rangle)$. Use S_{qj} from Lemma B.7. Recall that we are studying the gradient flow

$$
\begin{aligned}
\dot{\theta}_q &= -\nabla_{\theta_q} \mathcal{L}(\theta) \\
&= -\frac{1}{2} \sum_j \Bigg\{ S_{qj}(\bar{w}) \bar{w}'_{qj} \bar{\mathbf{z}}_j \\
&\quad + \sigma_\mu^2 \bar{w}'_{qj} \bar{\mathbf{z}}_j \sum_{i=1}^{2} \bar{w}'_{ij} \theta_i^\top \mathbb{E}_j[\Delta \mathbf{z}] \\
&\quad + \sigma_\mu^2 \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1 \right) \left[\bar{w}''_{qj} \bar{\mathbf{z}}_j \theta_q^\top \mathbb{E}_j[\Delta \mathbf{z}] + \bar{w}'_{qj} \mathbb{E}_j[\Delta \mathbf{z}] \right] \\
&\quad + \frac{\sigma_\mu^2}{n} \sum_{i=1}^{2} \bar{w}'_{ij} \left[\bar{w}''_{qj} \left(\theta_q^\top \mathbf{\Sigma}_Z^{(j)} \theta_i \right) \bar{\mathbf{z}}_j + \bar{w}'_{qj} \mathbf{\Sigma}_Z^{(j)} \theta_i \right] \\
&\quad + \frac{\sigma_\mu^2}{n} \bar{w}''_{qj} \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1 \right) \mathbf{\Sigma}_Z^{(j)} \theta_q \\
&\quad + \frac{\sigma_\mu^2}{2n} \bar{w}'_{qj} \left[\sum_{i=1}^{2} \bar{w}''_{ij} \left(\theta_i^\top \mathbf{\Sigma}_Z^{(j)} \theta_i \right) \right] \bar{\mathbf{z}}_j \\
&\quad + \frac{\sigma_\mu^2}{2n} \left(\sum_{i=1}^{2} \bar{w}_{ij} - 1 \right) \bar{w}'''_{qj} \left(\theta_q^\top \mathbf{\Sigma}_Z^{(j)} \theta_q \right) \bar{\mathbf{z}}_j + \bar{w}'_{qj} \mathcal{O}(n^{-3/2}) \Bigg\}
\end{aligned}
$$

We now re-parameterize the flow in terms of $\bar{w} = (\bar{w}_{ij})_{i\in[2],j\in\{\mathrm{G,A}\}}$, using $\bar{w}_{qk} = \sigma(\langle\theta_q, \bar{\mathbf{z}}_k\rangle)$ and the chain rule.

$$
\begin{aligned}
\frac{\partial \bar{w}_{qk}}{\partial t} &= \frac{\partial \bar{w}_{qk}}{\partial \theta_q}\frac{\partial \theta_q}{\partial t} \\
&= -\Bigg(\frac{1}{2}\sum_j \Bigg\{ S_{qj}(\bar{w})\bar{w}'_{qj}\bar{w}'_{qk}\,\langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k\rangle \\
&\quad + \sigma_\mu^2 \bar{w}'_{qj}\bar{w}'_{qk}\,\langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k\rangle \sum_{i=1}^{2} \bar{w}'_{ij}\theta_i^\top \mathbb{E}_j[\Delta\mathbf{z}] \\
&\quad + \sigma_\mu^2 \bar{w}'_{qk}\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\Big[\bar{w}''_{qj}\,\langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k\rangle\,\theta_q^\top \mathbb{E}_j[\Delta\mathbf{z}] \\
&\qquad\qquad + \bar{w}'_{qj}\,\langle \mathbb{E}_j[\Delta\mathbf{z}], \bar{\mathbf{z}}_k\rangle\Big] \\
&\quad + \frac{\bar{w}'_{qk}\sigma_\mu^2}{n}\sum_{i=1}^{2}\bar{w}'_{ij}\Big[\bar{w}''_{qj}\left(\theta_q^\top \mathbf{\Sigma}_Z^{(j)}\theta_i\right)\langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k\rangle \\
&\qquad\qquad + \bar{w}'_{qj}\left\langle \mathbf{\Sigma}_Z^{(j)}\theta_i, \bar{\mathbf{z}}_k\right\rangle\Big] \\
&\quad + \frac{\sigma_\mu^2 \bar{w}''_{qj}\bar{w}'_{qk}}{n}\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\left\langle \mathbf{\Sigma}_Z^{(j)}\theta_q, \bar{\mathbf{z}}_k\right\rangle \\
&\quad + \frac{\sigma_\mu^2}{2n}\bar{w}'_{qj}\bar{w}'_{qk}\left[\sum_{i=1}^{2}\bar{w}''_{ij}\left(\theta_i^\top \mathbf{\Sigma}_Z^{(j)}\theta_i\right)\right]\langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k\rangle \\
&\quad + \frac{\sigma_\mu^2}{2n}\bar{w}'_{qk}\left(\sum_{i=1}^{2}\bar{w}_{ij} - 1\right)\bar{w}'''_{qj}\left(\theta_q^\top \mathbf{\Sigma}_Z^{(j)}\theta_q\right)\langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k\rangle \\
&\quad + \bar{w}'_{qk}\bar{w}'_{qj}\mathcal{O}(n^{-3/2})\Bigg\}\Bigg)
\end{aligned}
$$

D.2 Concentration bounds

D.2.1 Uniform concentration bound for the empirical gradient

Lemma D.4 (Concentration bound for the gradient of θ)**.** *Let $B_d(r)$ denote the Euclidean ball of radius r in $\mathbb{R}^d$. Fix $j \in \{\mathrm{G,A}\}$ and sample the m tasks independently conditional on $\tau_b = j$, writing $\bar{\mathbf{z}} = \bar{\mathbf{z}}_j$ and $\bar{w}_k = \bar{w}_{kj}$. Fix σ_μ^2 and λ independently of n, with either data law from Definition 2.1. Let $0 < \beta < 1/2$, $0 < \delta < 1$, $r = \mathcal{O}(\log n)$, and suppose $m, \delta^{-1} \in \mathrm{poly}(n)$. For sufficiently large n, with probability at least $1-\delta$, simultaneously for all $\theta_1, \theta_2 \in B_d(r)$ and all $q \in \{1,2\}$ and $s \in [d]$,*

$$
\begin{aligned}
&\left|\frac{1}{m}\sum_{b=1}^{m}\left[\left(\sum_{k=1}^{2} w_k(\mathbf{X}_b)g_k(\mathbf{X}_b) - \mu_b\right) g_q(\mathbf{X}_b)w'_q(\mathbf{X}_b)z_s(\mathbf{X}_b)\right]\right. \\
&\qquad \left. - \mathbb{E}_j\left[\left(\sum_{k=1}^{2} w_k(\mathbf{X})g_k - \mu\right) g_q w'_q(\mathbf{X})z_s\right]\right| \\
&\le \left(\left|\sum_{k=1}^{2}\bar{w}_k - 1\right| + n^{-1/2+\beta}\right)\bar{w}_q(1-\bar{w}_q)\,\widetilde{\mathcal{O}}(m^{-1/2}).
\end{aligned}
$$

Proof. Set $M = n^{-1/2+\beta}$ and choose $R, C = \mathcal{O}(\sqrt{\log n})$, $C \ge 1$, with sufficiently large constants. Apply Lemma F.17 with $k = 1$ at the fixed $\pm\lambda_k$ and ω_s, using deviations equal to a sufficiently small constant multiple of M. Since $nM^2/(R^2 e^{2R|\lambda|}) = n^{2\beta - o(1)}$, symmetry of the centered observations,

the fixed d and $\boldsymbol{\omega}$, and the Gaussian tail of μ give

$$\mathcal{E} = \left\{ \max_k |g_k - \mu| \le M,\ |\mu| \le C,\ \|\mathbf{z} - \bar{\mathbf{z}}\|_2 \le M \right\}, \qquad \mathbb{P}_j(\mathcal{E}^c) = \mathcal{O}(n^{-A})$$

for any prescribed fixed $A > 0$. Apply the deterministic covers in Lemma D.5 to the integrands multiplied by $\mathbf{1}_{\mathcal{E}}$. The discarded normalized integrand is bounded by

$$\frac{3e^{2r}}{M} \left(|\mu| + \max_{\ell \in [n]} |X_\ell - \mu| \right)^2,$$

since $w'_q/\bar{w}'_q \le e^{2r}$ and $\Gamma_1 \ge M$. Its L^2 norm is polynomial in n, so Cauchy–Schwarz and sufficiently large A make the discarded expectation $o(m^{-1/2})$ and the probability of discarding any training task at most $\delta/2$. Use failure probability $\delta/2$ for the fixed number of coordinates in the helper bound, where $rM = o(1)$ and its remaining factors are logarithmic. □

Lemma D.5 (Concentration bound for the normalized gradient)**.** *Fix $j \in \{\mathrm{G}, \mathrm{A}\}$, a gate index $q \in [2]$, a coordinate $s \in [d]$, a radius $r > 0$ and a failure probability $\delta \in (0, 1)$, and sample m tasks independently conditional on $\tau_b = j$, writing $\bar{\mathbf{z}} = \bar{\mathbf{z}}_j$ and $\bar{w}_k = \bar{w}_{kj}$. Suppose $M > 0$, $C \ge 1$, and, almost surely for a single task and every k,*

$$|g_k(\mathbf{X}) - \mu| \le M, \qquad |\mu| \le C, \qquad \|\mathbf{z}(\mathbf{X}) - \bar{\mathbf{z}}\|_2 \le M, \qquad \|\bar{\mathbf{z}}\|_2 \le C.$$

Let

$$w_k(\mathbf{X}) = \sigma(\langle \theta_k, \mathbf{z}(\mathbf{X}) \rangle), \qquad \bar{w}_k = \sigma(\langle \theta_k, \bar{\mathbf{z}} \rangle),$$

and define

$$\Gamma_1 = \left| \sum_{k=1}^{2} \bar{w}_k - 1 \right| + M, \qquad \Gamma_2 = \bar{w}_q(1 - \bar{w}_q),$$

$$\Psi_{\theta,q,s}(\mathbf{X}, \mu) = \left(\sum_{k=1}^{2} w_k(\mathbf{X}) g_k(\mathbf{X}) - \mu \right) g_q(\mathbf{X}) w'_q(\mathbf{X}) z_s(\mathbf{X}).$$

Then, with probability at least $1 - \delta$,

$$\begin{aligned} &\sup_{\theta_1, \theta_2 \in B_d(r)} \frac{1}{\Gamma_1 \Gamma_2} \left| \frac{1}{m} \sum_{b=1}^{m} \Psi_{\theta,q,s}(\mathbf{X}_b, \mu_b) - \mathbb{E}_j[\Psi_{\theta,q,s}(\mathbf{X}, \mu)] \right| \\ &\lesssim 2(C + M + Cr) e^{rM} (M + C)^2 \\ &\qquad \times \left[\frac{2d \log\{2 + 2mr(C + M)(1 + M^{-1})\} + \log(2/\delta)}{m} \right]^{1/2}. \end{aligned}$$

Proof. Here L_∞ denotes the supremum over the full support of a single task. All covers below are deterministic. Define

$$\mathcal{F}_a = \left\{ f_\theta : (\mathbf{z}, \mathbf{g}, \mu) \mapsto \frac{1}{\Gamma_1} \left(\sum_{k=1}^{2} \sigma(\theta_k^\top \mathbf{z}) g_k - \mu \right) : \theta_k \in B_d(r) \right\}.$$

For $f_\theta \in \mathcal{F}_a$ we can write

$$f_\theta = \frac{1}{\Gamma_1} \left(\sum_{k=1}^{2} w_k(g_k - \mu) + \mu \sum_{k=1}^{2} (w_k - \bar{w}_k) + \mu \left(\sum_{k=1}^{2} \bar{w}_k - 1 \right) \right).$$

Since $|w_k - \bar{w}_k| \le rM/4$ and $C \ge 1$, this gives

$$\|f_\theta\|_\infty \le 2(C + M + Cr) \equiv B_a.$$

The normalizing factor also depends on θ, with

$$|\Gamma_1(\theta) - \Gamma_1(\theta')| \le \frac{C}{2} \max_k \|\theta_k - \theta'_k\|_2.$$

Using $\Gamma_1 \geq M$ and the sigmoid's $1/4$-Lipschitz bound, we obtain

$$\|f_\theta - f_{\theta'}\|_\infty \leq \frac{(C+M)^2 + B_a C}{2M} \max_k \|\theta_k - \theta'_k\|_2 \leq \frac{2B_a(C+M)}{M} \max_k \|\theta_k - \theta'_k\|_2.$$

A Euclidean cover of each $B_d(r)$ therefore gives

$$\log \mathcal{N}(u, \mathcal{F}_a, L_\infty) \lesssim 2d \log\left(2 + \frac{2rB_a(C+M)}{Mu}\right). \tag{16}$$

Define

$$\mathcal{F}_b = \left\{(\mathbf{z}, \mathbf{g}, \mu) \mapsto \frac{w'_q g_q z_s}{\Gamma_2} : \theta_q \in B_d(r)\right\}.$$

Since $|\frac{d}{du} \log \sigma'(u)| \leq 1$,

$$\frac{w'_q}{\Gamma_2} \leq e^{rM}, \qquad \left\|\nabla_{\theta_q} \log \frac{w'_q}{\Gamma_2}\right\|_2 \leq M + 2C.$$

Thus the functions in $\mathcal{F}_b$ are bounded in absolute value by $e^{rM}(M+C)^2$ and have Lipschitz constant at most $e^{rM}(M+C)^2(M+2C)$ in θ_q, including the dependence of Γ_2. Let $\mathcal{F}$ consist of the products $f_a f_b$ with the same parameters in both factors. These products are bounded in absolute value by $B_a e^{rM}(M+C)^2$. A deterministic parameter cover with mesh $[4m(C+M)(1+M^{-1})]^{-1}$ gives

$$\log \mathcal{N}\left(\frac{B_a e^{rM}(M+C)^2}{m}, \mathcal{F}, L_\infty\right) \lesssim 2d \log\left(2 + 2mr(C+M)(1+M^{-1})\right).$$

Hoeffding's inequality and a union bound over this cover, followed by approximation in L_∞, give, with probability at least $1-\delta$,

$$\sup_{f \in \mathcal{F}} \left|\frac{1}{m}\sum_{b=1}^{m} f(\mathbf{X}_b) - \mathbb{E}_j[f]\right| \lesssim B_a e^{rM}(M+C)^2 \times \left[\frac{2d \log\left(2 + 2mr(C+M)(1+M^{-1})\right) + \log(2/\delta)}{m}\right]^{1/2}.$$

The approximation contributes at most $2B_a e^{rM}(M+C)^2/m$, which is absorbed in this bound. □

D.2.2 Uniform concentration for the loss

Lemma D.6 (Local concentration for the loss)**.** *Under the assumptions of Lemma D.4, with probability at least $1-\delta$, uniformly over $\theta_1, \theta_2 \in B_d(r)$ satisfying $|\sum_k \bar{w}_k - 1| = \mathcal{O}(n^{-1})$ with a fixed implied constant,*

$$|\hat{\mathcal{L}}_m^{(j)}(\theta) - \mathcal{L}^{(j)}(\theta)| = \widetilde{\mathcal{O}}\left(\frac{n^{-1+\beta}}{\sqrt{m}} + \frac{n^{-1+2\beta}}{m}\right).$$

Proof. Use $\mathcal{E}$, M, and C from the proof of Lemma D.4, and apply the covers in Lemma D.7 to $\mathcal{F}_{sq}\mathbf{1}_{\mathcal{E}}$. The normalized squared residual has envelope

$$\frac{9}{M^2}\left(|\mu| + \max_{\ell \in [n]} |X_\ell - \mu|\right)^2.$$

Cauchy–Schwarz makes the discarded expectation $o(m^{-1})$, with the probability of discarding a training task absorbed in δ. The fixed-parameter moment-ratio argument in the proof of Lemma 3.5, applied also to the paired experts, and independence of μ from the centered observations give

$$\mathcal{L}^{(j)} = \mathcal{O}\left(\frac{(1+r)^2}{n}\right) = \widetilde{\mathcal{O}}(n^{-1}), \qquad \Gamma_1^2 U = \widetilde{\mathcal{O}}(n^{-1+2\beta}).$$

Substitute into Lemma D.7. The discarded expectation is absorbed by its linear term. □

Lemma D.7 (Uniform concentration for the loss)**.** *Fix $j \in \{\mathrm{G}, \mathrm{A}\}$ and sample the m tasks independently conditional on $\tau_b = j$, writing $\bar{\mathbf{z}} = \bar{\mathbf{z}}_j$ and $\bar{w}_k = \bar{w}_{kj}$. Suppose $M > 0$, $C \geq 1$, and the following bounds hold almost surely for a single task and every k:*

$$\begin{aligned} |g_k(\mathbf{X}) - \mu| &\leq M \\ |\mu| &\leq C \\ \|\mathbf{z}(\mathbf{X}) - \bar{\mathbf{z}}\|_2 &\leq M \\ \|\bar{\mathbf{z}}\|_2 &\leq C \\ w_k(\mathbf{X}) &= \sigma(\langle \theta_k, \mathbf{z}(\mathbf{X}) \rangle) \\ \bar{w}_k &= \sigma(\langle \theta_k, \bar{\mathbf{z}} \rangle) \\ \Gamma_1 &= \left| \sum_{k=1}^{2} \bar{w}_k - 1 \right| + M \\ U &= 4(C + M + Cr)^2 \end{aligned}$$

Let

$$\hat{\mathcal{L}}_m^{(j)}(\theta) = \frac{1}{2m} \sum_{b=1}^{m} \left(\sum_{k=1}^{2} w_k(\mathbf{X}_b) g_k(\mathbf{X}_b) - \mu_b \right)^2, \quad \mathcal{L}^{(j)}(\theta) = \frac{1}{2} \mathbb{E}_j \left[\left(\sum_{k=1}^{2} w_k(\mathbf{X}) g_k(\mathbf{X}) - \mu \right)^2 \right]$$

Then, with probability at least $1 - \delta$, the following holds for all $\theta_1, \theta_2 \in B_d(r)$:

$$\begin{aligned} \left| \hat{\mathcal{L}}_m^{(j)}(\theta) - \mathcal{L}^{(j)}(\theta) \right| \lesssim \sqrt{\mathcal{L}^{(j)}(\theta)\, \Gamma_1^2 U \, \frac{2d \log\left(2 + 2mr(C+M)(1+M^{-1})\right) + \log(2/\delta)}{m}} \\ + \Gamma_1^2 U \, \frac{2d \log\left(2 + 2mr(C+M)(1+M^{-1})\right) + \log(2/\delta)}{m}. \end{aligned}$$

Proof. Use $\mathcal{F}_a$ from the proof of Lemma D.5, with $B_a = 2(C + M + Cr)$, and let

$$\mathcal{F}_{sq} = \{ f_\theta^2 : f_\theta \in \mathcal{F}_a \}.$$

For every $u \in \mathcal{F}_{sq}$,

$$0 \leq u \leq B_a^2 = U, \qquad \mathrm{Var}(u) \leq \mathbb{E}_j[u^2] \leq U \mathbb{E}_j[u].$$

Since squaring is $2B_a$-Lipschitz on $[-B_a, B_a]$, (16) gives

$$\log \mathcal{N}(u, \mathcal{F}_{sq}, L_\infty) \lesssim 2d \log\left(2 + \frac{2rU(C+M)}{Mu}\right).$$

Choose $u_0 = U/m$ and a deterministic u_0-cover $\mathcal{C}_{u_0} \subset \mathcal{F}_{sq}$ in L_∞ over the full support. Then

$$\log |\mathcal{C}_{u_0}| \lesssim 2d \log\left(2 + 2mr(C+M)(1+M^{-1})\right).$$

Bernstein's inequality and a union bound give, simultaneously for every $v \in \mathcal{C}_{u_0}$, with probability at least $1 - \delta$,

$$\left| \frac{1}{m} \sum_{b=1}^{m} v(\mathbf{X}_b) - \mathbb{E}_j[v] \right| \leq \sqrt{\frac{2U \mathbb{E}_j[v] \log(2|\mathcal{C}_{u_0}|/\delta)}{m}} + \frac{2U \log(2|\mathcal{C}_{u_0}|/\delta)}{3m}.$$

For any $u \in \mathcal{F}_{sq}$ choose v in this cover with $\|u - v\|_\infty \leq U/m$. This controls both the empirical average and the expectation, so

$$\left| \frac{1}{m} \sum_{b=1}^{m} u(\mathbf{X}_b) - \mathbb{E}_j[u] \right| \leq \left| \frac{1}{m} \sum_{b=1}^{m} v(\mathbf{X}_b) - \mathbb{E}_j[v] \right| + \frac{2U}{m}, \qquad \mathbb{E}_j[v] \leq \mathbb{E}_j[u] + \frac{U}{m}.$$

The resulting approximation terms are absorbed by the linear term in Bernstein's bound. Finally, for $u = f_\theta^2$,

$$\hat{\mathcal{L}}_m^{(j)}(\theta) = \frac{\Gamma_1^2}{2m} \sum_{b=1}^{m} u(\mathbf{X}_b), \qquad \mathcal{L}^{(j)}(\theta) = \frac{\Gamma_1^2}{2} \mathbb{E}_j[u].$$

Multiplying by $\Gamma_1^2/2$ proves the claimed bound simultaneously for all parameters. □

D.3 Analysis of individual stationary points: complete flow

D.3.1 Proof of Lemma B.7

Proof. Since $Z \succ 0$,

$$\frac{d}{dt}\mathcal{L}^{\text{truncated}}(\bar{w}) = -\frac{1}{4}\sum_{q=1}^{2}\left\|\sum_j \bar{w}'_{qj} S_{qj}(\bar{w})\bar{\mathbf{z}}_j\right\|_2^2,$$
$$\dot{\bar{w}} = 0 \quad \Longleftrightarrow \quad \bar{w}'_{qj} S_{qj}(\bar{w}) = 0 \quad \text{for all } q, j.$$

For $k \neq k_j$, the strict hierarchy gives

$$S_{kj}(\bar{w}) - S_{k_j j}(\bar{w}) = \frac{\Sigma^{(j)}_{kk_j} - \Sigma^{(j)}_{k_j k_j}}{n}\bar{w}_{k_j j} + \frac{\Sigma^{(j)}_{kk} - \Sigma^{(j)}_{kk_j}}{n}\bar{w}_{kj} > 0$$

whenever $\bar{w}_{k_j j} + \bar{w}_{kj} > 0$, excluding two interior gates in the same context. For $q \neq k$ and $0 < \bar{w}^\circ_{qj} < 1$, stationarity requires

$$\bar{w}^\circ_{kj} = 1 \quad \Longrightarrow \quad 0 = S_{qj}(\bar{w}^\circ) = \left(\sigma_\mu^2 + \frac{\Sigma^{(j)}_{qq}}{n}\right)\bar{w}^\circ_{qj} + \frac{\Sigma^{(j)}_{qk}}{n} > 0,$$

$$\bar{w}^\circ_{kj} = 0 \quad \Longrightarrow \quad \bar{w}^\circ_{qj} = \frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \Sigma^{(j)}_{qq}}.$$

The four corners have $\bar{w}'_{qj} = 0$, completing the classification. The unique optimum follows from convexity, strict convexity on the selected coordinates, and

$$\frac{\partial \mathcal{L}^{\text{truncated}}}{\partial \bar{w}_{qj}} = \frac{1}{2}S_{qj}, \qquad S_{k_j j}(\bar{w}^*) = 0,$$
$$S_{kj}(\bar{w}^*) = \frac{\bar{w}^*_{k_j j}}{n}\left(\Sigma^{(j)}_{kk_j} - \Sigma^{(j)}_{k_j k_j}\right) > 0 \quad (k \neq k_j).$$

□

D.3.2 Proof of Lemma B.8

Proof of Lemma B.8. Whenever $\bar{w}_{qj} \in [1/4, 3/4]$, we have $\bar{w}'_{qj} \geq 3/16$, and therefore

$$\begin{aligned}\|\nabla_\theta \mathcal{L}^{\text{bias}}\|_2^2 &= \frac{\sigma_\mu^4}{4}\sum_{q=1}^{2}\left\|\sum_j (\bar{w}_{1j} + \bar{w}_{2j} - 1)\bar{w}'_{qj}\bar{\mathbf{z}}_j\right\|_2^2 \\ &\geq \frac{\sigma_\mu^4 \lambda_{\min}(Z)}{4}\sum_{q,j}(\bar{w}_{1j} + \bar{w}_{2j} - 1)^2(\bar{w}'_{qj})^2 \\ &\geq 2\alpha\mathcal{L}^{\text{bias}}.\end{aligned}$$

Consequently, locally the flow satisfies a PL inequality. Moreover,

$$-\frac{d}{dt}\sqrt{\mathcal{L}^{\text{bias}}} = \frac{\|\nabla_\theta \mathcal{L}^{\text{bias}}\|_2^2}{2\sqrt{\mathcal{L}^{\text{bias}}}} \geq \sqrt{\frac{\alpha}{2}}\,\|\dot{\theta}\|_2$$

Thus,

$$\|\theta(0) - \theta(t)\|_2 \leq \int_0^t \|\dot{\theta}(s)\|_2\, ds \leq \sqrt{\frac{2\mathcal{L}^{\text{bias}}(\theta(0))}{\alpha}}.$$

Since the map $\theta \mapsto \bar{w}$ is globally $\sqrt{\lambda_{\max}(Z)}/4$-Lipschitz, we obtain

$$\int_0^t \|\dot{\bar{w}}(s)\|_{\text{F}}\, ds \leq \frac{\sqrt{\lambda_{\max}(Z)}}{4}\sqrt{\frac{2\mathcal{L}^{\text{bias}}(\theta(0))}{\alpha}}.$$

Since $\max_{q,j} |\bar{w}_{qj}(0) - 1/2| \le \delta$, we have $\mathcal{L}^{\text{bias}}(\theta(0)) \le 2\sigma_\mu^2\delta^2$. Thus we obtain a travel bound on $\bar{w}$ of

$$\|\bar{w}(0) - \bar{w}(t)\|_{\text{F}} \le C\delta.$$

With $(1 + C)\delta < 1/4$, a first-exit argument shows that the flow remains in $\bar{w} \in [1/4, 3/4]^{2\times 2}$ for all time. Moreover,

$$\frac{d}{dt}\mathcal{L}^{\text{bias}} = -\|\nabla_\theta \mathcal{L}^{\text{bias}}\|_2^2 \le -2\alpha\mathcal{L}^{\text{bias}},$$

which gives the asserted exponential bound on the sum residuals.

The truncated gradient differs from the bias gradient by $\mathcal{O}(n^{-1})$ at fixed λ. The same bound holds for the full population gradient on bounded sets of θ, by the loss and gradient expansions. The PL inequality and $\|\nabla_\theta \mathcal{L}^{\text{bias}}\|_2 \le C\sqrt{\mathcal{L}^{\text{bias}}}$ give

$$\frac{d}{dt}\sqrt{\mathcal{L}^{\text{bias}}} \le -\alpha\sqrt{\mathcal{L}^{\text{bias}}} + \frac{C}{n}, \qquad \sqrt{\mathcal{L}^{\text{bias}}(\theta(t))} \le e^{-\alpha t}\sqrt{\mathcal{L}^{\text{bias}}(\theta(0))} + \frac{C}{\alpha n}.$$

At zero loss the inequality is interpreted using the upper right derivative. Integrating $\|\dot{\theta}\|_2 \le C\sqrt{\mathcal{L}^{\text{bias}}} + C/n$ now bounds the gate travel by $C\delta + C(1 + t)/n$. Choose $t_{\text{crit}} = 2\alpha^{-1}\log n$ and then n sufficiently large. The same first-exit argument keeps the gates in $[1/4, 3/4]$ up to t_{crit}, and the displayed bound gives the claimed $\mathcal{O}(n^{-1})$ residual. For the full population flow, the same travel bound keeps θ in a fixed bounded set up to t_{crit}.

For the empirical flow, apply Lemma D.4 with $\beta = \min\{\epsilon, 1\}/4$ and failure probabilities of order $1/n$, conditional on the family labels. Hoeffding's inequality bounds the deviations of the family proportions from $1/2$ by $\widetilde{\mathcal{O}}(m^{-1/2})$. On bounded sets of θ, the conditional gradient expansion bounds each family's population gradient by $C\max_j |\bar{w}_{1j} + \bar{w}_{2j} - 1| + C/n$. Combining these bounds gives, with probability at least $1 - 1/n$, uniformly on these sets,

$$\|\nabla_\theta(\hat{\mathcal{L}}_m - \mathcal{L})\|_2 \le \widetilde{\mathcal{O}}(n^{-1/2-\epsilon/2}) \max_j |\bar{w}_{1j} + \bar{w}_{2j} - 1| + \widetilde{\mathcal{O}}(n^{-1-\epsilon/4}). \tag{17}$$

For fixed $\sigma_\mu^2 > 0$, the first term is $o(1)\sqrt{\mathcal{L}^{\text{bias}}}$ and the second is $o(n^{-1})$. Up to first exit from the bounded parameter set and $[1/4, 3/4]^{2\times 2}$, the preceding PL inequality therefore gives

$$\frac{d}{dt}\sqrt{\mathcal{L}^{\text{bias}}} \le -\frac{\alpha}{2}\sqrt{\mathcal{L}^{\text{bias}}} + \frac{C}{n}, \qquad \sqrt{\mathcal{L}^{\text{bias}}(\theta(t))} \le e^{-\alpha t/2}\sqrt{\mathcal{L}^{\text{bias}}(\theta(0))} + \frac{2C}{\alpha n}.$$

The same integration of $\|\dot{\theta}\|_2$ bounds the travel by $C\delta + C(1 + t)/n$ and prevents exit up to $t_{\text{crit}} = 2\alpha^{-1}\log n$. This proves the empirical residual and parameter bounds. □

D.3.3 Proof of Lemma B.9

Proof of Lemma B.9. First, we check the assumptions of Lemma D.8. By Lemma F.5, almost surely over $\boldsymbol{\omega}$ we have $Z \succ 0$, $Z_{\text{GA}} \ge 0$ and $Z_{\text{GA}} < \min\{Z_{\text{GG}}, Z_{\text{AA}}\}$, with $\lambda_{\min}(Z)$ bounded below independently of n.

Set $s = (t - t_{\text{crit}})/n$ and

$$r_j = n\sigma_\mu^2(\bar{w}_{1j} + \bar{w}_{2j} - 1).$$

The truncated flow satisfies

$$\frac{1}{n\sigma_\mu^2}\frac{dr_k}{ds} = -\frac{1}{2}\bar{w}'_{1k}\sum_j Z_{jk}\bar{w}'_{1j}\left(2r_j + \sum_{i=1}^{2}\bar{w}_{ij}(\Sigma_{1i}^{(j)} + \Sigma_{2i}^{(j)})\right) + \mathcal{O}\left(\frac{1}{n\sigma_\mu^2}\right),$$

$$\frac{d}{ds}\left[\frac{\sigma^{-1}(\bar{w}_{1k}) - \sigma^{-1}(\bar{w}_{2k})}{2}\right] = -\frac{1}{4}\sum_j Z_{jk}\bar{w}'_{1j}n(S_{1j} - S_{2j})(\bar{w}) + \mathcal{O}\left(\frac{1}{n\sigma_\mu^2}\right),$$

where S_{qj} is defined in Lemma B.7. On bounded coordinate sets, the vector fields are bounded through two derivatives, and the remainder bounds hold through two derivatives.

Initialize the reduced flow in Lemma D.8 by

$$\bar{w}_{1j}^{\text{reduced}}(0) = \sigma\left(\frac{\sigma^{-1}(\bar{w}_{1j}(t_{\text{crit}})) - \sigma^{-1}(\bar{w}_{2j}(t_{\text{crit}}))}{2}\right), \qquad \bar{w}_{2j}^{\text{reduced}}(0) = 1 - \bar{w}_{1j}^{\text{reduced}}(0).$$

Then

$$\begin{aligned}\frac{d}{ds}\left(n\mathcal{L}^{\text{truncated}}(\bar{w}^{\text{reduced}})\right) &= -\frac{1}{8}\sum_{j,k}\bar{w}_{1j}^{\text{reduced}\,\prime} n(S_{1j}-S_{2j})(\bar{w}^{\text{reduced}}) \\ &\quad \times Z_{jk}\bar{w}_{1k}^{\text{reduced}\,\prime} n(S_{1k}-S_{2k})(\bar{w}^{\text{reduced}}) \\ &\le -\frac{\lambda_{\min}(Z)}{8}\sum_{j}(\bar{w}_{1j}^{\text{reduced}\,\prime} n(S_{1j}-S_{2j})(\bar{w}^{\text{reduced}}))^2 \le 0.\end{aligned}$$

The sum is integrable and uniformly continuous, hence tends to zero. Since $n(S_{1\mathrm{G}}-S_{2\mathrm{G}})<0<n(S_{1\mathrm{A}}-S_{2\mathrm{A}})$, connectedness of the accumulation set gives convergence to a corner. Lemma D.8, also with indices interchanged, excludes the two corners selecting the same expert. Loss monotonicity excludes the remaining corner, which selects the wrong expert in both contexts. Indeed, on the critical manifold

$$\mathcal{L}^{\text{truncated}}(\bar{w})=\frac{1}{4n}\sum_{j}q_j(\bar{w}_{1j}),\qquad q_j(\pi)=\pi^2\Sigma_{11}^{(j)}+2\pi(1-\pi)\Sigma_{12}^{(j)}+(1-\pi)^2\Sigma_{22}^{(j)},$$

and $q_j''=2\,(1,-1)\mathbf{\Sigma}^{(j)}(1,-1)^\top\ge 0$ because $\mathbf{\Sigma}^{(j)}$ is a covariance matrix. Each q_j is therefore convex on $[0,1]$ and attains its maximum at an endpoint. As $q_j(1)=\Sigma_{11}^{(j)}$ and $q_j(0)=\Sigma_{22}^{(j)}$, which the strict hierarchy separates, the maximum is attained at the endpoint selecting the wrong expert in family j, and strictly so over $\bar{w}_{1j}\in[c,1-c]$ for any $c>0$. Summing over j, the corner selecting the wrong expert in both contexts is the strict maximizer of $\mathcal{L}^{\text{truncated}}$ on the manifold. Since $\bar{w}(t_{\text{crit}})\in[1/4,3/4]^{2\times 2}$ gives bounded logits, $\bar{w}^{\text{reduced}}(0)$ lies in $[c,1-c]^2$ for some absolute $c>0$, so the nonincreasing reduced flow cannot converge to that corner. Thus $\bar{w}^{\text{reduced}}(s)\to I_2$, and compactness, continuous dependence, and loss monotonicity give common $S,\kappa>0$ with

$$n\mathcal{L}^{\text{truncated}}(\bar{w}^{\text{reduced}}(S))\le\min_{\bar{w}^\circ\text{ in (15)}} n\mathcal{L}^{\text{truncated}}(\bar{w}^\circ)-4\kappa.$$

Bounded logit derivatives give $\bar{w}_{1j}^{\text{reduced}}(s)\in[c,1-c]$ on $[0,S]$, uniformly for some $0<c<1/2$.

Following Theorem G.2, we set

$$\begin{aligned}y&=\left(\frac{\sigma^{-1}(\bar{w}_{1j})-\sigma^{-1}(\bar{w}_{2j})}{2}\right)_{j\in\{\mathrm{G,A}\}},\qquad \varepsilon=\frac{1}{n\sigma_\mu^2},\\ y^0(s)&=\left(\sigma^{-1}(\bar{w}_{1j}^{\text{reduced}}(s))\right)_{j\in\{\mathrm{G,A}\}}.\end{aligned}$$

On $\bar{w}_{2j}=1-\bar{w}_{1j}$, the limiting normal equilibrium and stability bound are

$$\begin{aligned}r_j^\star(y)&=-\frac{1}{2}\sum_{i=1}^{2}\bar{w}_{ij}(\Sigma_{1i}^{(j)}+\Sigma_{2i}^{(j)}),\\ \lambda_{\max}\left(\left[-\bar{w}_{1k}'Z_{jk}\bar{w}_{1j}'\right]_{k,j\in\{\mathrm{G,A}\}}\right)&\le -c^2(1-c)^2\lambda_{\min}(Z)<0.\end{aligned}$$

The affine normal equation gives, up to first exit from a bounded interior neighborhood,

$$\left\|\left(r_j(s)+\frac{1}{2}\sum_{i=1}^{2}\bar{w}_{ij}(t_{\text{crit}}+ns)(\Sigma_{1i}^{(j)}+\Sigma_{2i}^{(j)})\right)_{j\in\{\mathrm{G,A}\}}\right\|_2\le Ce^{-\kappa_0 n\sigma_\mu^2 s}+\frac{C}{n\sigma_\mu^2}.$$

This bounds r uniformly up to first exit. The limiting slow equation is independent of r, so its difference from the actual slow equation is uniformly $\mathcal{O}(\varepsilon)$ on this region. Since $y(0)=y^0(0)$, the Lipschitz bound and Gronwall's inequality give, up to first exit,

$$\begin{aligned}\|y(s)-y^0(s)\|_2&\le C\int_0^s\|y(u)-y^0(u)\|_2\,du+C\varepsilon s,\\ \sup_{0\le s\le S}\|y(s)-y^0(s)\|_2&\le C_S\varepsilon.\end{aligned}$$

Since $\bar{w}_{1j} + \bar{w}_{2j} - 1 = \varepsilon r_j$, smooth inversion of the coordinates on the interior region gives

$$\sup_{0\le s\le S} \|\bar{w}(t_{\text{crit}} + ns) - \bar{w}^{\text{reduced}}(s)\|_{\text{F}} \le \frac{C_S}{n\sigma_\mu^2}, \qquad \sup_{0\le s\le S} \|r(s)\|_2 \le C_S.$$

These bounds prevent exit before S, and

$$\begin{aligned} n\mathcal{L}^{\text{truncated}}(\bar{w}(t_{\text{crit}} + nS)) &= n\mathcal{L}^{\text{truncated}}(\bar{w}^{\text{reduced}}(S)) + \mathcal{O}\left(\frac{1}{n\sigma_\mu^2}\right) \\ &< \min_{\bar{w}^\circ \text{ in (15)}} n\mathcal{L}^{\text{truncated}}(\bar{w}^\circ) - \kappa \end{aligned}$$

for sufficiently large n.

For the full population flow, the loss and gradient expansions give, on bounded sets of θ,

$$\begin{aligned} |\mathcal{L} - \mathcal{L}^{\text{truncated}}| + \|\nabla_\theta(\mathcal{L} - \mathcal{L}^{\text{truncated}})\|_2 &\le C\left(\frac{\sigma_\mu^2}{n} + n^{-3/2}\right), \\ \left\|\frac{d\theta_q^\perp}{ds}\right\|_2 &\le C(\sigma_\mu^2 + n^{-1/2}), \end{aligned}$$

uniformly in n and $0 < \sigma_\mu^2 \le 1$, by Lemma F.6. By Lemma B.5, bounded logits and perpendicular components give bounded θ. The coordinate errors are $\mathcal{O}(\sigma_\mu^2 + n^{-1/2})$, so normal stability and Gronwall's inequality give, up to first exit,

$$\sup_{0\le s\le S} \|\bar{w}(t_{\text{crit}} + ns) - \bar{w}^{\text{reduced}}(s)\|_{\text{F}} \le C_S\left(\sigma_\mu^2 + n^{-1/2} + \frac{1}{n\sigma_\mu^2}\right), \qquad \sup_{0\le s\le S} \|r(s)\|_2 \le C_S.$$

For sufficiently small σ_μ^2 and then sufficiently large n, these bounds prevent exit before S, and

$$\begin{aligned} &\max\left\{n\mathcal{L}^{\text{truncated}}(\bar{w}(t_{\text{crit}} + nS)),\ n\mathcal{L}(\theta(t_{\text{crit}} + nS))\right\} \\ &\le n\mathcal{L}^{\text{truncated}}(\bar{w}^{\text{reduced}}(S)) + C_S\left(\sigma_\mu^2 + n^{-1/2} + \frac{1}{n\sigma_\mu^2}\right) \\ &< \min_{\bar{w}^\circ \text{ in (15)}} n\mathcal{L}^{\text{truncated}}(\bar{w}^\circ) - \kappa. \end{aligned}$$

For the empirical flow, use the uniform concentration event in (17). Up to first exit from the same bounded region in $(r, y, \theta^\perp)$, bounded r gives $\max_j |\bar{w}_{1j} + \bar{w}_{2j} - 1| = \mathcal{O}(n^{-1})$ for fixed $\sigma_\mu^2 > 0$, and hence

$$n\|\nabla_\theta(\hat{\mathcal{L}}_m - \mathcal{L})\|_2 = o(1).$$

In the rescaled equations this adds $o(1)$ to the right-hand sides of the normal equation, the slow equation, and the equation for $\theta_q^\perp$. The normal stability estimate and Gronwall's inequality for the slow equation consequently give

$$\begin{aligned} \sup_{0\le s\le S} \|\bar{w}(t_{\text{crit}} + ns) - \bar{w}^{\text{reduced}}(s)\|_{\text{F}} &\le C_S\left(\sigma_\mu^2 + n^{-1/2} + \frac{1}{n\sigma_\mu^2}\right) + o(1), \\ \sup_{0\le s\le S} \|r(s)\|_2 &\le C_S + o(1). \end{aligned}$$

Also, $\sup_{0\le s\le S} \|\theta_q^\perp(t_{\text{crit}} + ns) - \theta_q^\perp(t_{\text{crit}})\|_2 \le C_S\sigma_\mu^2 + o(1)$. Choose the bounds on r and $\theta^\perp$ strictly larger than these estimates and the logit bounds strictly larger than those of the reduced flow. For sufficiently small fixed σ_μ^2 and then sufficiently large n, the estimates prevent exit before S. The preceding loss comparison acquires only an additional $o(1)$ term after multiplication by n, so the strict loss margin is preserved. Bounded logits and perpendicular components give the stated parameter bound by Lemma B.5. □

Lemma D.8 (Exit from a neighborhood of a suboptimal stationary point)**.** *Let* $Z_{ij} = \langle \bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j \rangle$ *and assume*

$$Z_{\text{GA}} < \min\left\{Z_{\text{GG}}, Z_{\text{AA}}\right\}.$$

In the time variable $s = (t - t_{crit})/n$, consider the reduced flow on the critical manifold $\bar{w}_{1j} + \bar{w}_{2j} = 1$ given by

$$\dot{\bar{w}}_{1q} = -\frac{1}{4}\bar{w}'_{1q} \sum_j Z_{jq}\bar{w}'_{1j} n(S_{1j} - S_{2j})(\bar{w}),$$

where S_{qj} is defined in Lemma B.7, so that

$$n(S_{1j} - S_{2j})(\bar{w}) = (\Sigma_{11}^{(j)} - \Sigma_{12}^{(j)})\bar{w}_{1j} + (\Sigma_{12}^{(j)} - \Sigma_{22}^{(j)})\bar{w}_{2j}.$$

Suppose the expert hierarchy of Definition B.1 holds and consider the suboptimal stationary point $\bar{w}^\circ_{1\mathrm{G}} = \bar{w}^\circ_{1\mathrm{A}} = 1$, $\bar{w}^\circ_{2\mathrm{G}} = \bar{w}^\circ_{2\mathrm{A}} = 0$. The corner selecting the other expert in both contexts is covered by exchanging the expert indices. There exist $\delta \in (0, 1/2)$, $r_0 \in (0, 1)$, $p > 0$ and $\gamma > 0$, depending only on Z and $\{\mathbf{\Sigma}^{(j)}\}_{j\in\{\mathrm{G,A}\}}$, such that the following holds. Define

$$R_\delta = \{0 < \bar{w}_{2\mathrm{G}}, \bar{w}_{2\mathrm{A}} < \delta\}$$

$$W(\bar{w}) = \frac{1}{\bar{w}_{2\mathrm{A}}} \max\left\{1, \left(\frac{\bar{w}_{2\mathrm{G}}}{r_0 \bar{w}_{2\mathrm{A}}}\right)^p\right\}.$$

For every initialization $\bar{w}(0) \in R_\delta$, the first exit time satisfies

$$\inf\{s \geq 0 : \bar{w}(s) \notin R_\delta\} \leq \frac{W(\bar{w}(0))}{\gamma} < \infty.$$

Proof. Note that $\bar{w}'_{1q} = \bar{w}'_{2q}$. Also

$$n(S_{1\mathrm{G}} - S_{2\mathrm{G}})(\bar{w}) \leq 0, \qquad n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}^\circ) = \Sigma_{11}^{(\mathrm{A})} - \Sigma_{12}^{(\mathrm{A})} > 0.$$

Thus, for sufficiently small $\delta > 0$,

$$n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}) \geq \frac{1}{2} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}^\circ) > 0$$

for all $\bar{w} \in R_\delta$. Choose $p > 0$ such that

$$pZ_{\mathrm{GG}} - (p+1)Z_{\mathrm{GA}} > 0,$$
$$(p+1)Z_{\mathrm{AA}} - pZ_{\mathrm{GA}} > 0.$$

Such a p exists: since $Z_{\mathrm{AA}} > 0$ and $Z_{\mathrm{GA}} < Z_{\mathrm{AA}}$, the second inequality holds for every $p > 0$, and since $Z_{\mathrm{GA}} < Z_{\mathrm{GG}}$, the first holds for every $p > \max\{0, Z_{\mathrm{GA}}\}/(Z_{\mathrm{GG}} - Z_{\mathrm{GA}})$. Choose δ sufficiently small that

$$p(1 - \bar{w}_{2\mathrm{G}})Z_{\mathrm{GG}} - (p+1)(1 - \bar{w}_{2\mathrm{A}})Z_{\mathrm{GA}} > 0,$$
$$(p+1)(1 - \bar{w}_{2\mathrm{A}})Z_{\mathrm{AA}} - p(1 - \bar{w}_{2\mathrm{G}})Z_{\mathrm{GA}} > 0,$$

uniformly on R_δ.

We show $\dot{W} \leq -\gamma < 0$ for $\bar{w} \in R_\delta$.

If $\bar{w}_{2\mathrm{G}} \leq r_0 \bar{w}_{2\mathrm{A}}$, then $W(\bar{w}) = 1/\bar{w}_{2\mathrm{A}}$ and

$$\begin{aligned}\dot{W} &= -\frac{\dot{\bar{w}}_{2\mathrm{A}}}{\bar{w}_{2\mathrm{A}}^2} \\ &= -\frac{(1 - \bar{w}_{2\mathrm{A}})^2}{4}\left[Z_{\mathrm{GA}}\frac{\bar{w}'_{1\mathrm{G}}}{\bar{w}'_{1\mathrm{A}}} n(S_{1\mathrm{G}} - S_{2\mathrm{G}})(\bar{w}) + Z_{\mathrm{AA}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w})\right].\end{aligned}$$

Moreover,

$$\frac{\bar{w}'_{1\mathrm{G}}}{\bar{w}'_{1\mathrm{A}}} = \frac{\bar{w}_{2\mathrm{G}}(1 - \bar{w}_{2\mathrm{G}})}{\bar{w}_{2\mathrm{A}}(1 - \bar{w}_{2\mathrm{A}})} \leq \frac{r_0}{1 - \delta}.$$

Since $n(S_{1\mathrm{G}} - S_{2\mathrm{G}})$ is bounded, choose $r_0 \in (0, 1)$ such that

$$\max\{Z_{\mathrm{GA}}, 0\}\frac{r_0}{1 - \delta} \sup_{\bar{w}\in\overline{R_\delta}} |n(S_{1\mathrm{G}} - S_{2\mathrm{G}})(\bar{w})| \leq \frac{1}{4} Z_{\mathrm{AA}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}^\circ).$$

Then

$$\dot{W} \le -\frac{(1-\delta)^2 Z_{\mathrm{AA}} n (S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}^\circ)}{16} < 0. \tag{18}$$

Suppose next that $\bar{w}_{2\mathrm{G}} \ge r_0 \bar{w}_{2\mathrm{A}}$. For the branch $W = \bar{w}_{2\mathrm{G}}^p / (r_0^p \bar{w}_{2\mathrm{A}}^{p+1})$, the uniform bounds above give a constant $c > 0$, independent of $\bar{w} \in R_\delta$, such that

$$\begin{aligned}
4\frac{\dot{W}}{W} &= 4\left(p\frac{\dot{\bar{w}}_{2\mathrm{G}}}{\bar{w}_{2\mathrm{G}}} - (p+1)\frac{\dot{\bar{w}}_{2\mathrm{A}}}{\bar{w}_{2\mathrm{A}}}\right) \\
&= p\frac{\bar{w}'_{1\mathrm{G}}}{\bar{w}_{2\mathrm{G}}} \sum_j Z_{j\mathrm{G}} \bar{w}'_{1j} n(S_{1j} - S_{2j})(\bar{w}) - (p+1)\frac{\bar{w}'_{1\mathrm{A}}}{\bar{w}_{2\mathrm{A}}} \sum_j Z_{j\mathrm{A}} \bar{w}'_{1j} n(S_{1j} - S_{2j})(\bar{w}) \\
&= p(1-\bar{w}_{2\mathrm{G}}) \left[Z_{\mathrm{GG}} \bar{w}'_{1\mathrm{G}} n(S_{1\mathrm{G}} - S_{2\mathrm{G}})(\bar{w}) + Z_{\mathrm{GA}} \bar{w}'_{1\mathrm{A}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w})\right] \\
&\quad - (p+1)(1-\bar{w}_{2\mathrm{A}}) \left[Z_{\mathrm{GA}} \bar{w}'_{1\mathrm{G}} n(S_{1\mathrm{G}} - S_{2\mathrm{G}})(\bar{w}) + Z_{\mathrm{AA}} \bar{w}'_{1\mathrm{A}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w})\right] \\
&= \left[p(1-\bar{w}_{2\mathrm{G}}) Z_{\mathrm{GG}} - (p+1)(1-\bar{w}_{2\mathrm{A}}) Z_{\mathrm{GA}}\right] \bar{w}'_{1\mathrm{G}} n(S_{1\mathrm{G}} - S_{2\mathrm{G}})(\bar{w}) \\
&\quad + \left[p(1-\bar{w}_{2\mathrm{G}}) Z_{\mathrm{GA}} - (p+1)(1-\bar{w}_{2\mathrm{A}}) Z_{\mathrm{AA}}\right] \bar{w}'_{1\mathrm{A}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}) \\
&\le -\left[(p+1)(1-\bar{w}_{2\mathrm{A}}) Z_{\mathrm{AA}} - p(1-\bar{w}_{2\mathrm{G}}) Z_{\mathrm{GA}}\right] \bar{w}'_{1\mathrm{A}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}) \\
&\le -\frac{1}{2}\left[(p+1)(1-\bar{w}_{2\mathrm{A}}) Z_{\mathrm{AA}} - p(1-\bar{w}_{2\mathrm{G}}) Z_{\mathrm{GA}}\right] \bar{w}'_{1\mathrm{A}} n(S_{1\mathrm{A}} - S_{2\mathrm{A}})(\bar{w}^\circ) \\
&\le -4c\bar{w}'_{1\mathrm{A}}.
\end{aligned}$$

Thus

$$\begin{aligned}
\dot{W} &\le -c\bar{w}'_{1\mathrm{A}} W \\
&= -c(1-\bar{w}_{2\mathrm{A}}) \left(\frac{\bar{w}_{2\mathrm{G}}}{r_0 \bar{w}_{2\mathrm{A}}}\right)^p \\
&\le -c(1-\delta) < 0.
\end{aligned}$$

Both branch derivatives satisfy these bounds at $\bar{w}_{2\mathrm{G}} = r_0 \bar{w}_{2\mathrm{A}}$ as well. Since W is locally Lipschitz, $\dot{W} \le -\gamma$ almost everywhere along the trajectory for a common $\gamma > 0$. Apply Lemma G.1 with lower bound $C = 0$. □

Lemma D.9 (Gradient bound away from stationary points)**.** *Fix λ, suppose the expert hierarchy of Definition B.1 holds and $Z \succ 0$, and let $\sigma_\mu^2, c, \rho > 0$. Use $\mathcal{S}$ from Lemma B.7 and define*

$$R_{c,\rho} = \left\{ \bar{w} \in [0,1]^{2\times 2} : \mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \min_{\bar{w}^\circ \in \mathcal{S}} \mathcal{L}^{\mathrm{truncated}}(\bar{w}^\circ) \le \frac{c}{n}, \quad \min_{\bar{w}^\circ \in \mathcal{S}} \|\bar{w} - \bar{w}^\circ\|_{\mathrm{F}} \ge \rho \right\},$$

$$C_\Sigma = \max_{\substack{j \in \{\mathrm{G},\mathrm{A}\} \\ q \ne k_j}} \left\{ \frac{1}{\Sigma^{(j)}_{k_j k_j}}, \frac{1}{\Sigma^{(j)}_{q k_j} - \Sigma^{(j)}_{k_j k_j}}, \frac{1}{\Sigma^{(j)}_{qq} - \Sigma^{(j)}_{q k_j}} \right\}.$$

If

$$n\sigma_\mu^2 \ge \max\left\{ 16c + 4\sum_j \Sigma^{(j)}_{k_j k_j},\ \max_{i,k,j} \Sigma^{(j)}_{ik} \right\},$$

then every $\bar{w} \in R_{c,\rho}$ satisfies

$$\|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}(\bar{w})\|_2 \ge \frac{\sqrt{\lambda_{\min}(Z)}}{160 C_\Sigma} \frac{\rho}{n}.$$

Proof. The loss bound and $\min_{\bar{w}^\circ \in \mathcal{S}} \mathcal{L}^{\mathrm{truncated}}(\bar{w}^\circ) \le \mathcal{L}^{\mathrm{truncated}}(I_2)$ give

$$|\bar{w}_{1j} + \bar{w}_{2j} - 1|^2 \le \frac{4c + \sum_\ell \Sigma^{(\ell)}_{\ell\ell}}{n\sigma_\mu^2} \le \frac{1}{4}.$$

For each j, choose q with $\bar{w}_{qj} \geq \bar{w}_{kj}$, $k \neq q$, so $\bar{w}_{qj} \geq 1/4$ and $\bar{w}_{kj} \leq 3/4$. The hierarchy gives $|S_{qj} - S_{kj}| \geq 1/(2C_\Sigma n)$. If $1 - \bar{w}_{qj} \geq \bar{w}_{kj}/2$, then

$$\frac{\bar{w}_{kj}}{2C_\Sigma n} \leq \bar{w}_{kj}|S_{qj} - S_{kj}| \leq 12 \max_i |\bar{w}'_{ij} S_{ij}|.$$

Otherwise $S_{kj} \geq \Sigma^{(j)}_{kq} \bar{w}_{qj}/n \geq 1/(4C_\Sigma n)$, so $\bar{w}'_{kj} S_{kj} \geq \bar{w}_{kj}/(16 C_\Sigma n)$. Thus in both cases

$$\bar{w}_{kj} \leq 24 C_\Sigma n \max_i |\bar{w}'_{ij} S_{ij}|.$$

As a polynomial in $\bar{w}_{qj}$, $(1 - \bar{w}_{qj}) S_{qj}$ has roots

$$1, \qquad \frac{n\sigma_\mu^2 - (n\sigma_\mu^2 + \Sigma^{(j)}_{qk})\bar{w}_{kj}}{n\sigma_\mu^2 + \Sigma^{(j)}_{qq}},$$

separated by at least $\Sigma^{(j)}_{qq}/(n\sigma_\mu^2 + \Sigma^{(j)}_{qq})$. Using $\bar{w}_{qj} \geq 1/4$ and $(n\sigma_\mu^2 + \Sigma^{(j)}_{qk})/(n\sigma_\mu^2 + \Sigma^{(j)}_{qq}) \leq 2$ gives

$$\min\left\{1 - \bar{w}_{qj}, \left|\bar{w}_{qj} - \frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \Sigma^{(j)}_{qq}}\right|\right\} \leq 8 C_\Sigma n |\bar{w}'_{qj} S_{qj}| + 2\bar{w}_{kj}.$$

Setting $\bar{w}_{kj} = 0$ and choosing the nearer of these two values in each family produces a stationary point, hence

$$\rho \leq \min_{\bar{w}^\circ \in \mathcal{S}} \|\bar{w} - \bar{w}^\circ\|_{\mathrm{F}} \leq 80 C_\Sigma n \left(\sum_{i,j} (\bar{w}'_{ij} S_{ij})^2\right)^{1/2} \leq \frac{160 C_\Sigma n}{\sqrt{\lambda_{\min}(Z)}} \|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2.$$

□

Lemma D.10 (Population gradient perturbation)**.** *Assume the initial loss and covariance conditions of Lemma B.10, with its constant κ, at a point θ_0. Fix $B, c > 0$ independently of n and σ_μ^2, and assume $\|\theta_0\|_2 \leq B$ and $\mathcal{L}^{\mathrm{truncated}}(\bar{w}(\theta_0)) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*) \leq c/n$. Here $\bar{w}^*$ denotes the optimum, and $\mathcal{L}^{\mathrm{truncated}}$ is composed with $\bar{w}(\theta)$ when differentiating. For a constant C_0 and $\rho = C_0/(n\sigma_\mu^2)$, set*

$$E(\theta) = \mathcal{L}^{\mathrm{truncated}}(\bar{w}(\theta)) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*) + \frac{1}{n^2\sigma_\mu^2}, \qquad E_0 = \frac{c + \kappa}{n} + \frac{1}{n^2\sigma_\mu^2},$$

and let $\mathcal{D}$ be the closed set of θ satisfying

$$\mathcal{L}^{\mathrm{truncated}}(\bar{w}(\theta)) \leq \mathcal{L}^{\mathrm{truncated}}(\bar{w}(\theta_0)) + \frac{\kappa}{2n}, \qquad \min_{\bar{w}^\circ \in \mathcal{S}} \|\bar{w}(\theta) - \bar{w}^\circ\|_{\mathrm{F}} \geq \rho,$$
$$\|\theta\|_2 \leq B + 1 + 4C_0 \log \frac{E_0}{E(\theta)}.$$

There is such a C_0, independent of n and σ_μ^2, for which the following hold for sufficiently small fixed $\sigma_\mu^2 > 0$ and then sufficiently large n.

(a) $\mathcal{D}$ is contained in the ball of radius $r = B + 1 + 4C_0 \log(1 + (c + \kappa)n\sigma_\mu^2) = \mathcal{O}(\log n)$, and uniformly over $\theta \in \mathcal{D}$,

$$\|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2 \geq \frac{E(\theta)}{C_0},$$
$$\sum_{q,j} \bar{w}'_{qj} \leq C_0 n \|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2,$$
$$\|\nabla_\theta(\mathcal{L} - \mathcal{L}^{\mathrm{truncated}})\|_2 \leq \left[C_0 \sigma_\mu^2 n E(\theta)(1 + \|\theta\|_2)^2 + \widetilde{\mathcal{O}}(n^{-1})\right] \|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2$$
$$\leq \frac{1}{4} \|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2.$$

(b) Let $\theta(t)$, $t \geq t_{crit}$, be absolutely continuous with $\theta(t_{crit}) = \theta_0$ and

$$\dot{\theta} = -\nabla_\theta \mathcal{L}(\theta) - \zeta(t),$$
$$\|\zeta(t)\|_2 \leq \min\left\{\frac{1}{4}, \frac{1}{\log n}\right\} \|\nabla_\theta \mathcal{L}^{\text{truncated}}(\theta(t))\|_2 \quad \textit{whenever } \theta(t) \in \mathcal{D},$$

Let t_ρ be the first time $t \geq t_{crit}$ with $\min_{\bar{w}^\circ \in \mathcal{S}} \|\bar{w}(\theta(t)) - \bar{w}^\circ\|_{\mathrm{F}} \leq \rho$. Then $t_\rho \leq t_{crit} + 2C_0^2 n^2 \sigma_\mu^2$, $\theta(t) \in \mathcal{D}$ for $t_{crit} \leq t < t_\rho$, and for $t_{crit} \leq t \leq t_\rho$,

$$\frac{d}{dt}\mathcal{L}^{\text{truncated}}(\bar{w}(\theta(t))) \leq -\frac{1}{2}\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2^2, \qquad \|\dot{\theta}\|_2 \leq \frac{3}{2}\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2,$$
$$\|\theta(t)\|_2 \leq B + 3C_0 \log \frac{E(\theta_0)}{E(\theta(t))} \leq B + 3C_0 \log(1 + cn\sigma_\mu^2),$$
$$\|\theta_q^\perp(t) - \theta_q^\perp(t_{crit})\|_2 \leq \mathcal{O}(\sigma_\mu^2) + \widetilde{\mathcal{O}}(n^{-1}) + \frac{2C_0}{\log n} \log(1 + cn\sigma_\mu^2).$$

The full population flow is the case $\zeta \equiv 0$.

Proof. Part (a). On $\mathcal{D}$, the truncated loss is $\mathcal{O}(n^{-1})$, lies at least $\kappa/(2n)$ below the loss of every point in (15), and has excess at most $(c+\kappa)/n$. In particular, $1/(n^2\sigma_\mu^2) \leq E(\theta) \leq E_0$, which gives the radius r. The excess-loss bound (20) of Lemma D.11 gives

$$\max_j \left[\sum_q \bar{w}'_{qj} + \left|\sum_q \bar{w}_{qj} - 1\right|\right] \lesssim n[\mathcal{L}^{\text{truncated}}(\bar{w}) - \mathcal{L}^{\text{truncated}}(\bar{w}^*)] + \frac{1}{n\sigma_\mu^2} = nE(\theta).$$

Minimizing the quadratic loss over the other gates shows that the loss gap below each wrong selection bounds $\bar{w}_{k_j j}$ and $1 - \bar{w}_{qj}$ away from zero for $q \neq k_j$. The same two-case calculation as in Lemma D.9 then gives

$$\frac{\bar{w}_{qj}}{n} + \sigma_\mu^2 \left(\sum_i \bar{w}_{ij} - 1\right)^2 \lesssim \|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2 + \frac{1}{n^2\sigma_\mu^2}.$$

Expanding the quadratic loss and applying Lemma D.9 with $\rho = C_0/(n\sigma_\mu^2)$ proves the first estimate for sufficiently large C_0. Combined with the first display, it gives the second estimate after enlarging C_0. These two estimates depend only on $\bar{w}$, so C_0 is fixed before σ_μ^2 is chosen.

Taylor-expand the conditional gradient through cubic order in the centered empirical moments. Their signed cubic moments and fourth absolute moments are $\mathcal{O}(n^{-2})$, so retaining the sigmoid derivatives gives

$$\|\nabla_\theta(\mathcal{L} - \mathcal{L}^{\text{truncated}})\|_2 \lesssim \sum_{q,j} \bar{w}'_{qj} \left[\frac{\sigma_\mu^2(1 + \|\theta\|_2)^2}{n}\left(\sum_i \bar{w}'_{ij} + \left|\sum_i \bar{w}_{ij} - 1\right|\right) + \frac{(1 + \|\theta\|_2)^4}{n^2}\right].$$

This bound is uniform for $\|\theta\|_2 \leq r$ and $0 < \sigma_\mu^2 \leq 1$, using the moment bounds and truncation from Lemma D.4. The first display and the second estimate turn it into the first inequality of the third estimate. On $\mathcal{D}$, $1 + \|\theta\|_2 \leq 2 + B + 4C_0 \log(E_0/E(\theta))$, so with $u = E(\theta)/E_0 \in (0, 1]$ and $nE_0 = \mathcal{O}(1)$,

$$C_0\sigma_\mu^2 nE(\theta)(1 + \|\theta\|_2)^2 \leq C_0\sigma_\mu^2 nE_0 \sup_{0<u\leq 1} u\left(2 + B + 4C_0 \log\frac{1}{u}\right)^2 = \mathcal{O}(\sigma_\mu^2).$$

Choose σ_μ^2 small and then n large to make the bracket at most $1/4$.

Part (b). Write $\dot{\theta} = -\nabla_\theta \mathcal{L}^{\text{truncated}} - p$ with $p = \nabla_\theta(\mathcal{L} - \mathcal{L}^{\text{truncated}}) + \zeta$. While $\theta(t) \in \mathcal{D}$, part (a) gives $\|p\|_2 \leq \frac{1}{2}\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2$, hence the displayed descent and speed bounds, and, by the first estimate,

$$\frac{d}{dt}\log E(\theta(t)) \leq -\frac{\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2^2}{2E(\theta(t))} \leq -\frac{\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2}{2C_0}.$$

Integrating,

$$\int_{t_{\text{crit}}}^{t} \|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2 \, ds \le 2C_0 \log \frac{E(\theta_0)}{E(\theta(t))}, \qquad \int_{t_{\text{crit}}}^{t} \|\dot\theta(s)\|_2 \, ds \le 3C_0 \log \frac{E(\theta_0)}{E(\theta(t))}. \tag{19}$$

We first show that the trajectory cannot leave $\mathcal{D}$ before t_ρ. Let T be the supremum of times $t \le t_\rho$ with $\theta([t_{\text{crit}}, t]) \subset \mathcal{D}$, and suppose $T < t_\rho$. Since $\mathcal{D}$ is closed, (19) holds up to T, so

$$\mathcal{L}^{\text{truncated}}(\bar w(\theta(T))) \le \mathcal{L}^{\text{truncated}}(\bar w(\theta_0)) < \mathcal{L}^{\text{truncated}}(\bar w(\theta_0)) + \frac{\kappa}{2n},$$

$$\|\theta(T)\|_2 \le B + 3C_0 \log \frac{E(\theta_0)}{E(\theta(T))} < B + 1 + 4C_0 \log \frac{E_0}{E(\theta(T))},$$

using $E(\theta_0) \le E_0$, while the distance to $\mathcal{S}$ exceeds ρ because $T < t_\rho$. All three conditions defining $\mathcal{D}$ hold strictly at $\theta(T)$, so by continuity they hold on a neighborhood of T, contradicting the definition of T.

For finite entry, set $W = -1/E$. On $[t_{\text{crit}}, t_\rho)$,

$$\frac{d}{dt} W(\theta(t)) = \frac{\dot E}{E^2} \le -\frac{\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2^2}{2E^2} \le -\frac{1}{2C_0^2}, \qquad -n^2\sigma_\mu^2 \le W < 0,$$

so, as in Lemma G.1, $t_\rho \le t_{\text{crit}} + 2C_0^2 n^2 \sigma_\mu^2$. The parameter bound follows from (19), $\|\theta_0\|_2 \le B$, and $n^2\sigma_\mu^2 E(\theta_0) \le 1 + cn\sigma_\mu^2 \le n^2\sigma_\mu^2 E_0$.

Since the truncated gradient has zero perpendicular component, $\|\dot\theta_q^\perp\|_2 \le \|\nabla_\theta(\mathcal{L} - \mathcal{L}^{\text{truncated}})\|_2 + \|\zeta\|_2$. Along the trajectory, E decreases and $\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2 \, ds \le -2C_0 \, dE/E$. Substituting the third estimate of part (a) and the parameter bound of $\mathcal{D}$,

$$\begin{aligned}\int_{t_{\text{crit}}}^{t} \|\nabla_\theta(\mathcal{L} - \mathcal{L}^{\text{truncated}})\|_2 \, ds &\le 2C_0^2\sigma_\mu^2 n E_0 \int_0^1 \left(2 + B + 4C_0 \log \frac{1}{u}\right)^2 du \\ &\quad + \widetilde{\mathcal{O}}(n^{-1}) \log(1 + cn\sigma_\mu^2) \\ &= \mathcal{O}(\sigma_\mu^2) + \widetilde{\mathcal{O}}(n^{-1}).\end{aligned}$$

Finally, $\int_{t_{\text{crit}}}^{t} \|\zeta\|_2 \, ds \le \frac{2C_0}{\log n} \log(1 + cn\sigma_\mu^2)$ by (19). □

D.3.4 Proof of Lemma B.10

Proof. Use $\mathcal{S}$ from Lemma B.7 and C_Σ from Lemma D.9. Until entry within distance $1/n$ of $\mathcal{S}$,

$$\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2 \ge \frac{\sqrt{\lambda_{\min}(Z)}}{160 C_\Sigma n^2},$$

$$\frac{d}{dt} \mathcal{L}^{\text{truncated}}(\bar w(t)) = -\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2^2 \le -\frac{\lambda_{\min}(Z)}{(160C_\Sigma)^2 n^4}.$$

Since $\mathcal{L}^{\text{truncated}} \ge 0$, entry occurs by

$$t_\theta \le t_{\text{crit}} + \frac{(160C_\Sigma)^2 n^4}{\lambda_{\min}(Z)} \mathcal{L}^{\text{truncated}}(\bar w(t_{\text{crit}})) = t_{\text{crit}} + \mathcal{O}(n^3).$$

For $\rho = \mathcal{O}(n^{-1})$, ρ-neighborhoods of stationary points with both gates 0 or 1 in a context have

$$\mathcal{L}^{\text{truncated}}(\bar w) \ge \frac{\sigma_\mu^2}{4} - \mathcal{O}(n^{-1}) > \mathcal{L}^{\text{truncated}}(\bar w(t_{\text{crit}})).$$

For every $\bar w^\circ$ in (15) and every $\bar w$ with $\|\bar w - \bar w^\circ\|_{\text{F}} = \mathcal{O}(n^{-1})$, the quadratic expansion gives

$$\begin{aligned}\mathcal{L}^{\text{truncated}}(\bar w) &\ge \mathcal{L}^{\text{truncated}}(\bar w^\circ) - \mathcal{O}(n^{-2}) \\ &> \mathcal{L}^{\text{truncated}}(\bar w(t_{\text{crit}})).\end{aligned}$$

Thus, by Lemma B.7, loss descent permits entry only near stationary points with

$$\bar{w}^{\circ}_{\cdot j} \in \{e_{k_j}, \bar{w}^{*}_{\cdot j}\}, \qquad \|e_{k_j} - \bar{w}^{*}_{\cdot j}\|_2 = \frac{\Sigma^{(j)}_{k_j k_j}}{n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j}}.$$

Adding the entry radius $1/n$ bounds $\|\bar{w}_{\cdot j}(t_\theta) - e_{k_j}\|_2$ as claimed.

For the full flow, take B and c as the assumed bounds on $\|\theta(t_{\text{crit}})\|_2$ and $n\mathcal{L}^{\text{truncated}}(\bar{w}(t_{\text{crit}}))$, and apply part (b) of Lemma D.10 with $\theta_0 = \theta(t_{\text{crit}})$ and $\zeta \equiv 0$. Set $t_\theta = t_\rho \le t_{\text{crit}} + 2C_0^2 n^2 \sigma_\mu^2$. It gives $\mathcal{L}^{\text{truncated}}(\bar{w}(t_\theta)) \le \mathcal{L}^{\text{truncated}}(\bar{w}(t_{\text{crit}}))$ and $\|\theta(t)\|_2 \le B + 3C_0 \log(1 + cn\sigma_\mu^2) = \mathcal{O}(\log n)$ for $t_{\text{crit}} \le t \le t_\theta$. Since $\rho = \mathcal{O}(n^{-1})$, the exclusion above applies to a point of $\mathcal{S}$ within distance ρ of $\bar{w}(t_\theta)$, and the triangle inequality gives the full-flow gate bound. □

D.3.5 Proof of Lemma B.11

Proof. Let $\mathcal{D}$ and $r = \mathcal{O}(\log n)$ be as in Lemma D.10 with $\theta_0 = \theta(t_{\text{crit}})$. The radius r depends only on B, c, κ, and C_0, so the concentration events below do not depend on the data-dependent initial point. Write $m^{(j)} = \#\{b : \tau_b = j\}$ and $\mathcal{L}^{(j)} = \frac{1}{2}\mathbb{E}_j[(\widehat{\mu} - \mu)^2]$, with $\hat{\mathcal{L}}^{(j)}_{m^{(j)}}$ its empirical average over the tasks of family j, so that

$$\mathcal{L} = \frac{1}{2}\sum_j \mathcal{L}^{(j)}, \qquad \hat{\mathcal{L}}_m = \sum_j \frac{m^{(j)}}{m}\hat{\mathcal{L}}^{(j)}_{m^{(j)}}.$$

Concentration. Hoeffding's inequality gives, with failure probability at most $1/(3n)$,

$$\max_j \left|\frac{m^{(j)}}{m} - \frac{1}{2}\right| = \widetilde{\mathcal{O}}(m^{-1/2}),$$

so $m^{(j)} \ge m/4$ for large n. Conditioning on the family labels, apply Lemma D.4 and Lemma D.6 with radius r and $\beta = \min\{\epsilon, 1\}/4$ to the $m^{(j)}$ tasks of each family, each with failure probability $1/(6n)$. The total failure probability is at most $1/n$, and all bounds below hold on this event, uniformly over the ball of radius r.

Relative gradient error. On $\mathcal{D}$, since $\mathbf{\Sigma}^{(j)} \succeq 0$,

$$\max_j \left|\sum_k \bar{w}_{kj} - 1\right|^2 \le \frac{4}{\sigma_\mu^2}\mathcal{L}^{\text{truncated}}(\bar{w}) \le \frac{4}{\sigma_\mu^2}\left(\mathcal{L}^{\text{truncated}}(\bar{w}(t_{\text{crit}})) + \frac{\kappa}{2n}\right) = \mathcal{O}(n^{-1}),$$

so the prefactor in Lemma D.4 is at most $2n^{-1/2+\beta}$. Summing its coordinatewise bound over the fixed number of coordinates and weighting by $m^{(j)}/m$,

$$\left\|\sum_j \frac{m^{(j)}}{m}\nabla_\theta\left(\hat{\mathcal{L}}^{(j)}_{m^{(j)}} - \mathcal{L}^{(j)}\right)\right\|_2 \le \widetilde{\mathcal{O}}\left(\frac{n^{-1/2+\beta}}{\sqrt{m}}\right)\sum_{q,j}\bar{w}'_{qj}.$$

The truncation bounds in the proof of Lemma D.4 also give

$$\left\|\sum_j \left(\frac{m^{(j)}}{m} - \frac{1}{2}\right)\nabla_\theta \mathcal{L}^{(j)}\right\|_2 \le \widetilde{\mathcal{O}}\left(\frac{n^{-1/2+\beta}}{\sqrt{m}}\right)\sum_{q,j}\bar{w}'_{qj}.$$

By the second estimate of Lemma D.10(a), uniformly over $\theta \in \mathcal{D}$,

$$\frac{\|\nabla_\theta(\hat{\mathcal{L}}_m - \mathcal{L})\|_2}{\|\nabla_\theta \mathcal{L}^{\text{truncated}}\|_2} \le C_0 n\,\widetilde{\mathcal{O}}\left(\frac{n^{-1/2+\beta}}{\sqrt{m}}\right) = \widetilde{\mathcal{O}}\left(\frac{n^{1/2+\beta}}{\sqrt{m}}\right) \le \min\left\{\frac{1}{4}, \frac{1}{\log n}\right\}.$$

For the last inequality, write the preceding bound as $C(\log n)^a n^{1/2+\beta}/\sqrt{m}$ with C and a fixed. Since $\beta = \min\{\epsilon, 1\}/4$ and $m \ge n^{1+\epsilon}$, it is at most $C(\log n)^a n^{-\min\{\epsilon,1\}/4}$, hence at most $\min\{1/4, 1/\log n\}$ for sufficiently large n.

Entry, parameter bound, and gate accuracy. The empirical flow is $\dot\theta = -\nabla_\theta \mathcal{L} - \zeta(t)$ with $\zeta(t) = \nabla_\theta(\hat{\mathcal{L}}_m - \mathcal{L})(\theta(t))$, which satisfies $\|\zeta(t)\|_2 \le \min\{1/4, 1/\log n\}\|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2$ whenever $\theta(t) \in \mathcal{D}$. Part (b) of Lemma D.10 applies. Combined with part (a), it gives, before entry,

$$\|\nabla_\theta(\hat{\mathcal{L}}_m - \mathcal{L}^{\mathrm{truncated}})\|_2 \le \frac{1}{2}\|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2,$$
$$\frac{d}{dt}\mathcal{L}^{\mathrm{truncated}} \le -\frac{1}{2}\|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2^2,$$
$$\|\dot\theta\|_2 \le \frac{3}{2}\|\nabla_\theta \mathcal{L}^{\mathrm{truncated}}\|_2.$$

Set $t_\theta = t_\rho \le t_{\mathrm{crit}} + 2C_0^2 n^2 \sigma_\mu^2$. Part (b) gives loss descent and the stated parameter bound on $[t_{\mathrm{crit}}, t_\theta]$, and

$$\|\theta_q^\perp(t)\|_2 \le B + \mathcal{O}(\sigma_\mu^2) + \widetilde{\mathcal{O}}(n^{-1}) + \frac{2C_0}{\log n}\log(1 + cn\sigma_\mu^2) = \mathcal{O}(1).$$

Since $\rho = \mathcal{O}(n^{-1})$ and $\mathcal{L}^{\mathrm{truncated}}(\bar w(t_\theta)) \le \mathcal{L}^{\mathrm{truncated}}(\bar w(t_{\mathrm{crit}}))$, the exclusion in the proof of Lemma B.10 applies to a point of $\mathcal{S}$ within distance ρ of $\bar w(t_\theta)$. Its columns are e_{k_j} or $\bar w^*_{\cdot j}$, and the triangle inequality gives the gate bound. Since $\|e_{k_j} - \bar w^*_{\cdot j}\|_2 = \Sigma^{(j)}_{k_j k_j}/(n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j})$,

$$\|\bar w_{\cdot j}(t_\theta) - e_{k_j}\|_1 \le \sqrt{2}\left(\rho + \frac{2\Sigma^{(j)}_{k_j k_j}}{n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j}}\right) = \mathcal{O}(n^{-1}).$$

Risk. By Lemma F.6, $\mathbb{E}_j[(g_{k_j} - \mu)^2] = \mathcal{O}(n^{-1})$. For the 1-sub-Gaussian centered observations, $\mathbb{E}_j[\max_{\ell\in[n]}|X_\ell - \mu|^4] = \mathcal{O}(\log^2 n)$. Apply Lemma 3.5 with $i = k_j$ and $\max_k \|\theta_k(t_\theta)\|_2 = \mathcal{O}(\log n)$:

$$\sqrt{\mathbb{E}_j[(\widehat\mu - \mu)^2]} \le \mathcal{O}(n^{-1/2}) + \mathcal{O}(n^{-1})\left(\sigma_\mu^2 + \mathcal{O}(\log n)\right)^{1/2} = \mathcal{O}(n^{-1/2}).$$

Hence $\mathcal{L}^{(j)}(\theta(t_\theta)) = \mathcal{O}(n^{-1})$ for each j, and $\mathcal{L}(\theta(t_\theta)) = \mathcal{O}(n^{-1})$. Moreover, $|\sum_k \bar w_{kj}(t_\theta) - 1| \le \|\bar w_{\cdot j}(t_\theta) - e_{k_j}\|_1 = \mathcal{O}(n^{-1})$ with a fixed implied constant, and $\|\theta(t_\theta)\|_2 \le r$. Hence Lemma D.6 and $m^{(j)} \ge m/4$ give

$$\left|\hat{\mathcal{L}}^{(j)}_{m^{(j)}}(\theta(t_\theta)) - \mathcal{L}^{(j)}(\theta(t_\theta))\right| = \widetilde{\mathcal{O}}\left(\frac{n^{-1+\beta}}{\sqrt{m}} + \frac{n^{-1+2\beta}}{m}\right) = \mathcal{O}(n^{-3/2}),$$

so $\hat{\mathcal{L}}_m(\theta(t_\theta)) \le \max_j \hat{\mathcal{L}}^{(j)}_{m^{(j)}}(\theta(t_\theta)) = \mathcal{O}(n^{-1})$. All concentration bounds are uniform over the ball of radius r, so they apply at the data-dependent time t_θ. □

Lemma D.11 (Loss barrier near the optimal stationary point)**.** *Fix λ and let $\sigma_\mu^2 > 0$. Suppose the expert hierarchy of Definition B.1 holds and use Lemma B.7 and the truncated loss from Definition B.6. For each family j, choose the positive threshold*

$$\delta_j = \frac{1}{16n^2}\min\left\{\sigma_\mu^2\left(\Sigma^{(j)}_{k_j k_j}\right)^2,\ \bar w^*_{k_j j}\Sigma^{(j)}_{k_j k_j}\min_{k\ne k_j}\left(\Sigma^{(j)}_{k k_j} - \Sigma^{(j)}_{k_j k_j}\right)\right\},$$

and let $\delta = \min_{j\in\{\mathrm{G},\mathrm{A}\}}\delta_j$.

Define the region

$$N_\delta = \left\{\bar w \in [0,1]^{2\times 2} : \mathcal{L}^{\mathrm{truncated}}(\bar w) - \mathcal{L}^{\mathrm{truncated}}(\bar w^*) \le \delta\right\}.$$

Every point in N_δ satisfies, for each j,

$$\|\bar w_{\cdot j} - \bar w^*_{\cdot j}\|_2 \le \frac{\Sigma^{(j)}_{k_j k_j}}{n}.$$

Moreover, for every $\bar w \in [0,1]^{2\times 2}$ and all sufficiently large n,

$$\max_j\left[\sum_q \bar w'_{qj} + \left|\sum_q \bar w_{qj} - 1\right|\right] \lesssim n\left[\mathcal{L}^{\mathrm{truncated}}(\bar w) - \mathcal{L}^{\mathrm{truncated}}(\bar w^*)\right] + \frac{1}{n\sigma_\mu^2}, \tag{20}$$

with an implied constant depending only on $\mathbf{\Sigma}^{(\mathrm{G})}$ and $\mathbf{\Sigma}^{(\mathrm{A})}$. The truncated gradient flow

$$\dot{\bar{w}}_{qk} = -\frac{1}{2}\bar{w}'_{qk}\sum_j Z_{jk}\bar{w}'_{qj}S_{qj}(\bar{w}), \qquad Z_{jk} = \langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k \rangle ,$$

cannot exit N_δ once it enters it.

For the full population gradient flow $\dot{\theta} = -\nabla_\theta \mathcal{L}$ at fixed λ, the same region cannot be exited after time t_{crit} provided

$$\mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit})) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*) \le \delta/4$$

and, for every θ whose gates $\bar{w}$ lie in N_δ,

$$\left|\mathcal{L}(\theta) - \mathcal{L}(\theta(t_{crit})) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}) + \mathcal{L}^{\mathrm{truncated}}(\bar{w}(t_{crit}))\right| \le \delta/4.$$

The same statement holds for the empirical gradient flow with $\mathcal{L}$ replaced by $\hat{\mathcal{L}}_m$ in both the flow and this bound.

Proof. Since $\mathcal{L}^{\mathrm{truncated}}$ is quadratic and $S_{k_j j}(\bar{w}^*) = 0$, its expansion at $\bar{w}^*$ is exact:

$$\begin{aligned} 2\big(\mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*)\big) = \sum_j \Bigg[& \frac{\bar{w}^*_{k_j j}}{n} \sum_{k \neq k_j} \big(\Sigma^{(j)}_{k k_j} - \Sigma^{(j)}_{k_j k_j}\big)\bar{w}_{kj} \\ & + \frac{\sigma_\mu^2}{2}\left(\sum_{k=1}^{2}(\bar{w}_{kj} - \bar{w}^*_{kj})\right)^2 \\ & + \frac{1}{2n}(\bar{w}_{\cdot j} - \bar{w}^*_{\cdot j})^\top \mathbf{\Sigma}^{(j)}(\bar{w}_{\cdot j} - \bar{w}^*_{\cdot j})\Bigg]. \end{aligned}$$

Every term on the right is nonnegative by the expert hierarchy and positive semidefiniteness of the covariance matrices.

We first deduce (20). Bound each group of terms in the expansion by $2[\mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*)]$. Since $\bar{w}^*_{k_j j} \ge 1/2$ for large n and the hierarchy gaps are positive constants, the first group gives $\sum_{k \neq k_j} \bar{w}_{kj} \lesssim n[\mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*)]$. The second gives

$$\begin{aligned} \left|\sum_k (\bar{w}_{kj} - \bar{w}^*_{kj})\right| &\le 2\sqrt{[\mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*)]/\sigma_\mu^2} \\ &= 2\sqrt{n[\mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*)] \cdot \tfrac{1}{n\sigma_\mu^2}} \\ &\le n[\mathcal{L}^{\mathrm{truncated}}(\bar{w}) - \mathcal{L}^{\mathrm{truncated}}(\bar{w}^*)] + \frac{1}{n\sigma_\mu^2} \end{aligned}$$

by the arithmetic–geometric mean inequality. As $\sum_k \bar{w}^*_{kj} = \bar{w}^*_{k_j j}$ and $1 - \bar{w}^*_{k_j j} = \Sigma^{(j)}_{k_j k_j}/(n\sigma_\mu^2 + \Sigma^{(j)}_{k_j k_j}) \lesssim (n\sigma_\mu^2)^{-1}$, the triangle inequality bounds $|\sum_k \bar{w}_{kj} - 1|$ by the right-hand side of (20). Combining the same two bounds controls $|\bar{w}_{k_j j} - \bar{w}^*_{k_j j}|$ and hence $1 - \bar{w}_{k_j j}$, so $\bar{w}'_{k_j j} \le 1 - \bar{w}_{k_j j}$ and $\bar{w}'_{qj} \le \bar{w}_{qj}$ for $q \neq k_j$ are bounded by the same quantity. This proves (20).

For each family j, the bound $\delta \le \delta_j$ gives, on N_δ,

$$\sum_{k \neq k_j} \bar{w}_{kj} \le \frac{2n\delta}{\bar{w}^*_{k_j j} \min_{k \neq k_j}(\Sigma^{(j)}_{k k_j} - \Sigma^{(j)}_{k_j k_j})} \le \frac{\Sigma^{(j)}_{k_j k_j}}{8n},$$

and

$$\left|\sum_{k=1}^{2}(\bar{w}_{kj} - \bar{w}^*_{kj})\right| \le \sqrt{\frac{4\delta}{\sigma_\mu^2}} \le \frac{\Sigma^{(j)}_{k_j k_j}}{2n}.$$

Consequently,

$$\begin{aligned}\|\bar{w}_{\cdot j} - \bar{w}^*_{\cdot j}\|_2 &\le |\bar{w}_{k_j j} - \bar{w}^*_{k_j j}| + \sum_{k \neq k_j} \bar{w}_{kj} \\ &\le \left| \sum_{k=1}^{2} (\bar{w}_{kj} - \bar{w}^*_{kj}) \right| + 2 \sum_{k \neq k_j} \bar{w}_{kj} \\ &\le \sqrt{\frac{4\delta}{\sigma_\mu^2}} + \frac{4n\delta}{\bar{w}^*_{k_j j} \min_{k \neq k_j} (\Sigma^{(j)}_{k k_j} - \Sigma^{(j)}_{k_j k_j})} \le \frac{3\Sigma^{(j)}_{k_j k_j}}{4n}.\end{aligned}$$

The truncated flow preserves $[0,1]^{2\times 2}$ because $\bar{w}'_{qk} = 0$ on each coordinate boundary. Moreover, $\frac{\partial \mathcal{L}^{\text{truncated}}}{\partial \bar{w}_{qj}} = S_{qj}/2$, so

$$\frac{d}{dt}\mathcal{L}^{\text{truncated}}(\bar{w}(t)) = -\frac{1}{4}\sum_{q=1}^{2} \left\| \sum_j S_{qj}(\bar{w}) \bar{w}'_{qj} \bar{\mathbf{z}}_j \right\|_2^2 \le 0.$$

Since λ is fixed, $\mathcal{L}^{\text{truncated}}(\bar{w}^*)$ and δ are constant. Hence the truncated flow cannot exit N_δ.

For the full flow, a first exit would require a time t with $\mathcal{L}^{\text{truncated}}(\bar{w}(t)) - \mathcal{L}^{\text{truncated}}(\bar{w}^*) = \delta$. The assumed bound would then imply

$$\mathcal{L}(\theta(t)) - \mathcal{L}(\theta(t_{\text{crit}})) \ge \mathcal{L}^{\text{truncated}}(\bar{w}(t)) - \mathcal{L}^{\text{truncated}}(\bar{w}(t_{\text{crit}})) - \delta/4 \ge \delta/2 > 0,$$

contradicting $d\mathcal{L}/dt = -\|\nabla_\theta \mathcal{L}\|_2^2 \le 0$. Replacing $\mathcal{L}$ by $\hat{\mathcal{L}}_m$ gives the empirical statement. □

D.3.6 Proof of Corollary B.12

Proof of Corollary B.12. Since g_1 is the sample mean at $\lambda_1 = 0$, integrating the Gaussian gradient expansion in Lemma F.1 gives, uniformly for $|\lambda_1| \le 2$,

$$\mathbb{E}_{\mathrm{G}}[(g_1 - \mu)^2] = \frac{\cosh(\lambda_1^2) + \lambda_1^2 \sinh(\lambda_1^2)}{n} + \mathcal{O}(n^{-2}).$$

At t_λ, Lemma F.1, Lemma F.2, and Lemma F.3 give

$$\mathbb{E}_{\mathrm{G}}[(g_1 - \mu)^2] = \frac{1 + o(1)}{n}, \qquad \mathbb{E}_{\mathrm{A}}[(g_2 - \mu)^2] = \begin{cases} \mathcal{O}((n|\lambda_2|)^{-1}), & \text{Uniform,} \\ \mathcal{O}(\sigma_n^2/n), & \text{Gaussian mixture,} \end{cases}$$

since $|\lambda_1(t_\lambda)| = o(1)$ and $e^{-4\sqrt{1-\sigma_n^2}|\lambda_2|} = \sigma_n^2$ for the mixture. The sample-maximum fourth moment is $\mathcal{O}(1)$ for Uniform data and $\mathcal{O}((\log n)^2)$ for Gaussian and Gaussian-mixture data. Apply Lemma 3.5 with $i = 1$ for $j = \mathrm{G}$ and $i = 2$ for $j = \mathrm{A}$, and $\|\bar{w}_{\cdot j} - e_i\|_1 = \mathcal{O}(n^{-1})$. Its proof gives $\|\widehat{\mu} - g_1\|_{L^2(\mathbb{P}_{\mathrm{G}})} = \mathcal{O}(\sqrt{\log n}/n) = o(n^{-1/2})$, so the triangle inequality in both directions preserves the Gaussian leading constant. In the Gaussian-mixture case, this discrepancy is of lower order than the expert risk, since

$$\frac{n^{-2}\log n}{\sigma_n^2/n} \le n^{-\epsilon}\log n = o(1).$$

For the bounds at fixed μ, the translation identities and the gate bound give

$$
\begin{aligned}
w_k(\mathbf{X}+\mu\mathbf{1}) &= w_k(\mathbf{X}), \qquad g_k(\mathbf{X}+\mu\mathbf{1}) = g_k(\mathbf{X})+\mu,\\
\widehat{\mu}(\mathbf{X}+\mu\mathbf{1})-\mu &= \widehat{\mu}(\mathbf{X}) + \mu\left(\sum_k w_k(\mathbf{X})-1\right),\\
\mathbb{E}_{\mathbf{X}\sim P_{0,j}^{\otimes n}}\left[\left(\sum_k w_k(\mathbf{X})-1\right)^2\right] &\le \mathbb{E}_j\left[\|w(\mathbf{X})-e_i\|_1^2\right]\\
&\le C\|\bar{w}_{\cdot j}-e_i\|_1^2 = \mathcal{O}(n^{-2}),\\
R_{j,n}(\widehat{\mu}) &= \mathbb{E}_{\mathbf{X}\sim P_{0,j}^{\otimes n}}[\widehat{\mu}(\mathbf{X})^2]\\
&\quad + \sigma_\mu^2 \mathbb{E}_{\mathbf{X}\sim P_{0,j}^{\otimes n}}\left[\left(\sum_k w_k(\mathbf{X})-1\right)^2\right],\\
\mathbb{E}_{\mathbf{X}\sim P_{\mu,j}^{\otimes n}}[(\widehat{\mu}(\mathbf{X})-\mu)^2] &\le 2\mathbb{E}_{\mathbf{X}\sim P_{0,j}^{\otimes n}}[\widehat{\mu}(\mathbf{X})^2]\\
&\quad + 2\mu^2 \mathbb{E}_{\mathbf{X}\sim P_{0,j}^{\otimes n}}\left[\left(\sum_k w_k(\mathbf{X})-1\right)^2\right]\\
&\le 2R_{j,n}(\widehat{\mu}) + \frac{C\mu^2}{n^2},\\
\sum_k w_k = 1 \quad &\Longrightarrow \quad \mathbb{E}_{\mathbf{X}\sim P_{\mu,j}^{\otimes n}}[(\widehat{\mu}(\mathbf{X})-\mu)^2] = R_{j,n}(\widehat{\mu}).
\end{aligned}
$$

□

D.3.7 Proof of Corollary 3.2

Proof of Corollary 3.2. The population gates $\bar{w}_{\cdot j}$ depend on the data only through $\bar{\mathbf{z}}_j$, which does not depend on the context length. Hence $\max\{\|\bar{w}_{\cdot \mathrm{G}}-e_1\|_1, \|\bar{w}_{\cdot \mathrm{A}}-e_2\|_1\} = \mathcal{O}(n^{-1})$ and $\max_q \|\theta_q\|_2 = \mathcal{O}(\log n) = \mathcal{O}(\log n')$ at the trained parameters, and Lemma 3.5 applies with n' in place of n. The hypotheses of Lemma F.2 hold at length n', since $|\lambda_2(t_\lambda)| = n^{1-\epsilon} \le n'^{1-\epsilon}$, and those of Lemma F.3 hold since $n'^{-1+\epsilon} \le n^{-1+\epsilon} \le \sigma_n^2$. The proof of Corollary B.12 with n' test observations then gives the claim, with sample-maximum fourth moment $\mathcal{O}(1)$ for Uniform data and $\mathcal{O}((\log n')^2)$ otherwise. □

E Proof of the softmax MoE part of Lemma 3.3

For the proof, write $\vartheta = \theta_1 - \theta_2$. The softmax MoE weights in (3) satisfy $w_1(\mathbf{X}) = \sigma(\langle \mathbf{z}(\mathbf{X}), \vartheta\rangle)$ and $w_2(\mathbf{X}) = 1 - w_1(\mathbf{X})$, with the same identities at $\bar{\mathbf{z}}_j$. Writing the loss as a function of ϑ, the original population flow satisfies $\dot{\vartheta} = -2\nabla_\vartheta \mathcal{L}$ and $\frac{d}{dt}(\theta_1+\theta_2) = 0$. The same identities hold for the empirical loss and for $\mathcal{L}^{\mathrm{sm}}$. Since λ is fixed in the first stage, $|\lambda_k(0)| \le 1$ and $0 < |\lambda_1(0)| < |\lambda_2(0)|$ make expert 1 better in family G and expert 2 better in family A by Lemma B.2, Lemma B.3 and Lemma B.4 for sufficiently large n, as in Section B.3. The strict covariance gaps do not depend on n for the Gaussian and uniform families; for the Gaussian mixture, (37) converges as $\sigma_n \to 0$ to $\operatorname{sech}^2(\lambda_i)\operatorname{sech}^2(\lambda_k)$, whose gaps are positive, so they are bounded below independently of n.

Proof of the softmax MoE part of Lemma 3.3. The gate and parameter bounds follow from the empirical part of Lemma E.2 with $B = 1$. Since $\theta_1 + \theta_2$ is constant and bounded at initialization, the bound on ϑ gives $\max_k \|\theta_k(t)\|_2 = \mathcal{O}(\log n)$ throughout training of the router. □

Lemma E.1 (Taylor expansion of the loss)**.** *Under the assumptions of Lemma 3.3, for every fixed $C_0 > 0$ and all sufficiently large n,*

$$|\mathcal{L}(\vartheta) - \mathcal{L}^{\mathrm{sm}}(\vartheta)| \le \frac{C(1+\|\vartheta\|_2)^4}{n^{3/2}}, \tag{21}$$

$$\|\nabla_\vartheta \mathcal{L} - \nabla_\vartheta \mathcal{L}^{\mathrm{sm}}\|_2 \le \frac{C(1+\|\vartheta\|_2)^4}{n^{3/2}} \sum_j \bar{w}'_{1j}, \tag{22}$$

uniformly on $\|\vartheta\|_2 \le C_0 \log n$, where $\mathcal{L}^{\mathrm{sm}}$ is defined in (8)*.*

Proof. $w_1 + w_2 = 1$ gives

$$\widehat{\mu} - \mu = \sum_{i=1}^{2} w_i (g_i - \mu).$$

Write $\Delta\mathbf{z} = \mathbf{z} - \bar{\mathbf{z}}_j$ as in Appendix D. The empirical-moment expansions in Lemma F.6 and Lemma F.1, together with the layer-normalization argument in Appendix D, give

$$\begin{aligned}
\mathbb{E}_j[(g_i-\mu)(g_k-\mu)] &= \frac{\Sigma^{(j)}_{ik}}{n} + \mathcal{O}(n^{-3/2}),\\
\mathbb{E}_j[|g_i-\mu|^4] = \mathcal{O}(n^{-2}), &\qquad \mathbb{E}_j[\|\Delta\mathbf{z}\|_2^8] = \mathcal{O}(n^{-4}),\\
\|\mathbb{E}_j[(g_i-\mu)(g_k-\mu)\Delta\mathbf{z}]\|_2 &\le \big(\mathbb{E}_j[|g_i-\mu|^4]\mathbb{E}_j[|g_k-\mu|^4]\big)^{1/4}\,\mathbb{E}_j[\|\Delta\mathbf{z}\|_2^2]^{1/2}\\
&= \mathcal{O}(n^{-3/2}).
\end{aligned}$$

The exponential-moment bound in the proof of Lemma 3.5 and Hölder's inequality give

$$\begin{aligned}
\mathbb{E}_j\Big[e^{8|\langle\vartheta,\Delta\mathbf{z}\rangle|}\Big] &\le C,\\
\mathbb{E}_j\Big[|g_i-\mu||g_k-\mu|\|\Delta\mathbf{z}\|_2^2 e^{2|\langle\vartheta,\Delta\mathbf{z}\rangle|}\Big] &\le \big(\mathbb{E}_j[|g_i-\mu|^4]\mathbb{E}_j[|g_k-\mu|^4]\big)^{1/4}\\
&\quad\cdot\Big(\mathbb{E}_j[\|\Delta\mathbf{z}\|_2^8]\mathbb{E}_j[e^{8|\langle\vartheta,\Delta\mathbf{z}\rangle|}]\Big)^{1/4} = \mathcal{O}(n^{-2}).
\end{aligned}$$

Taylor expansion of the gate factors at $\bar{\mathbf{z}}_j$ gives

$$\begin{aligned}
w_i w_k &= \bar{w}_{ij}\bar{w}_{kj} + \nabla_{\mathbf{z}}(w_i w_k)|_{\bar{\mathbf{z}}_j}^{\top}\Delta\mathbf{z} + \mathcal{O}\big((1+\|\vartheta\|_2)^2\|\Delta\mathbf{z}\|_2^2\big),\\
w_i w'_1\mathbf{z} &= \bar{w}_{ij}\bar{w}'_{1j}\bar{\mathbf{z}}_j + \bar{w}_{ij}\bar{w}'_{1j}\Delta\mathbf{z} + \bar{\mathbf{z}}_j\,\nabla_{\mathbf{z}}(w_i w'_1)|_{\bar{\mathbf{z}}_j}^{\top}\Delta\mathbf{z}\\
&\quad + \bar{w}'_{1j}\mathcal{O}\Big((1+\|\vartheta\|_2)^2 e^{2|\langle\vartheta,\Delta\mathbf{z}\rangle|}\|\Delta\mathbf{z}\|_2^2\Big),
\end{aligned}$$

where

$$\Big\|\nabla_{\mathbf{z}}(w_i w_k)|_{\bar{\mathbf{z}}_j}\Big\|_2 \le C\|\vartheta\|_2, \qquad \Big\|\nabla_{\mathbf{z}}(w_i w'_1)|_{\bar{\mathbf{z}}_j}\Big\|_2 \le C\bar{w}'_{1j}\|\vartheta\|_2.$$

Substituting these expansions yields

$$\begin{aligned}
\nabla_\vartheta\mathcal{L} &= \frac{1}{2}\sum_j \mathbb{E}_j\left[\left(\sum_{i=1}^{2} w_i(g_i-\mu)\right)(g_1-g_2)w'_1\mathbf{z}\right] &(23)\\
&= \frac{1}{2n}\sum_j \bar{w}'_{1j}\bar{\mathbf{z}}_j\sum_{i=1}^{2}\bar{w}_{ij}(\Sigma^{(j)}_{i1}-\Sigma^{(j)}_{i2}) + \mathcal{O}\left(\frac{(1+\|\vartheta\|_2)^4}{n^{3/2}}\right)\sum_j \bar{w}'_{1j}\\
&= \nabla_\vartheta\mathcal{L}^{\mathrm{sm}}(\vartheta) + \mathcal{O}\left(\frac{(1+\|\vartheta\|_2)^4}{n^{3/2}}\right)\sum_j \bar{w}'_{1j}, &(24)
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}(\vartheta) &= \frac{1}{4}\sum_j \mathbb{E}_j\left[\left(\sum_{i=1}^{2} w_i(g_i-\mu)\right)^2\right]\\
&= \frac{1}{4n}\sum_j\sum_{i,k}\bar{w}_{ij}\bar{w}_{kj}\Sigma^{(j)}_{ik} + \mathcal{O}\left(\frac{(1+\|\vartheta\|_2)^4}{n^{3/2}}\right)\\
&= \mathcal{L}^{\mathrm{sm}}(\vartheta) + \mathcal{O}\left(\frac{(1+\|\vartheta\|_2)^4}{n^{3/2}}\right).
\end{aligned}$$

□

Lemma E.2 (Convergence of the softmax router)**.** *Under the assumptions of Lemma 3.3, consider the population flow $\dot{\vartheta} = -2\nabla_\vartheta \mathcal{L}$ with the expert parameters fixed. For every fixed $B > 0$ and all sufficiently large n, there is a finite time $t_\theta = \mathcal{O}(n^{B+1})$ such that*

$$\max\{\|\bar{w}_{\cdot\mathrm{G}}(t_\theta) - e_1\|_1, \|\bar{w}_{\cdot\mathrm{A}}(t_\theta) - e_2\|_1\} = \mathcal{O}(n^{-B}), \qquad \sup_{0\le t\le t_\theta} \|\vartheta(t)\|_2 = \mathcal{O}(\log n).$$

The reduced flow $\dot{\vartheta} = -2\nabla_\vartheta \mathcal{L}^{\mathrm{sm}}$ satisfies $\bar{w}(t) \to I_2$ as $t \to \infty$. Under the sample-size conditions of Lemma 3.3, the same finite-time bounds hold for $\dot{\vartheta} = -2\nabla_\vartheta \hat{\mathcal{L}}_m$ with probability at least $1 - 1/n$.

Proof. In time $s = t/n$, the reduced gate flow is the flow of Lemma D.8, with time multiplied by four. Its hypotheses hold since $Z_{\mathrm{GG}} = Z_{\mathrm{AA}} = 1$ and $0 \le Z_{\mathrm{GA}} < 1$ by Lemma F.5, which also bounds $\lambda_{\min}(Z)$ below by a positive constant independent of n. The reduced convergence argument there, using avoidance and loss monotonicity, gives $\bar{w}^{\mathrm{reduced}}(s) \to I_2$ and fixed $S, \kappa > 0$ such that

$$n\mathcal{L}^{\mathrm{sm}}(\vartheta^{\mathrm{reduced}}(S)) \le \min_{\bar{w}^{\mathrm{suboptimal}}} n\mathcal{L}^{\mathrm{sm}}(\bar{w}^{\mathrm{suboptimal}}) - 3\kappa, \tag{25}$$

where the minimum is over the three incorrect corners and $\vartheta^{\mathrm{reduced}}(0) = \vartheta(0)$.

The rescaled vector field is uniformly Lipschitz on a bounded neighborhood of $\vartheta^{\mathrm{reduced}}([0, S])$, so Lemma E.1 and Gronwall's inequality give

$$\sup_{0\le s\le S} \|\vartheta(ns) - \vartheta^{\mathrm{reduced}}(s)\|_2 \le \frac{C_S}{\sqrt{n}}. \tag{26}$$

This prevents exit from that neighborhood for sufficiently large n, gives $\|\vartheta(nS)\|_2 = \mathcal{O}(1)$, and transfers (25) to

$$n\mathcal{L}^{\mathrm{sm}}(\vartheta(nS)) \le \min_{\bar{w}^{\mathrm{suboptimal}}} n\mathcal{L}^{\mathrm{sm}}(\bar{w}^{\mathrm{suboptimal}}) - 2\kappa. \tag{27}$$

On this sublevel set, $\bar{w}_{1\mathrm{G}}$ and $\bar{w}_{2\mathrm{A}}$ are bounded below by a positive constant, since the minimum on each face $\bar{w}_{1\mathrm{G}} = 0$ and $\bar{w}_{2\mathrm{A}} = 0$ is attained at an incorrect corner. The strict hierarchy and $Z \succ 0$ therefore give

$$\|\nabla_\vartheta \mathcal{L}^{\mathrm{sm}}\|_2 \asymp \frac{1}{n}\sum_j \bar{w}'_{1j} \asymp \mathcal{L}^{\mathrm{sm}} - \mathcal{L}^{\mathrm{sm}}(I_2), \tag{28}$$

$$\max\{\|\bar{w}_{\cdot\mathrm{G}} - e_1\|_1, \|\bar{w}_{\cdot\mathrm{A}} - e_2\|_1\} \le Cn\left[\mathcal{L}^{\mathrm{sm}} - \mathcal{L}^{\mathrm{sm}}(I_2)\right]. \tag{29}$$

By Lemma E.1, while $\|\vartheta\|_2 \le C_0 \log n$,

$$\frac{\|\nabla_\vartheta(\mathcal{L} - \mathcal{L}^{\mathrm{sm}})\|_2}{\|\nabla_\vartheta \mathcal{L}^{\mathrm{sm}}\|_2} \le \frac{C(1 + \|\vartheta\|_2)^4}{\sqrt{n}} \le \frac{1}{2} \tag{30}$$

for sufficiently large n, hence

$$\begin{aligned} \frac{d}{dt}\mathcal{L}^{\mathrm{sm}} &\le -\|\nabla_\vartheta \mathcal{L}^{\mathrm{sm}}\|_2^2 \le -c\left[\mathcal{L}^{\mathrm{sm}} - \mathcal{L}^{\mathrm{sm}}(I_2)\right]^2, \\ \|\dot{\vartheta}\|_2 &\le 3\|\nabla_\vartheta \mathcal{L}^{\mathrm{sm}}\|_2. \end{aligned} \tag{31}$$

These bounds preserve (27) and, by the parameter-travel calculation in Lemma D.10, give

$$\|\vartheta(t) - \vartheta(nS)\|_2 \le C \log \frac{\mathcal{L}^{\mathrm{sm}}(\vartheta(nS)) - \mathcal{L}^{\mathrm{sm}}(I_2)}{\mathcal{L}^{\mathrm{sm}}(\vartheta(t)) - \mathcal{L}^{\mathrm{sm}}(I_2)} \le CB \log n + C \tag{32}$$

until the excess loss first reaches n^{-B-1}. Choosing C_0 larger than this coefficient prevents exit, and integrating (31) gives attainment by $t_\theta = nS + \mathcal{O}(n^{B+1})$; if the target already holds at nS, take $t_\theta = nS$. The gate bound follows from (29).

For empirical flow, apply Lemma D.4 with $\beta = \min\{\epsilon, 1\}/4$ to the sigmoid parameters $(\vartheta, -\vartheta)$ and subtract the two gradient blocks. Since $\bar{w}_{1j} + \bar{w}_{2j} = 1$, the normalization-error term vanishes.

Conditioning on the family counts as in Section D.3.5, and using the classwise bound $\|\nabla_\vartheta \mathcal{L}^{(j)}\|_2 \le C\bar{w}'_{1j}/n$ from (23), gives, with probability at least $1 - 1/n$, uniformly on $\|\vartheta\|_2 \le C_0 \log n$,

$$\|\nabla_\vartheta(\hat{\mathcal{L}}_m - \mathcal{L})\|_2 \le \widetilde{\mathcal{O}}\Big(\frac{n^{-1/2+\beta}}{\sqrt{m}}\Big) \sum_j \bar{w}'_{1j}.$$

On $[0, S]$, (26) consequently holds with right-hand side $\mathcal{O}(n^{-1/2}) + \widetilde{\mathcal{O}}(n^{1/2+\beta}/\sqrt{m}) = o(1)$, as in Section D.3.5, which preserves (27). On that sublevel set, (28) gives

$$\frac{\|\nabla_\vartheta(\hat{\mathcal{L}}_m - \mathcal{L}^{\mathrm{sm}})\|_2}{\|\nabla_\vartheta \mathcal{L}^{\mathrm{sm}}\|_2} \le \frac{C(1 + \|\vartheta\|_2)^4}{\sqrt{n}} + \widetilde{\mathcal{O}}\Big(\frac{n^{1/2+\beta}}{\sqrt{m}}\Big) \le \frac{1}{2}.$$

Thus (31) and (32) also hold for $\dot{\vartheta} = -2\nabla_\vartheta \hat{\mathcal{L}}_m$, giving the same stopping-time and parameter bounds. □

F Attention-expert dynamics

F.1 Proof of Lemma 3.4

Proof of Lemma 3.4. Since g_k is even in λ_k, it suffices to consider $\lambda_2(t_\theta) > 0$. For sufficiently large n, the uniform event in Lemma F.4 has probability at least $1 - \eta$. On this event, the vector field points inward at $\lambda_1 = \pm 2$ and λ_2 increases, so the flow stays in the stated ranges of λ_1, λ_2 until the first target time. Throughout this interval,

$$\dot{\lambda}_2 \ge \begin{cases} 1/(Cn\lambda_2^2), & \text{Uniform,} \\ e^{-4\sqrt{1-\sigma_n^2}\lambda_2}/(Cn), & \text{Gaussian mixture.} \end{cases}$$

Integration of $\lambda_2^2\, d\lambda_2$ or $e^{4\sqrt{1-\sigma_n^2}\lambda_2}\, d\lambda_2$, respectively, gives the asserted finite times. The initial loss bound in Lemma F.4 and $d\hat{\mathcal{L}}_m/dt = -\|\dot{\lambda}\|_2^2$ give

$$|\lambda_2(t_\lambda) - \lambda_2(t_\theta)|^2 \le (t_\lambda - t_\theta) \int_{t_\theta}^{t_\lambda} \|\dot{\lambda}(t)\|_2^2\, dt \le \frac{C(t_\lambda - t_\theta)}{n}.$$

Both target values diverge, so $(t_\lambda - t_\theta)/n \to \infty$. For every fixed $\delta \in (0, 1)$, the uniform λ_1 gradient expansion in Lemma F.4 gives $\mathrm{sgn}(\lambda_1)\partial_{\lambda_1}\hat{\mathcal{L}}_m \ge \delta^3/n$ whenever $\delta \le |\lambda_1| \le 2$, for sufficiently large n. Thus $|\lambda_1|$ enters $[0, \delta]$ within $2n/\delta^3$ time after t_θ and cannot leave before t_λ. Since δ is arbitrary, $|\lambda_1(t_\lambda)| = o(1)$. □

F.2 Expert-parameter gradients

Throughout this subsection the data follow Definition 2.1, and in the Gaussian-mixture case we assume

$$n^{-1+\epsilon} \le \sigma_n^2, \qquad \sigma_n^2 = o(1)$$

for a constant $\epsilon > 0$ independent of n. Fix $C < \infty$, $p \ge 2$ and, for Uniform data, $\epsilon \in (0, 1)$ independently of n. In the next three lemmas, the covariance Σ_g is evaluated under the indicated distribution. Constants depend on ϵ, C, p, and the gradient inequalities hold for sufficiently large n.

Lemma F.1 (Expert moments and normal gradients)**.** *For $q \in [2]$ and $j \in \{\mathrm{G}, \mathrm{A}\}$, uniformly for $|\lambda_q| \le 2$,*

$$\|g_q - \mu\|_{L^p(\mathbb{P}_j)} + \Big\|\frac{\partial g_q}{\partial \lambda_q}\Big\|_{L^p(\mathbb{P}_j)} \le C_p n^{-1/2}.$$

Uniformly for $|\lambda_1| \le 2$,

$$\mathbb{E}_{\mathrm{G}}\Big[(g_1 - \mu)\frac{\partial g_1}{\partial \lambda_1}\Big] = \frac{1}{2n}\frac{d}{d\lambda_1}\Sigma_g(\lambda_1, \lambda_1) + \mathcal{O}(n^{-2}).$$

At $\lambda_1 = \pm 2$, this gives

$$\mathrm{sgn}(\lambda_1)\mathbb{E}_{\mathrm{G}}\Big[(g_1 - \mu)\frac{\partial g_1}{\partial \lambda_1}\Big] \ge C^{-1}/n.$$

Proof. With $\xi_i = X_i - \mu$ from Definition 2.1 and the population moment ratios from Lemma F.7, the linear terms for $h(\mathbf{X}, \lambda) - \mu - m_j(\lambda)$ and $\frac{\partial h(\mathbf{X},\lambda)}{\partial \lambda} - m_j'(\lambda)$ are the sample averages of

$$\frac{e^{\lambda \xi_i}}{M_Y(\lambda)}(\xi_i - m_j(\lambda)), \qquad \frac{e^{\lambda \xi_i}}{M_Y(\lambda)}\big((\xi_i - m_j(\lambda))^2 - m_j'(\lambda)\big), \tag{33}$$

respectively. The second expression is the derivative of the first with respect to λ. Pairing at $\pm\lambda_q$ cancels the population means, since $m_j(-\lambda) = -m_j(\lambda)$ and $m_j'(-\lambda) = m_j'(\lambda)$. The covariance of the paired linear terms is therefore $(2n)^{-1} d\Sigma_g(\lambda_q, \lambda_q)/d\lambda_q$. For $|\lambda_q| \le 2$, the moment summands have uniformly bounded fixed-order moments and population denominators bounded below. Expansion of the ratios gives the stated L^p bounds in both contexts. Outside a fixed neighborhood of the population moments, higher moment bounds and the sample maximum control the contribution.

For Gaussian data, let

$$V_i = \left(\frac{e^{\nu \xi_i}}{C_\nu}, \frac{\xi_i e^{\nu \xi_i}}{C_\nu}, \frac{\xi_i^2 e^{\nu \xi_i}}{C_\nu} \right)_{\nu = \lambda_q, -\lambda_q}, \qquad \mathbf{V}_n = \frac{1}{n} \sum_i V_i, \quad \mu_V = \mathbb{E}_j[V_i],$$

where $C_\nu = e^{\nu^2/2}$. The normalized population denominators are 1, and all fixed-order moments of V_i are uniformly bounded. Write $f(V)$ for the product $(g_q - \mu)\frac{\partial g_q}{\partial \lambda_q}$ expressed in these moments and $\Delta \mathbf{V}_n = \mathbf{V}_n - \mu_V$. Since $f(\mu_V) = 0$ and $\nabla f(\mu_V) = 0$, Taylor expansion gives

$$f(\mathbf{V}_n) = \frac{1}{2} \Delta \mathbf{V}_n^\top \nabla^2 f(\mu_V) \Delta \mathbf{V}_n + \frac{1}{6} \nabla^3 f(\mu_V) \Delta \mathbf{V}_n^{\otimes 3} + R_{4,n}.$$

Independence and centering give

$$\mathbb{E}_j \left[\frac{1}{2} \Delta \mathbf{V}_n^\top \nabla^2 f(\mu_V) \Delta \mathbf{V}_n \right] = \frac{1}{2n} \frac{d}{d\lambda_q} \Sigma_g(\lambda_q, \lambda_q),$$
$$\mathbb{E}_j[\Delta \mathbf{V}_n^{\otimes 3}] = \mathcal{O}(n^{-2}), \qquad \mathbb{E}_j[\|\Delta \mathbf{V}_n\|_2^4] = \mathcal{O}(n^{-2}).$$

Bounded derivatives near μ_V give $\mathbb{E}_j[|R_{4,n}|] = \mathcal{O}(n^{-2})$ there. Outside that neighborhood, use higher moments and

$$|g_q - \mu| \le \max_i |\xi_i|, \qquad \left|\frac{\partial g_q}{\partial \lambda_q}\right| \le C \max_i |\xi_i|^2.$$

For Gaussian data, the covariance formula gives

$$\frac{1}{2n} \frac{d}{d\lambda_1} \Sigma_g(\lambda_1, \lambda_1) = \frac{\lambda_1}{n} \left[2 \sinh(\lambda_1^2) + \lambda_1^2 \cosh(\lambda_1^2) \right].$$

The claimed sign at $\lambda_1 = \pm 2$ follows. □

Lemma F.2 (Uniform expert moments and gradients)**.** *Uniformly for $C^{-1} \le \lambda_2 \le n^{1-\epsilon}$,*

$$\|g_2 - \mu\|_{L^p(\mathbb{P}_U)} \le \frac{C_p}{\sqrt{n\lambda_2}}, \qquad \left\|\frac{\partial g_2}{\partial \lambda_2}\right\|_{L^p(\mathbb{P}_U)} \le \frac{C_p}{\sqrt{n\lambda_2^3}},$$

$$\mathbb{E}_U\left[(g_2 - \mu)\frac{\partial g_2}{\partial \lambda_2}\right] = \frac{1}{2n} \frac{d}{d\lambda_2} \Sigma_g(\lambda_2, \lambda_2) + \mathcal{O}\left(\frac{1}{n^{3/2}\lambda_2^{3/2}}\right) \le -\frac{1}{Cn\lambda_2^2}.$$

Proof. With $\xi_i = X_i - \mu$ as in the proof of Lemma F.1,

$$\mathbb{E}_U[e^{\lambda_2(\xi_i - \sqrt{3})}] = \frac{1 - e^{-2\sqrt{3}\lambda_2}}{2\sqrt{3}\lambda_2} = \Theta(\lambda_2^{-1}),$$
$$\left\| \frac{\frac{1}{n}\sum_i e^{\lambda_2(\xi_i - \sqrt{3})}}{\mathbb{E}_U[e^{\lambda_2(\xi_i - \sqrt{3})}]} - 1 \right\|_{L^p(\mathbb{P}_U)} \le C_p \sqrt{\lambda_2/n}.$$

The linear terms for the centered ratio and its derivative, given by (33), have L^p norms $\mathcal{O}((n\lambda_2)^{-1/2})$ and $\mathcal{O}((n\lambda_2^3)^{-1/2})$, respectively. These follow by integrating the bounded functions $e^{-\lambda_2 u}$ times

powers of u over $0 \le u \le 2\sqrt{3}$, and applying the moment bounds for independent averages. Dividing by one plus the relative denominator fluctuation leaves remainders $\mathcal{O}(n^{-1})$ and $\mathcal{O}((n\lambda_2)^{-1})$, respectively. For the derivative, the additional squared linear numerator has the latter order as well. The denominator lies outside a fixed relative neighborhood of its expectation with probability at most $C\exp(-n/(C\lambda_2))$, by Bernstein's inequality. Since $n/\lambda_2 \ge n^{\epsilon}$, this event contributes less than any inverse power of n to the stated moments. The product remainders are bounded by

$$C\left(\frac{1}{n\sqrt{n\lambda_2^3}} + \frac{1}{n\lambda_2\sqrt{n\lambda_2}} + \frac{1}{n^2\lambda_2}\right) \le \frac{C}{n^{3/2}\lambda_2^{3/2}}.$$

The Uniform covariance and its derivative in Lemma F.11 satisfy

$$-\frac{d}{d\lambda_2}\Sigma_g(\lambda_2,\lambda_2) = \frac{\sqrt{3}}{4\lambda_2^2} + \mathcal{O}\left(\lambda_2^2 e^{-2\sqrt{3}\lambda_2}\right) \qquad (\lambda_2 \to \infty).$$

Positivity and continuity on $[C^{-1},\infty)$ therefore give

$$\inf_{\lambda_2 \ge C^{-1}}\left(-\lambda_2^2\frac{d}{d\lambda_2}\Sigma_g(\lambda_2,\lambda_2)\right) > 0, \qquad \frac{n^{-3/2}\lambda_2^{-3/2}}{1/(n\lambda_2^2)} = \sqrt{\lambda_2/n} = o(1).$$

□

Lemma F.3 (Gaussian-mixture expert moments and gradients)**.** *Assume $n^{-1+\epsilon} \le \sigma_n^2$ and $\sigma_n^2 = o(1)$ for a fixed $\epsilon > 0$. Uniformly for $C^{-1} \le \lambda_2 \le \log(1/\sigma_n^2)/(4\sqrt{1-\sigma_n^2})$,*

$$\mathbb{E}_{\mathrm{M}}[(g_2-\mu)^2] = \frac{\Sigma_g(\lambda_2,\lambda_2)}{n} + \mathcal{O}(n^{-2}),$$

$$\mathbb{E}_{\mathrm{M}}[(g_2-\mu)\frac{\partial g_2}{\partial\lambda_2}] = \frac{1}{2n}\frac{d}{d\lambda_2}\Sigma_g(\lambda_2,\lambda_2) + \mathcal{O}(n^{-2}) \le -\frac{e^{-4\sqrt{1-\sigma_n^2}\lambda_2}}{Cn},$$

with

$$\|g_2-\mu\|_{L^p(\mathbb{P}_{\mathrm{M}})} + \|\frac{\partial g_2}{\partial\lambda_2}\|_{L^p(\mathbb{P}_{\mathrm{M}})} \le C_p(1+\lambda_2)^C\left(\frac{\sigma_n + e^{-2\sqrt{1-\sigma_n^2}\lambda_2}}{\sqrt{n}} + \frac{1}{n}\right).$$

Proof. Use V_i from the proof of Lemma F.1, with $j = \mathrm{M}$ and $q = 2$ and $C_\nu = e^{\sqrt{1-\sigma_n^2}|\nu| + \sigma_n^2\nu^2/2}$. For $\xi_i = \sqrt{1-\sigma_n^2}S_i + \sigma_n G_i$,

$$\mathbb{E}_{\mathrm{M}}\left[\frac{e^{\nu\xi_i}}{C_\nu}\right] = e^{-\sqrt{1-\sigma_n^2}|\nu|}\cosh(\sqrt{1-\sigma_n^2}\nu) \in [1/2,1],$$

$$\mathbb{E}_{\mathrm{M}}\left[\left(\frac{e^{\nu\xi_i}}{C_\nu}\right)^p\right] = e^{-p\sqrt{1-\sigma_n^2}|\nu|}\cosh(p\sqrt{1-\sigma_n^2}\nu)e^{(p^2-p)\sigma_n^2\nu^2/2} \le e^{(p^2-p)\sigma_n^2\nu^2/2}.$$

Including the polynomial factors by Gaussian tilting preserves boundedness because $\sigma_n\lambda_2 = \mathcal{O}(\sigma_n\log(1/\sigma_n^2)) = o(1)$. The Taylor expansion in the proof of Lemma F.1, applied to $(g_2-\mu)\frac{\partial g_2}{\partial\lambda_2}$ and $(g_2-\mu)^2$, gives the two covariance expansions.

For the Gaussian mixture, (37) gives

$$\Sigma_g(\lambda_2,\lambda_2) = \sigma_n^2 + (1-\sigma_n^2)\operatorname{sech}^4(\sqrt{1-\sigma_n^2}\lambda_2) + \mathcal{O}\big(\sigma_n^4(1+\lambda_2)^4\big),$$

$$\frac{d}{d\lambda_2}\Sigma_g(\lambda_2,\lambda_2) = -4(1-\sigma_n^2)^{3/2}\operatorname{sech}^4(\sqrt{1-\sigma_n^2}\lambda_2)\tanh(\sqrt{1-\sigma_n^2}\lambda_2) + \mathcal{O}\big(\sigma_n^4(1+\lambda_2)^5\big).$$

Since $e^{-4\sqrt{1-\sigma_n^2}\lambda_2} \ge \sigma_n^2$,

$$\frac{\sigma_n^4(1+\lambda_2)^5}{e^{-4\sqrt{1-\sigma_n^2}\lambda_2}} \le \sigma_n^2(1+\lambda_2)^5$$

$$\le C\sigma_n^2(1+\log(1/\sigma_n^2))^5 = o(1),$$

$$\frac{n^{-2}}{e^{-4\sqrt{1-\sigma_n^2}\lambda_2}/n} \le (n\sigma_n^2)^{-1} \le n^{-\epsilon} = o(1).$$

For the refined moment bounds, substitute $\xi_i = \sqrt{1-\sigma_n^2}S_i + \sigma_n G_i$ in the linear terms of (33). At $\sigma_n = 0$ their paired values are bounded by $Ce^{-2\sqrt{1-\sigma_n^2}\lambda_2}$. The difference for $\sigma_n > 0$ has L^p norm at most $C_p(1+\lambda_2)^C\sigma_n$, by Gaussian moments and $\sigma_n\lambda_2 = o(1)$. Averaging the centered linear terms gives the factor $n^{-1/2}$, and the ratio remainders are $\mathcal{O}(n^{-1})$ in L^p. □

Lemma F.4 (Empirical gradients with fixed gating parameters)**.** *Assume the data model of Definition 2.1 and fix $C < \infty$ and $\eta \in (0,1)$ independently of n. For Uniform data, also fix $\epsilon > 0$. For Gaussian-mixture data, fix $\epsilon > 0$ and assume $n^{-1+\epsilon} \le \sigma_n^2$ and $\sigma_n^2 = o(1)$. Consider (θ, λ) satisfying*

$$\max\{\|\bar{w}_{\cdot\mathrm{G}} - e_1\|_1, \|\bar{w}_{\cdot\mathrm{A}} - e_2\|_1\} \le C/n, \qquad \max_q \|\theta_q\|_2 \le C\log n, \qquad |\lambda_1| \le 2,$$

with $C^{-1} \le \lambda_2 \le n^{1-\epsilon}$ for Uniform data and $C^{-1} \le \lambda_2 \le \log(1/\sigma_n^2)/(4\sqrt{1-\sigma_n^2})$ for the mixture. For $m = \widetilde{\Omega}(n^{1+\epsilon})$ independent training tasks and sufficiently large n, with probability at least $1-\eta$, uniformly over this class,

$$\frac{\partial \hat{\mathcal{L}}_m}{\partial \lambda_2} \le \begin{cases} -1/(Cn\lambda_2^2), & \textit{Uniform}, \\ -e^{-4\sqrt{1-\sigma_n^2}\lambda_2}/(Cn), & \textit{Gaussian mixture}. \end{cases}$$

On the same event, uniformly over all (θ, λ) satisfying the bounds above,

$$\frac{\partial \hat{\mathcal{L}}_m}{\partial \lambda_1} = \frac{\lambda_1}{2n}\left[2\sinh(\lambda_1^2) + \lambda_1^2\cosh(\lambda_1^2)\right] + \mathcal{O}(n^{-3/2+o(1)}).$$

On the same event, $\hat{\mathcal{L}}_m(\theta,\lambda) \le C/n$ whenever also $\lambda_2 \le 1$. Constants may depend on ϵ, C, d, and $\boldsymbol{\omega}$.

Proof. For every fixed $p \ge 2$, the feature moment bounds and the sigmoid inequalities in the proof of Lemma 3.5 give

$$\|1-w_1\|_{L^p(\mathbb{P}_{\mathrm{G}})} + \|w_2\|_{L^p(\mathbb{P}_{\mathrm{G}})} + \|1-w_2\|_{L^p(\mathbb{P}_{\mathrm{A}})} + \|w_1\|_{L^p(\mathbb{P}_{\mathrm{A}})} \le C_p/n,$$
$$\|w_k - \bar{w}_{kj}\|_{L^p(\mathbb{P}_j)} \le C_p \log n / n^{3/2}.$$

For $q, k \in [2]$, the fixed $\boldsymbol{\omega}$ give

$$\frac{\partial w_k(\mathbf{X})}{\partial \lambda_q} = w_k(\mathbf{X})(1-w_k(\mathbf{X}))\left\langle \theta_k, \frac{\partial \mathbf{z}(\mathbf{X})}{\partial \lambda_q}\right\rangle = 0,$$
$$\frac{\partial h(\mathbf{X},\lambda)}{\partial \lambda} = \frac{\sum_i X_i^2 e^{\lambda X_i}}{\sum_i e^{\lambda X_i}} - h(\mathbf{X},\lambda)^2,$$
$$\frac{\partial g_q(\mathbf{X})}{\partial \lambda_q} = \frac{1}{2}\left(\frac{\partial h(\mathbf{X},\lambda)}{\partial \lambda}\Big|_{\lambda=\lambda_q} - \frac{\partial h(\mathbf{X},\lambda)}{\partial \lambda}\Big|_{\lambda=-\lambda_q}\right),$$
$$\frac{\partial \widehat{\mu}(\mathbf{X})}{\partial \lambda_q} = w_q(\mathbf{X})\frac{\partial g_q(\mathbf{X})}{\partial \lambda_q}.$$

Consequently,

$$\frac{\partial}{\partial \lambda_q}\frac{1}{2}(\widehat{\mu}-\mu)^2 = (\widehat{\mu}-\mu)w_q\frac{\partial g_q}{\partial \lambda_q}$$
$$= \left[\sum_k w_k(g_k-\mu) + \mu\left(\sum_k w_k - 1\right)\right] w_q\frac{\partial g_q}{\partial \lambda_q},$$
$$\frac{\partial \hat{\mathcal{L}}_m}{\partial \lambda_q} = \frac{1}{m}\sum_{b=1}^m (\widehat{\mu}(\mathbf{X}_b) - \mu_b) w_q(\mathbf{X}_b)\frac{\partial g_q(\mathbf{X}_b)}{\partial \lambda_q}.$$

Since the gates and centered experts are independent of μ,

$$\mathbb{E}_j\left[\mu\left(\sum_k w_k - 1\right) w_q\frac{\partial g_q}{\partial \lambda_q}\right] = \mathbb{E}[\mu]\,\mathbb{E}_j\left[\left(\sum_k w_k - 1\right) w_q\frac{\partial g_q}{\partial \lambda_q}\right] = 0,$$
$$\frac{\partial \mathcal{L}}{\partial \lambda_q} = \frac{1}{2}\sum_j\sum_k \mathbb{E}_j\left[w_q w_k (g_k-\mu)\frac{\partial g_q}{\partial \lambda_q}\right].$$

For $\lambda > 0$ and $u \geq 0$,

$$\sum_{i:\,\max_\ell X_\ell - X_i \geq u} \frac{e^{\lambda X_i}}{\sum_\ell e^{\lambda X_\ell}} \leq \min\{1, ne^{-\lambda u}\},$$

$$\begin{aligned} 0 \leq \frac{\partial h(\mathbf{X},\lambda)}{\partial \lambda} &= \sum_i \frac{e^{\lambda X_i}}{\sum_\ell e^{\lambda X_\ell}} (X_i - h(\mathbf{X},\lambda))^2 \\ &\leq \sum_i \frac{e^{\lambda X_i}}{\sum_\ell e^{\lambda X_\ell}} (\max_\ell X_\ell - X_i)^2 \\ &\leq \int_0^\infty 2u \min\{1, ne^{-\lambda u}\}\, du \leq \frac{C(1+\log n)^2}{\lambda^2}. \end{aligned}$$

Applying the same bound to $-\mathbf{X}$ gives

$$\left|\frac{\partial g_q(\mathbf{X})}{\partial \lambda_q}\right| \leq \frac{C(1+\log n)^2}{\lambda_q^2}.$$

For Uniform data, the gate moment bounds, Hölder's inequality, and Lemma F.2 give

$$\begin{aligned} &\left|\mathbb{E}_{\mathrm{U}}\left[\frac{\partial}{\partial \lambda_2}\frac{1}{2}(\widehat{\mu}-\mu)^2\right] - \mathbb{E}_{\mathrm{U}}\left[(g_2-\mu)\frac{\partial g_2}{\partial \lambda_2}\right]\right| \\ &= \left|\mathbb{E}_{\mathrm{U}}\left[\left((w_2^2-1)(g_2-\mu) + w_1 w_2 (g_1-\mu)\right)\frac{\partial g_2}{\partial \lambda_2}\right]\right| \\ &\leq \frac{C}{n}\left(\frac{1}{\sqrt{n\lambda_2}}\frac{1}{\sqrt{n\lambda_2^3}} + \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n\lambda_2^3}}\right) = C\left(\frac{1}{n^2\lambda_2^2} + \frac{1}{n^2\lambda_2^{3/2}}\right). \end{aligned}$$

In family G, the same bounds give

$$\left|\sum_k \mathbb{E}_{\mathrm{G}}\left[w_2 w_k (g_k - \mu)\frac{\partial g_2}{\partial \lambda_2}\right]\right| \leq \frac{C(1+\log n)^C}{n^{3/2}\lambda_2^2}.$$

Consequently,

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} \leq -\frac{1}{Cn\lambda_2^2}.$$

For the Gaussian mixture, Lemma F.3 gives

$$\left|\mathbb{E}_{\mathrm{M}}\left[\left((w_2^2-1)(g_2-\mu) + w_1 w_2 (g_1-\mu)\right)\frac{\partial g_2}{\partial \lambda_2}\right]\right| \leq \frac{C(1+\log n)^C}{n^2}.$$

In family G, symmetry gives

$$\mathbb{E}_{\mathrm{G}}[g_1 - \mu] = \mathbb{E}_{\mathrm{G}}\left[\frac{\partial g_2}{\partial \lambda_2}\right] = 0.$$

Integration by parts and Cauchy–Schwarz give

$$\left|\mathbb{E}_{\mathrm{G}}\left[(g_1-\mu)\frac{\partial g_2}{\partial \lambda_2}\right]\right| \leq \left\|\max_i \left|\frac{\partial g_1}{\partial X_i}\right|\right\|_{L^2(\mathbb{P}_{\mathrm{G}})} \left\|\sum_i \left|\frac{\partial^2 g_2}{\partial X_i \partial \lambda_2}\right|\right\|_{L^2(\mathbb{P}_{\mathrm{G}})}.$$

Differentiating

$$\frac{\partial h(\mathbf{X},\lambda)}{\partial X_i} = \frac{e^{\lambda X_i}}{\sum_\ell e^{\lambda X_\ell}}\big(1 + \lambda(X_i - h(\mathbf{X},\lambda))\big)$$

gives

$$\left\|\max_i \left|\frac{\partial g_1}{\partial X_i}\right|\right\|_{L^2(\mathbb{P}_{\mathrm{G}})} \leq \frac{1}{n}\left\|C(1+\max_i |X_i-\mu|)e^{C\max_i |X_i-\mu|}\right\|_{L^2(\mathbb{P}_{\mathrm{G}})} = n^{-1+o(1)},$$

$$\left\|\sum_i \left|\frac{\partial^2 g_2}{\partial X_i \partial \lambda_2}\right|\right\|_{L^2(\mathbb{P}_{\mathrm{G}})} \leq C(1+\lambda_2)\left\|(1+\max_i |X_i-\mu|)^2\right\|_{L^2(\mathbb{P}_{\mathrm{G}})} \leq C(1+\log n)^2.$$

Using the gate moment bounds,

$$\left|\sum_k \mathbb{E}_{\mathrm{G}}\left[w_2 w_k (g_k-\mu)\frac{\partial g_2}{\partial \lambda_2}\right]\right| \le \frac{C}{n}\left|\mathbb{E}_{\mathrm{G}}\left[(g_1-\mu)\frac{\partial g_2}{\partial \lambda_2}\right]\right| + \frac{C\log n}{n^{3/2}} n^{-1/2+o(1)} + n^{-2+o(1)}$$
$$\le n^{-2+o(1)}.$$

It follows that

$$\frac{\partial \mathcal{L}}{\partial \lambda_2} \le -\frac{e^{-4\sqrt{1-\sigma_n^2}\lambda_2}}{Cn} + n^{-2+o(1)} \le -\frac{e^{-4\sqrt{1-\sigma_n^2}\lambda_2}}{2Cn}.$$

The same estimates, including the mean term, give

$$\left\|\frac{\partial}{\partial \lambda_2}\frac{1}{2}(\widehat{\mu}-\mu)^2\right\|_{L^p(\mathbb{P}_j)} \le \begin{cases} \dfrac{C_p(1+\log n)^C}{n\lambda_2^2}, & \text{Uniform, } j \in \{\mathrm{G},\mathrm{U}\}, \\ n^{-1+o(1)}(\sigma_n^2 + e^{-4\sqrt{1-\sigma_n^2}\lambda_2}), & \text{Gaussian mixture, } j=\mathrm{M}, \\ n^{-3/2+o(1)}, & \text{Gaussian mixture, } j=\mathrm{G}. \end{cases}$$

For $m = \widetilde{\Omega}(n^{1+\epsilon})$ independent tasks and sufficiently large n, the moment inequality for centered independent sums and the preceding bounds give, for every fixed $\beta > 0$, uniformly on any fixed polynomial-size net in (θ, λ), with probability at least $1 - \eta/4$,

$$\left|\frac{\partial \hat{\mathcal{L}}_m}{\partial \lambda_2} - \frac{\partial \mathcal{L}}{\partial \lambda_2}\right| \le \frac{1}{\sqrt{m}} \begin{cases} \dfrac{n^\beta}{n\lambda_2^2}, & \text{Uniform}, \\ n^{-1+\beta}(\sigma_n^2 + e^{-4\sqrt{1-\sigma_n^2}\lambda_2}) + n^{-3/2+\beta}, & \text{Gaussian mixture}. \end{cases}$$

For a net with at most n^C points, choose $p > 2(C+1)/\beta$. Markov's inequality and a union bound give

$$\mathbb{P}\left(\text{the preceding bound fails at some net point}\right) \le C_p n^C n^{-p\beta/2} \le C_p/n.$$

Since $\|\mathbf{z}\|_2 = 1$ and $\mathbf{z}$ is independent of λ,

$$\sup_{\theta,\lambda}\left\|\nabla_{(\theta,\lambda)}\left[(\widehat{\mu}-\mu)w_q\frac{\partial g_q}{\partial \lambda_q}\right]\right\|_2 \le C(1+|\mu|+\max_i|X_i-\mu|)^4.$$

The expectation of the right-hand side is $\mathcal{O}((1+\log n)^2)$, so Markov's inequality gives, for sufficiently large n, with probability at least $1-\eta/4$,

$$\frac{1}{m}\sum_{b=1}^m (1+|\mu_b|+\max_{\ell\in[n]}|\mathbf{X}_{b\ell}-\mu_b|)^4 \le n.$$

(θ, λ) has $2d+2$ coordinates with ranges of polynomial length. A net of sufficiently fine inverse-polynomial mesh therefore extends the gradient inequalities to the full stated domain. Choose $\beta < \min\{1/4, \epsilon/2\}$. Relative to the respective population gradient lower bounds, the empirical errors are at most

$$\begin{cases} Cn^{-1/2-\epsilon/2+\beta}, & \text{Uniform}, \\ C\left(n^{-1/2-\epsilon/2+\beta} + n^{-3\epsilon/2+\beta}\right), & \text{Gaussian mixture}, \end{cases}$$

which tend to zero.

Uniformly for $|\lambda_1| \le 2$, the gate moment bounds, Lemma F.1, and $|g_2 - \mu| \le \max_i |X_i - \mu|$ give

$$\left\|(\widehat{\mu}-\mu)w_1\frac{\partial g_1}{\partial \lambda_1}\right\|_{L^p(\mathbb{P}_j)} \le C_p/n.$$

Hölder's inequality bounds the population error, while Markov's inequality and a union bound on the parameter net above bound the empirical error at the net points. Extending between net points using the derivative bound above gives, for sufficiently large n, with probability at least $1-\eta/4$, uniformly over all (θ, λ) satisfying the bounds in Lemma F.4,

$$\left|\frac{\partial \mathcal{L}}{\partial \lambda_1} - \frac{1}{2}\mathbb{E}_{\mathrm{G}}\left[(g_1-\mu)\frac{\partial g_1}{\partial \lambda_1}\right]\right| + \left|\frac{\partial \hat{\mathcal{L}}_m}{\partial \lambda_1} - \frac{\partial \mathcal{L}}{\partial \lambda_1}\right| \le n^{-3/2+o(1)}.$$

The Gaussian gradient formula in Lemma F.1 proves the expansion. When also $\lambda_2 \le 1$, the bounded-parameter expert moments and the gate moment bounds give $\|(\widehat{\mu}-\mu)^2\|_{L^p(\mathbb{P}_j)} \le C_p/n$. Applying the same net argument to the loss gives $\hat{\mathcal{L}}_m \le C/n$ uniformly over these parameters, for sufficiently large n, with probability at least $1-\eta/4$. A union bound gives all conclusions jointly with probability at least $1-\eta$. □

F.3 Geometry of $\mathbf{z}$

Lemma F.5 (Geometry of $\mathbf{z}$). *Let $d \ge 3$ and sample $\boldsymbol{\omega}$ as in Definition 2.2. Recall m_j, $\bar{\boldsymbol{\phi}}_j$, and $\bar{\mathbf{z}}_j$ from Section 3.1, with $m_j(0) = 0$ by continuity, and let $Z_{jk} = \langle \bar{\mathbf{z}}_j, \bar{\mathbf{z}}_k \rangle$. Explicitly,*

$$
\begin{aligned}
&m_{\mathrm{G}}(a) = a, \\
&m_{\mathrm{A}}(a) = \sqrt{3}\coth(\sqrt{3}a) - 1/a && \textit{(uniform alternative)}, \\
&m_{\mathrm{A}}(a) = \sigma_n^2 a + \sqrt{1-\sigma_n^2}\tanh\left(\sqrt{1-\sigma_n^2}\,a\right) && \textit{(Gaussian-mixture alternative)}.
\end{aligned}
$$

Then, almost surely over $\boldsymbol{\omega}$,

$$
\|P_d\bar{\boldsymbol{\phi}}_j\|_2 > 0, \qquad Z_{\mathrm{GG}} = Z_{\mathrm{AA}} = 1, \qquad 0 < Z_{\mathrm{GA}} < 1.
$$

In particular $Z \succ 0$, $Z_{\mathrm{GA}} \ge 0$, and $Z_{\mathrm{GA}} < \min\{Z_{\mathrm{GG}}, Z_{\mathrm{AA}}\}$. For the uniform alternative these quantities do not depend on n. For the Gaussian-mixture alternative they depend on n only through σ_n, and if $\sigma_n \to 0$ then, almost surely, Z_{GA} converges to a limit strictly below one. Consequently, $\lambda_{\min}(Z)$ is bounded below by a positive constant independent of n, for all sufficiently large n.

Proof. Let $M_j(a) = \mathbb{E}_j[e^{a(X-\mu)}]$, which is finite for every a because the centered observations are 1-sub-Gaussian. Then $m_j = (\log M_j)'$ and $m_j'(a) = \mathrm{Var}(Y_a) > 0$, where Y_a denotes the centered observation tilted by a. Hence, each m_j is strictly increasing on $\mathbb{R}$. The displayed formulas are the cumulant generating functions of $\mathcal{N}(0,1)$, of $\mathrm{Unif}[-\sqrt{3},\sqrt{3}]$, and of the mixture in Lemma F.14, differentiated once.

The ω_r are almost surely distinct, so strict monotonicity makes the entries of $\bar{\boldsymbol{\phi}}_j$ distinct. In particular, they are not all equal and $P_d\bar{\boldsymbol{\phi}}_j \ne 0$. This proves the first claim, makes $\bar{\mathbf{z}}_j$ well defined, and gives $Z_{jj} = \|\bar{\mathbf{z}}_j\|_2^2 = 1$.

For the lower bound, write $\overline{m_j} = d^{-1}\sum_r m_j(\omega_r)$ and use the identity

$$
\begin{aligned}
\left\langle P_d\bar{\boldsymbol{\phi}}_{\mathrm{G}}, P_d\bar{\boldsymbol{\phi}}_{\mathrm{A}} \right\rangle &= \sum_{r=1}^{d} \left(m_{\mathrm{G}}(\omega_r) - \overline{m_{\mathrm{G}}}\right)\left(m_{\mathrm{A}}(\omega_r) - \overline{m_{\mathrm{A}}}\right) \\
&= \frac{1}{2d}\sum_{r,r'=1}^{d} \left(m_{\mathrm{G}}(\omega_r) - m_{\mathrm{G}}(\omega_{r'})\right)\left(m_{\mathrm{A}}(\omega_r) - m_{\mathrm{A}}(\omega_{r'})\right).
\end{aligned}
$$

Both m_{G} and m_{A} are strictly increasing, so every summand with $\omega_r \ne \omega_{r'}$ is strictly positive and the rest vanish. Since the ω_r are almost surely distinct and $d \ge 2$, the sum is strictly positive, whence $Z_{\mathrm{GA}} > 0$.

For the upper bound, Cauchy–Schwarz gives $Z_{\mathrm{GA}} \le 1$ with equality only if $P_d\bar{\boldsymbol{\phi}}_{\mathrm{A}}$ is a nonnegative multiple of $P_d\bar{\boldsymbol{\phi}}_{\mathrm{G}}$, that is, only if there exist $\alpha, \beta \in \mathbb{R}$ with

$$
m_{\mathrm{A}}(\omega_r) = \alpha\,\omega_r + \beta \qquad \text{for every } r \in [d],
$$

using $m_{\mathrm{G}}(a) = a$. In both cases m_{A} is real analytic on $\mathbb{R}$ and is not affine there: for the uniform alternative $m_{\mathrm{U}}'(a) = a^{-2} - 3\,\mathrm{csch}^2(\sqrt{3}a)$ is nonconstant, and for the mixture $m_{\mathrm{M}}''(a) = -2(1-\sigma_n^2)^{3/2}\,\mathrm{sech}^2(\sqrt{1-\sigma_n^2}a)\tanh(\sqrt{1-\sigma_n^2}a)$ vanishes only at $a = 0$. Condition on (ω_1, ω_2) with $\omega_1 \ne \omega_2$, which determines α and β. The remaining requirement is that ω_3 be a zero of $a \mapsto m_{\mathrm{A}}(a) - \alpha a - \beta$, a real-analytic function that is not identically zero and therefore has at most countably many zeros in $\mathbb{R}$. As $d \ge 3$ and ω_3 is independent of (ω_1, ω_2) with a density, this event has probability zero. Hence $Z_{\mathrm{GA}} < 1$ almost surely, and $\lambda_{\min}(Z) = 1 - Z_{\mathrm{GA}} > 0$.

Only m_{A} depends on n, and only through σ_n in the Gaussian-mixture case. If $\sigma_n \to 0$ then $m_{\mathrm{M}}(\omega_r) \to \tanh(\omega_r)$ for each of the finitely many sampled coordinates $r \in [d]$, so Z_{GA} converges

to the value obtained from $\tanh$, which is strictly below one by the same argument because $\tanh$ is real analytic and not affine. Hence Z_{GA} is bounded away from one, and $\lambda_{\min}(Z)$ away from zero, for all sufficiently large n, with constants depending only on $\boldsymbol{\omega}$. □

F.4 Expert covariance

Lemma F.6 (Expert covariance approximation)**.** *Here, $\boldsymbol{\Sigma}$ is the covariance obtained from the multivariate delta method. Y is sub-Gaussian.*

$$\mathbb{E}\left[n\left(\frac{\sum_{i=1}^{n}\exp(\lambda_k Y_i)Y_i}{\sum_{i=1}^{n}\exp(\lambda_k Y_i)}-\frac{\mathbb{E}\left[\exp(\lambda_k Y)Y\right]}{\mathbb{E}\left[\exp(\lambda_k Y)\right]}\right)^2-\Sigma_{kk}\right]=\mathcal{O}\left(\frac{1}{n}\right) \tag{34}$$

Proof. Set

$$a=\frac{\mathbb{E}[e^{\lambda_k Y}Y]}{\mathbb{E}[e^{\lambda_k Y}]},\qquad B_i=e^{\lambda_k Y_i},\qquad A_i=B_i(Y_i-a),\qquad B^*=\mathbb{E}[B_i],$$

and write $A_n=n^{-1}\sum_i A_i$ and $B_n=n^{-1}\sum_i B_i$. Then $\mathbb{E}[A_i]=0$ and the expression inside the expectation in (34) is $n(A_n/B_n)^2$. Taylor's theorem gives

$$B_n^{-2}=(B^*)^{-2}-2(B^*)^{-3}(B_n-B^*)+3\tilde{B}^{-4}(B_n-B^*)^2, \tag{35}$$

where $\tilde{B}$ lies between B_n and B^*. The constant term in (35) contributes

$$\frac{n\mathbb{E}[A_n^2]}{(B^*)^2}=\frac{\mathrm{Var}(e^{\lambda_k Y}(Y-a))}{\mathbb{E}[e^{\lambda_k Y}]^2}=\Sigma_{kk}.$$

For the linear term, independence and centering leave only coincident indices, so

$$-\frac{2n}{(B^*)^3}\mathbb{E}\big[A_n^2(B_n-B^*)\big]=-\frac{2}{n(B^*)^3}\mathbb{E}\big[A_1^2(B_1-B^*)\big]=\mathcal{O}(n^{-1}).$$

It remains to bound the quadratic remainder. Because $\tilde{B}$ lies between two positive numbers, $\tilde{B}^{-8}\le B_n^{-8}+(B^*)^{-8}$. Jensen's inequality gives $B_n\ge e^{\lambda_k\bar{Y}_n}$, and sub-Gaussianity therefore implies

$$\mathbb{E}[\tilde{B}^{-8}]\le\mathbb{E}[e^{-8\lambda_k\bar{Y}_n}]+(B^*)^{-8}=\mathcal{O}(1).$$

The variables A_i and B_i have moments of every fixed order, and the standard moment bound for centered i.i.d. averages yields

$$\mathbb{E}[A_n^8]=\mathcal{O}(n^{-4}),\qquad\mathbb{E}[(B_n-B^*)^8]=\mathcal{O}(n^{-4}).$$

Hölder's inequality with exponents $2,4,4$ now gives

$$\begin{aligned}3n\mathbb{E}\Big[\tilde{B}^{-4}A_n^2(B_n-B^*)^2\Big]&\le 3n\,\mathbb{E}[\tilde{B}^{-8}]^{1/2}\mathbb{E}[A_n^8]^{1/4}\mathbb{E}[(B_n-B^*)^8]^{1/4}\\&=\mathcal{O}(n^{-1}).\end{aligned}$$

Combining the three terms proves (34). □

F.4.1 Expert hierarchy

The goal of this section is to verify the expert hierarchy of Definition B.1 for Gaussian, uniform and Gaussian-mixture data.

We denote the moment generating function as

$$M_Y(\lambda)=\mathbb{E}\left[\exp\left(\lambda Y\right)\right].$$

Lemma F.7 (Asymptotic normality)**.** *This lemma concerns the full stacked vector $(\lambda_1,-\lambda_1,\lambda_2,-\lambda_2,\omega_1,\dots,\omega_d)\in\mathbb{R}^{4+d}$. For readability, index this stacked vector by $\lambda_1,\dots,\lambda_{4+d}$ within this lemma only. Let $Y_1,\dots,Y_n\overset{\text{i.i.d.}}{\sim}Q_j$. Recall the one-sided attention output h*

from (1). *Write* $\mathbf{h}(\mathbf{Y}) = (h(\mathbf{Y}, \lambda_1), \ldots, h(\mathbf{Y}, \lambda_{4+d}))^\top$ *for these stacked head outputs. For* $\lambda = (\lambda_1, \ldots, \lambda_{4+d})$*,*

$$\sqrt{n}\,(\mathbf{h}(\mathbf{Y}) - m_j(\lambda)) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma}_h),$$

where $m_j(\lambda)$ *is the vector with components* $m_j(\lambda_k) = \frac{M'_Y(\lambda_k)}{M_Y(\lambda_k)}$*. The covariance matrix* $\mathbf{\Sigma}_h$ *is given by:*

$$\Sigma_h(\lambda_k, \lambda_l) = \frac{M''_Y(\lambda_k + \lambda_l) - (m_j(\lambda_k) + m_j(\lambda_l))M'_Y(\lambda_k + \lambda_l) + m_j(\lambda_k)m_j(\lambda_l)M_Y(\lambda_k + \lambda_l)}{M_Y(\lambda_k)M_Y(\lambda_l)}.$$

Proof. By the law of large numbers, as $n \to \infty$, the sample averages converge to expectations:

$$h(\mathbf{Y}, \lambda) \xrightarrow{a.s.} \frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} = \frac{M'_Y(\lambda)}{M_Y(\lambda)} = m_j(\lambda).$$

To find the asymptotic variance, we use the multivariate delta method. Let $V_i(\lambda) = \begin{pmatrix} Y_i e^{\lambda Y_i} \\ e^{\lambda Y_i} \end{pmatrix}$. By the central limit theorem, $\sqrt{n}(\bar{V} - \mathbb{E}[V]) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Sigma}_V)$. Define the mapping $\phi(u, v) = \frac{u}{v}$. The gradient is:

$$\nabla\phi(u, v) = \begin{pmatrix} 1/v \\ -u/v^2 \end{pmatrix}.$$

Evaluated at the expectations $(u, v) = (M'_Y(\lambda), M_Y(\lambda))$, the gradient is:

$$\nabla\phi = \frac{1}{M_Y(\lambda)} \begin{pmatrix} 1 \\ -m_j(\lambda) \end{pmatrix}.$$

The linear approximation (influence function) for the estimator at λ is:

$$\text{IF}_i(\lambda) = \nabla\phi^\top (V_i(\lambda) - \mathbb{E}[V_i(\lambda)]) = \frac{e^{\lambda Y_i}}{M_Y(\lambda)}(Y_i - m_j(\lambda)).$$

The asymptotic covariance between the estimators at λ_k and λ_l is the covariance of their respective influence functions:

$$\begin{aligned}\Sigma_h(\lambda_k, \lambda_l) &= \text{Cov}(\text{IF}(\lambda_k), \text{IF}(\lambda_l)) = \mathbb{E}[\text{IF}(\lambda_k)\text{IF}(\lambda_l)] \\ &= \mathbb{E}\left[\left(\frac{e^{\lambda_k Y}}{M_Y(\lambda_k)}(Y - m_j(\lambda_k))\right)\left(\frac{e^{\lambda_l Y}}{M_Y(\lambda_l)}(Y - m_j(\lambda_l))\right)\right]\end{aligned}$$

Expanding yields

$$\begin{aligned}\Sigma_h(\lambda_k, \lambda_l) &= \frac{1}{M_Y(\lambda_k)M_Y(\lambda_l)}\mathbb{E}\left[e^{(\lambda_k+\lambda_l)Y}(Y - m_j(\lambda_k))(Y - m_j(\lambda_l))\right] \\ &= \frac{1}{M_Y(\lambda_k)M_Y(\lambda_l)}\mathbb{E}\left[e^{(\lambda_k+\lambda_l)Y}(Y^2 - Y(m_j(\lambda_k) + m_j(\lambda_l)) + m_j(\lambda_k)m_j(\lambda_l))\right]\end{aligned}$$

Using the property that $\mathbb{E}[Y^k e^{tY}] = M_Y^{(k)}(t)$, we obtain the final expression. □

F.4.2 Gaussian expert covariance

Lemma F.8 (Gaussian distribution)**.** *If* $Y \sim \mathcal{N}(0, 1)$*, then* $M_Y(t) = e^{t^2/2}$*,*

$$m_{\text{G}}(\lambda) = \lambda,$$
$$\Sigma_h(\lambda_k, \lambda_l) = (1 + \lambda_k\lambda_l)e^{\lambda_k\lambda_l}.$$

Proof. Let $Y \sim \mathcal{N}(0, 1)$. The moment generating function is $M_Y(t) = e^{t^2/2}$. The first two derivatives are:

$$M'_Y(t) = te^{t^2/2} = tM_Y(t),$$
$$M''_Y(t) = (1 + t^2)e^{t^2/2} = (1 + t^2)M_Y(t).$$

The mean function is $m_{\mathrm{G}}(\lambda) = M_Y'(\lambda)/M_Y(\lambda) = \lambda$. We apply the general covariance formula from Lemma F.7:

$$\Sigma_h(\lambda_k, \lambda_l) = \frac{M_Y''(\lambda_k+\lambda_l) - (m_{\mathrm{G}}(\lambda_k)+m_{\mathrm{G}}(\lambda_l))M_Y'(\lambda_k+\lambda_l) + m_{\mathrm{G}}(\lambda_k)m_{\mathrm{G}}(\lambda_l)M_Y(\lambda_k+\lambda_l)}{M_Y(\lambda_k)M_Y(\lambda_l)}.$$

Substituting the derivatives and $m_{\mathrm{G}}(\lambda) = \lambda$:

$$\begin{aligned}
&M_Y''(\lambda_k+\lambda_l) - (m_{\mathrm{G}}(\lambda_k)+m_{\mathrm{G}}(\lambda_l))M_Y'(\lambda_k+\lambda_l) + m_{\mathrm{G}}(\lambda_k)m_{\mathrm{G}}(\lambda_l)M_Y(\lambda_k+\lambda_l)\\
=&(1+(\lambda_k+\lambda_l)^2)M_Y(\lambda_k+\lambda_l) - (\lambda_k+\lambda_l)^2 M_Y(\lambda_k+\lambda_l) + \lambda_k\lambda_l M_Y(\lambda_k+\lambda_l)\\
=&M_Y(\lambda_k+\lambda_l)\left[1+(\lambda_k+\lambda_l)^2 - (\lambda_k+\lambda_l)^2 + \lambda_k\lambda_l\right]\\
=&M_Y(\lambda_k+\lambda_l)(1+\lambda_k\lambda_l).
\end{aligned}$$

Dividing by the denominator $M_Y(\lambda_k)M_Y(\lambda_l) = e^{\lambda_k^2/2}e^{\lambda_l^2/2}$:

$$\begin{aligned}
\Sigma_h(\lambda_k, \lambda_l) &= (1+\lambda_k\lambda_l)\frac{e^{(\lambda_k+\lambda_l)^2/2}}{e^{(\lambda_k^2+\lambda_l^2)/2}}\\
&= (1+\lambda_k\lambda_l)\exp\left(\frac{\lambda_k^2+2\lambda_k\lambda_l+\lambda_l^2-\lambda_k^2-\lambda_l^2}{2}\right)\\
&= (1+\lambda_k\lambda_l)e^{\lambda_k\lambda_l}.
\end{aligned}$$

□

Definition F.9 (Symmetrized expert covariance)**.** Recall $g_k(\mathbf{X}) = \frac{1}{2}\left(h(\mathbf{X}, \lambda_k) + h(\mathbf{X}, -\lambda_k)\right)$. For $a, b \in \mathbb{R}$, the asymptotic covariance of $h(\cdot, a)$ and $h(\cdot, -a)$'s symmetrization against the analogous pair at b is denoted $\Sigma_g(a, b)$. Since each Q_j is symmetric about the origin, it follows from the linearity of the influence functions that:

$$\Sigma_g(a, b) = \frac{1}{2}\left(\Sigma_h(a, b) + \Sigma_h(a, -b)\right) \tag{36}$$

where $\Sigma_h(\cdot, \cdot)$ is the one-sided covariance defined in Lemma F.7.

F.4.3 Uniform expert covariance

Lemma F.10 (Uniform distribution)**.** *If* $Y \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$*, then*

$$m_{\mathrm{U}}(\lambda) = \sqrt{3}\coth(\sqrt{3}\lambda) - \frac{1}{\lambda},$$

and

$$\begin{aligned}
\Sigma_h(\lambda_k, \lambda_l) = &\frac{2\sqrt{3}\lambda_k\lambda_l\{\coth(\sqrt{3}\lambda_k) + \coth(\sqrt{3}\lambda_l)\}}{(\lambda_k+\lambda_l)^3}\\
&- \frac{3\{\lambda_k^2\,\mathrm{csch}^2(\sqrt{3}\lambda_k) + \lambda_l^2\,\mathrm{csch}^2(\sqrt{3}\lambda_l)\}}{(\lambda_k+\lambda_l)^2}.
\end{aligned}$$

Proof. Put $a = \sqrt{3}$. The moment generating function and its tilted mean are

$$M_Y(\lambda) = \frac{\sinh(a\lambda)}{a\lambda}, \qquad m_{\mathrm{U}}(\lambda) = \frac{M_Y'(\lambda)}{M_Y(\lambda)} = a\coth(a\lambda) - \frac{1}{\lambda},$$

with $m_{\mathrm{U}}'(\lambda) = -a^2\,\mathrm{csch}^2(a\lambda) + \lambda^{-2}$. The general covariance identity from Lemma F.7 can be written

$$\begin{aligned}
\Sigma_h(\lambda_k, \lambda_l) = \frac{M_Y(\lambda_k+\lambda_l)}{M_Y(\lambda_k)M_Y(\lambda_l)}\Big[&m_{\mathrm{U}}'(\lambda_k+\lambda_l)\\
&+ \{m_{\mathrm{U}}(\lambda_k+\lambda_l) - m_{\mathrm{U}}(\lambda_k)\}\{m_{\mathrm{U}}(\lambda_k+\lambda_l) - m_{\mathrm{U}}(\lambda_l)\}\Big].
\end{aligned}$$

Substitute the displayed formulas for $M_Y, m_{\mathrm{U}}, m_{\mathrm{U}}'$ and use

$$\frac{\sinh(a(\lambda_k+\lambda_l))}{\sinh(a\lambda_k)\sinh(a\lambda_l)} = \coth(a\lambda_k) + \coth(a\lambda_l).$$

Collecting the rational and hyperbolic terms gives exactly the stated expression. □

Lemma F.11 (Monotonicity for uniform distribution)**.** *Let $Y \sim \mathrm{Unif}[-\sqrt{3},\sqrt{3}]$. Then $\Sigma_g(\lambda,\lambda)$ is strictly decreasing in $|\lambda|$.*

Proof. For $t=\sqrt{3}\lambda>0$, Lemma F.10 and symmetry give

$$\Sigma_g(t/\sqrt{3},t/\sqrt{3})=\frac{3}{4}\left(\frac{\coth t}{t}-\operatorname{csch}^2 t\right)+\frac{1}{2}t^2\operatorname{csch}^2 t\left[1-3\left(\coth t-\frac{1}{t}\right)^2\right].$$

$$\begin{aligned}&-64t^2\sinh^5 t\,\frac{d}{dt}\Sigma_g(t/\sqrt{3},t/\sqrt{3})\\&\quad=3\cosh(5t)-(9+120t^2+32t^4)\cosh(3t)\\&\qquad+(6+120t^2-352t^4)\cosh t+(12t+128t^3)\sinh(3t)\\&\qquad+(-36t+384t^3)\sinh t=\sum_{m\ge 0}A_m t^m.\end{aligned}$$

$$A_0=A_2=A_4=A_6=0,\qquad A_{2k+1}=0\quad(k\ge 0),$$

m	8	10	12	14	16	18
A_m	$\frac{256}{5}$	$\frac{3712}{315}$	$\frac{6368}{4725}$	$\frac{3952}{31185}$	$\frac{539606}{42567525}$	$\frac{1979}{1658475}$

For even $m\ge 20$, dropping positive terms and bounding falling factorials by powers gives

$$\frac{m!A_m}{5^m}\ge 3-\left(\frac{3}{5}\right)^m\left(9+\frac{40}{3}m^2+\frac{32}{81}m^4\right)-\frac{352m^4+36m}{5^m}.$$

Each subtracted term decreases for $m\ge 20$, since its ratio at consecutive integers is at most

$$\frac{3}{5}\left(\frac{m+1}{m}\right)^4\le\frac{3}{5}\left(\frac{21}{20}\right)^4=\frac{583443}{800000}<1.$$

Thus

$$\frac{m!A_m}{5^m}\ge 3-\frac{239026832372329}{95367431640625}>\frac{1}{3},\qquad\frac{d}{dt}\Sigma_g(t/\sqrt{3},t/\sqrt{3})<0\quad(t>0).$$

Evenness and continuity at zero give the claim. □

Lemma F.12 (Symmetrized cross-covariance for uniform distribution)**.** *Let $Y\sim\mathrm{Unif}[-\sqrt{3},\sqrt{3}]$ and $\lambda_1,\lambda_2\in\mathbb{R}$. Then, with the right-hand side interpreted by continuity when $\lambda_1\lambda_2=0$ or $\lambda_1=\pm\lambda_2$,*

$$\begin{aligned}\Sigma_g(\lambda_1,\lambda_2)=\sqrt{3}\lambda_1\lambda_2&\left[\frac{\coth(\sqrt{3}\lambda_1)+\coth(\sqrt{3}\lambda_2)}{(\lambda_1+\lambda_2)^3}-\frac{\coth(\sqrt{3}\lambda_1)-\coth(\sqrt{3}\lambda_2)}{(\lambda_1-\lambda_2)^3}\right]\\&-\frac{3}{2}\left[\lambda_1^2\operatorname{csch}^2(\sqrt{3}\lambda_1)+\lambda_2^2\operatorname{csch}^2(\sqrt{3}\lambda_2)\right]\\&\qquad\times\left[\frac{1}{(\lambda_1+\lambda_2)^2}+\frac{1}{(\lambda_1-\lambda_2)^2}\right].\end{aligned}$$

Proof. By symmetry, $\Sigma_g(\lambda_1,\lambda_2)=\frac{1}{2}\{\Sigma_h(\lambda_1,\lambda_2)+\Sigma_h(\lambda_1,-\lambda_2)\}$. Substitute the formula in Lemma F.10 and use $\coth(-x)=-\coth(x)$ and $\operatorname{csch}^2(-x)=\operatorname{csch}^2(x)$. Collecting the common terms gives the displayed expression. Continuity of the covariance formula in Lemma F.7 gives the stated extensions. For $\lambda_1=\pm\lambda_2$, evenness gives the self-variance $\Sigma_g(\lambda_1,\lambda_1)$. □

Corollary F.13 (Expert hierarchy for the uniform distribution)**.** *For the Uniform distribution and $\lambda_i\in[-1,1]$, Σ_g satisfies the expert hierarchy of Definition B.1.*

Proof. Σ_g is even in each argument, so it suffices to treat $0 \le \lambda_1 < \lambda_2 \le 1$. Write

$$M_Y(\lambda) = \frac{\sinh(\sqrt{3}\lambda)}{\sqrt{3}\lambda}, \qquad \chi_\lambda(x) = M_Y(\lambda) x \cosh(\lambda x) - M_Y'(\lambda) \sinh(\lambda x),$$

with continuous extensions at zero. For $u > 0$,

$$\begin{aligned} \frac{d}{du}(u \coth u) &= \frac{\sinh(2u) - 2u}{2\sinh^2 u} > 0, \\ \frac{2u}{3} - \frac{d}{du}(u \coth u) &= \frac{2u\cosh(2u) + 4u - 3\sinh(2u)}{6\sinh^2 u} \\ &= \frac{1}{6\sinh^2 u} \sum_{k \ge 2} \frac{2^{2k+1}(2k-2)}{(2k+1)!} u^{2k+1} > 0. \end{aligned}$$

Hence

$$\lim_{t \downarrow 0} t \coth t = 1, \qquad 0 < s\coth s - t \coth t < \frac{s^2 - t^2}{3} \quad (0 \le t < s).$$

For $0 < x < \sqrt{3}$ and $0 < \lambda \le 1$,

$$\begin{aligned} \frac{\chi_\lambda(x)}{M_Y(\lambda)^2} &= \frac{\sinh(\lambda x)}{\lambda M_Y(\lambda)} \left[1 - \sqrt{3}\lambda \coth(\sqrt{3}\lambda) + \lambda x \coth(\lambda x)\right] \\ &> \frac{\sinh(\lambda x)}{\lambda M_Y(\lambda)} (1 - \lambda^2) \ge 0, \\ \frac{\partial}{\partial \lambda} \frac{\chi_\lambda(x)}{M_Y(\lambda)^2} &= \frac{\sinh(\lambda x)}{\lambda^2 M_Y(\lambda)} \Big[\lambda^2(x^2 - 3) + 2\{\sqrt{3}\lambda \coth(\sqrt{3}\lambda) - 1\} \\ &\qquad \times \{\sqrt{3}\lambda \coth(\sqrt{3}\lambda) - \lambda x \coth(\lambda x)\}\Big] \\ &\le -\frac{(3 - x^2)\sinh(\lambda x)}{M_Y(\lambda)} \left(1 - \frac{2\lambda^2}{3}\right) < 0. \end{aligned}$$

Using $\chi_0(x) = x$, oddness in x, and Lemma F.7,

$$\begin{aligned} \Sigma_g(\lambda_2, \lambda_2) &= \frac{1}{\sqrt{3}} \int_0^{\sqrt{3}} \frac{\chi_{\lambda_2}(x)^2}{M_Y(\lambda_2)^4}\, dx \\ &< \frac{1}{\sqrt{3}} \int_0^{\sqrt{3}} \frac{\chi_{\lambda_2}(x)\chi_{\lambda_1}(x)}{M_Y(\lambda_2)^2 M_Y(\lambda_1)^2}\, dx = \Sigma_g(\lambda_1, \lambda_2) \\ &< \frac{1}{\sqrt{3}} \int_0^{\sqrt{3}} \frac{\chi_{\lambda_1}(x)^2}{M_Y(\lambda_1)^4}\, dx = \Sigma_g(\lambda_1, \lambda_1). \end{aligned}$$

□

Taking $\lambda_1 = \lambda_2$ above also gives $\frac{d}{d\lambda}\Sigma_g(\lambda, \lambda) < 0$ on $(0, 1)$. Lemma F.11 is needed for the larger λ reached in the second stage, where the hierarchy itself fails and no bounded-range argument applies.

F.4.4 GAUSSIAN-MIXTURE EXPERT COVARIANCE

We now give the analogous proof for the Gaussian mixture in Definition 2.1. Write $a_n = \sqrt{1 - \sigma_n^2}$. The centered observation has the representation $Y \overset{D}{=} a_n S + \sigma_n G$, where S is Rademacher, G is standard Gaussian, and S, G are independent. Thus Y has mean zero and variance one.

Lemma F.14 (Mixture moments)**.**

$$\begin{aligned} M_Y(t) &= e^{\sigma_n^2 t^2/2} \cosh(a_n t), \\ m_{\mathrm{M}}(t) &= \sigma_n^2 t + a_n \tanh(a_n t), \\ m_{\mathrm{M}}'(t) &= \sigma_n^2 + a_n^2 \operatorname{sech}^2(a_n t). \end{aligned}$$

Moreover, Y is 1-sub-Gaussian.

Proof. Independence of S and G gives

$$M_Y(t) = \mathbb{E}[e^{ta_nS}]\mathbb{E}[e^{t\sigma_nG}] = \cosh(a_nt)e^{\sigma_n^2t^2/2}.$$

Differentiating $\log M_Y$ twice gives $m_{\mathrm{M}}(t) = M_Y'(t)/M_Y(t)$ and $m_{\mathrm{M}}'(t) = M_Y''(t)/M_Y(t) - m_{\mathrm{M}}(t)^2$. Finally, $\cosh(a_nt) \le e^{a_n^2t^2/2}$ gives $M_Y(t) \le e^{t^2/2}$. □

Lemma F.15 (Gaussian mixture expert covariance)**.**

$$\Sigma_h(\lambda_k,\lambda_l) = \frac{M_Y(\lambda_k+\lambda_l)}{M_Y(\lambda_k)M_Y(\lambda_l)}\Big[m_{\mathrm{M}}'(\lambda_k+\lambda_l) + (m_{\mathrm{M}}(\lambda_k+\lambda_l) - m_{\mathrm{M}}(\lambda_k))(m_{\mathrm{M}}(\lambda_k+\lambda_l) - m_{\mathrm{M}}(\lambda_l))\Big].$$

Writing $\Sigma_g(\lambda_k,\lambda_l) = \frac{1}{2}\{\Sigma_h(\lambda_k,\lambda_l) + \Sigma_h(\lambda_k,-\lambda_l)\}$,

$$\begin{aligned}\Sigma_g(\lambda_k,\lambda_l) = {}& \cosh(\sigma_n^2\lambda_k\lambda_l)\Big[a_n^2\operatorname{sech}^2(a_n\lambda_k)\operatorname{sech}^2(a_n\lambda_l) + \sigma_n^2 \\ &+ \sigma_n^4\lambda_k\lambda_l\tanh(a_n\lambda_k)\tanh(a_n\lambda_l)\Big] \\ &+ \sinh(\sigma_n^2\lambda_k\lambda_l)\Big[\sigma_n^2\tanh(a_n\lambda_k)\tanh(a_n\lambda_l) + \sigma_n^4\lambda_k\lambda_l \\ &+ \sigma_n^2a_n\{\lambda_k\operatorname{sech}^2(a_n\lambda_k)\tanh(a_n\lambda_l) + \lambda_l\operatorname{sech}^2(a_n\lambda_l)\tanh(a_n\lambda_k)\}\Big].\end{aligned} \tag{37}$$

In particular, $\Sigma_g(\lambda_k,0) = \sigma_n^2 + a_n^2\operatorname{sech}^2(a_n\lambda_k)$ *and* $\Sigma_g(0,0) = 1$.

Proof. The formula for Σ_h is Lemma F.7 rewritten using Lemma F.14, exactly as in the uniform case. Exponentially tilting $a_nS + \sigma_nG$ by $\lambda_k + \lambda_l$ shifts the Gaussian mean by $\sigma_n^2(\lambda_k+\lambda_l)$ and tilts the Rademacher probabilities. For $S \in \{-1,1\}$,

$$\frac{e^{a_n\lambda_kS}}{\cosh(a_n\lambda_k)}\big(a_nS - a_n\tanh(a_n\lambda_k)\big) = a_n\operatorname{sech}^2(a_n\lambda_k)S.$$

Substitution gives

$$\begin{aligned}\Sigma_h(\lambda_k,\lambda_l) = e^{\sigma_n^2\lambda_k\lambda_l}\Big[&a_n^2\operatorname{sech}^2(a_n\lambda_k)\operatorname{sech}^2(a_n\lambda_l) \\ &+ (\sigma_n^2 + \sigma_n^4\lambda_k\lambda_l)\{1 + \tanh(a_n\lambda_k)\tanh(a_n\lambda_l)\} \\ &+ \sigma_n^2a_n\{\lambda_k\operatorname{sech}^2(a_n\lambda_k)\tanh(a_n\lambda_l) + \lambda_l\operatorname{sech}^2(a_n\lambda_l)\tanh(a_n\lambda_k)\}\Big].\end{aligned}$$

Averaging this expression with its value at $-\lambda_l$ and using the definitions of $\cosh$ and $\sinh$ gives (37). □

Lemma F.16 (Expert hierarchy for the Gaussian mixture)**.** *For every fixed* $0 < \delta < 1$ *there exists* $\sigma_\delta > 0$ *such that, for all* $0 < \sigma_n < \sigma_\delta < 1/\sqrt{2}$ *and*

$$0 \le |\lambda_1| < |\lambda_2| \le L_{\max} = (1-\delta)\log(1/\sigma_n),$$

we have

$$\Sigma_g(\lambda_2,\lambda_2) < \Sigma_g(\lambda_2,\lambda_1) < \Sigma_g(\lambda_1,\lambda_1).$$

Quantitatively,

$$\min\{\Sigma_g(\lambda_2,\lambda_1) - \Sigma_g(\lambda_2,\lambda_2), \Sigma_g(\lambda_1,\lambda_1) - \Sigma_g(\lambda_2,\lambda_1)\} \ge \frac{e^{-4L_{\max}}}{8(1+L_{\max})}(\lambda_2^2 - \lambda_1^2).$$

Thus the expert hierarchy of Definition B.1 holds throughout $[-L_{\max}, L_{\max}]$*, and* $L_{\max} \to \infty$ *as* $\sigma_n \downarrow 0$ *for any fixed* δ.

Proof. Write $u = \lambda_2^2$, $v = \lambda_1^2$, and separate

$$\Sigma_g(\sqrt{u}, \sqrt{v}) = a_n^2 \operatorname{sech}^2(a_n\sqrt{u}) \operatorname{sech}^2(a_n\sqrt{v}) + \sigma_n^2 + R_{\sigma_n}(u, v).$$

For $L \geq 1$ and $0 \leq u, v \leq L^2$, expanding (37), using $\cosh t - 1 \leq t^2 e^t/2$ and $1 \leq \sinh(t)/t \leq e^t$, and bounding the resulting derivatives gives a universal C such that

$$|\partial_u R_{\sigma_n}(u, v)| + |\partial_v R_{\sigma_n}(u, v)| \leq C\sigma_n^4 (1 + L)^8 e^{\sigma_n^2 L^2}.$$

For the dominant term, $a_n^4 \geq 1/4$, $\operatorname{sech} t \geq e^{-t}$, and $\tanh(t)/t \geq 1/(1 + t)$ imply

$$\begin{aligned} &-\frac{\partial}{\partial u}\{a_n^2 \operatorname{sech}^2(a_n\sqrt{u}) \operatorname{sech}^2(a_n\sqrt{v})\} \\ &\quad = a_n^4 \frac{\tanh(a_n\sqrt{u})}{a_n\sqrt{u}} \operatorname{sech}^2(a_n\sqrt{u}) \operatorname{sech}^2(a_n\sqrt{v}) \\ &\quad \geq \frac{e^{-4L}}{4(1 + L)}. \end{aligned}$$

Hence $\partial_u \Sigma_g(\sqrt{u}, \sqrt{v}) \leq -e^{-4L}/\{8(1 + L)\}$, and symmetrically in v, whenever

$$8C\sigma_n^4 (1 + L)^9 e^{4L + \sigma_n^2 L^2} \leq 1.$$

This condition holds at $L = L_{\max} = (1 - \delta) \log(1/\sigma_n)$ for sufficiently small σ_n, since $\sigma_n^4 e^{4L_{\max}} = \sigma_n^{4\delta} \to 0$. Integrating the derivative bounds in one coordinate between λ_1^2 and λ_2^2 gives both quantitative gaps and hence the strict hierarchy. □

F.5 Concentration bounds

F.5.1 Softmax concentration

Lemma F.17 (Concentration of softmax)**.** *Let $Y_1, \ldots, Y_n$ be independent copies of a mean-zero 1-sub-Gaussian variable Y, and fix $\lambda \in \mathbb{R}$ and an integer $k \geq 1$. There are constants $C(\lambda), C_2(\lambda) > 0$, depending only on λ, k, such that, for $0 < M \leq 1$ and $R \geq 1$ with $\mathbb{P}(|Y| > R) \leq M^2 C_2(\lambda)$,*

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n e^{\lambda Y_i} Y_i^k}{\sum_{i=1}^n e^{\lambda Y_i}} - \frac{\mathbb{E}[Y^k e^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]}\right| \geq MC(\lambda)\right) \leq 2ne^{-R^2/2} + 4\exp\left(-\frac{nM^2}{32R^{2k}e^{2R|\lambda|}}\right).$$

Proof. Set $A_i = Y_i^k e^{\lambda Y_i} \mathbf{1}_{\{|Y_i| \leq R\}}$ and $B_i = e^{\lambda Y_i} \mathbf{1}_{\{|Y_i| \leq R\}}$, and write $A_n = n^{-1} \sum_i A_i$, $B_n = n^{-1} \sum_i B_i$, $A^* = \mathbb{E}[Y^k e^{\lambda Y}]$, and $B^* = \mathbb{E}[e^{\lambda Y}] \geq 1$. The A_i are independent, as are the B_i, and satisfy $|A_i| \leq R^k e^{R|\lambda|}$ and $0 \leq B_i \leq e^{R|\lambda|}$. The second moments of $Y^k e^{\lambda Y}$ and $e^{\lambda Y}$ are bounded in terms of λ, k by the sub-Gaussian assumption. Cauchy–Schwarz and a sufficiently small $C_2(\lambda)$ give $|\mathbb{E}A_n - A^*|, |\mathbb{E}B_n - B^*| \leq M/4$. Hoeffding's inequality therefore gives $|A_n - A^*|, |B_n - B^*| \leq M/2$, except with probability at most $4\exp(-nM^2/(32R^{2k}e^{2R|\lambda|}))$. On this event, $B_n \geq 1/2$, so

$$\left|\frac{A_n}{B_n} - \frac{A^*}{B^*}\right| \leq \frac{|A_n - A^*|}{B_n} + \frac{|A^*||B_n - B^*|}{B_n B^*} \leq M(1 + |A^*|). \tag{38}$$

The sub-Gaussian moment bound allows $C(\lambda) > 1 + |A^*|$ to be chosen depending only on λ, k. Finally, $\mathbb{P}(\max_i |Y_i| > R) \leq 2ne^{-R^2/2}$, and outside this event the truncated and original ratios agree. □

G Additional results

Lemma G.1 (Lyapunov exit time)**.** *Consider $\dot{x} = f(x)$ and a region R. Suppose W is locally Lipschitz, $W(x) > C$ on R, and $\frac{d}{dt}W(x(t)) \leq -\gamma < 0$ almost everywhere while $x(t) \in R$. If $x(0) \in R$ and the solution exists up to exit, then its first exit time is at most*

$$t^* = \frac{W(x(0)) - C}{\gamma}.$$

Moreover, $W(x(t)) \leq W(x(0))$ until exit.

Proof. While the trajectory remains in R,

$$C < W(x(t)) \le W(x(0)) - \gamma t.$$

At $t = t^*$ these inequalities contradict each other. □

Theorem G.2 (Tikhonov's theorem)**.** *Consider*

$$\varepsilon \dot{r} = f(r, y), \qquad \dot{y} = g(r, y), \qquad (r(0), y(0)) = (r_0, y_0).$$

Let $\mathcal{N}_0$ be open and $r^\star$ be a twice continuously differentiable function on $\mathcal{N}_0$ with $f(r^\star(y), y) = 0$ and graph contained in an open set $\mathcal{N}$. Assume that, for every $(r, y) \in \mathcal{N}$ with $y \in \mathcal{N}_0$, the line segment from $(r^\star(y), y)$ to (r, y) lies in $\mathcal{N}$. Assume that f, g and their derivatives through order two are uniformly bounded in $\mathcal{N}$, and that all eigenvalues of $\partial_r f(r^\star(y), y)$ have real parts at most $-a_0 < 0$, uniformly for $y \in \mathcal{N}_0$. Let

$$\dot{y}^0 = g(r^\star(y^0), y^0), \qquad y^0(0) = y_0,$$

and let $L_g > 0$ be a Lipschitz constant for $y \mapsto g(r^\star(y), y)$ on $\mathcal{N}_0$. There exist $\varepsilon_0, c_0, \dots, c_4, \kappa_0 > 0$ such that, for $0 < \varepsilon < \varepsilon_0$ and $(r_0, y_0) \in \mathcal{N}$ with $\|r_0 - r^\star(y_0)\|_2 < c_0$,

$$\|r(s) - r^\star(y(s))\|_2 \le c_1\varepsilon + c_2\|r_0 - r^\star(y_0)\|_2 e^{-\kappa_0 s/\varepsilon},$$
$$\|y(s) - y^0(s)\|_2 \le c_3\varepsilon e^{L_g s} + c_4\|r_0 - r^\star(y_0)\|_2 e^{-\kappa_0 s/\varepsilon},$$

for all $s \ge 0$ such that $(r(u), y(u)) \in \mathcal{N}$ and $y(u), y^0(u) \in \mathcal{N}_0$ for $0 \le u \le s$.

Proof. See Berglund (2001, Theorem 4.1.2). □

H Lower bound for the Gaussian mixture

H.1 Gaussian and uniform benchmarks

Proposition H.1 (Gaussian and uniform benchmarks)**.** *For the variance-one Gaussian location family, $R^\star_{\mathrm{G},n} = n^{-1}$, attained by the sample mean. For the uniform location family, $R^\star_{\mathrm{U},n} \asymp n^{-2}$, and the sample mid-range has risk*

$$\mathbb{E}_{\mu,\mathrm{U}}\left[\left(\frac{X_{(1)} + X_{(n)}}{2} - \mu\right)^2\right] = \frac{6}{(n+1)(n+2)}.$$

The sample mean instead has risk n^{-1} under either family.

Proof. See (Lehmann & Casella, 1998, Theorem 3.5) or (Shao, 2003, Section 2). □

H.2 Gaussian-mixture benchmark

Lemma H.2 (Fisher information and lower bounds for the location)**.** *For known $\sigma_n \in (0, 1)$, the observations have density*

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(x-\mu)^2 + 1 - \sigma_n^2}{2\sigma_n^2}\right) \cosh\left(\frac{\sqrt{1-\sigma_n^2}(x-\mu)}{\sigma_n^2}\right), \tag{39}$$

and the Fisher information in one observation is

$$I_{\sigma_n} = \frac{1}{\sigma_n^2} - \frac{1-\sigma_n^2}{\sigma_n^4}\mathbb{E}\left[\operatorname{sech}^2\left(\frac{\sqrt{1-\sigma_n^2}(X-\mu)}{\sigma_n^2}\right) \mid \mu\right]. \tag{40}$$

As $\sigma_n \downarrow 0$,

$$I_{\sigma_n} = \frac{1}{\sigma_n^2}\left[1 - \sqrt{\frac{\pi}{2}}\frac{\sqrt{1-\sigma_n^2}}{\sigma_n} e^{-(1-\sigma_n^2)/(2\sigma_n^2)}\left(1 + \mathcal{O}(\sigma_n^2)\right)\right]. \tag{41}$$

By the Cramér–Rao inequality, every unbiased estimator of μ based on n i.i.d. observations satisfies

$$\mathrm{Var}_\mu(\widehat{\mu}) \ge \frac{1}{nI_{\sigma_n}} \sim \frac{\sigma_n^2}{n} \qquad \text{as } \sigma_n \downarrow 0. \tag{42}$$

$$R^\star_{\mathrm{M},n} \ge \frac{\sigma_n^2}{n}.$$

Proof. Differentiating (39) with respect to μ gives

$$\frac{\partial}{\partial \mu} \log f_\mu(x) = \frac{x - \mu - \sqrt{1-\sigma_n^2}\tanh(\sqrt{1-\sigma_n^2}(x-\mu)/\sigma_n^2)}{\sigma_n^2},$$
$$\frac{\partial^2}{\partial \mu^2} \log f_\mu(x) = -\frac{1}{\sigma_n^2} + \frac{1-\sigma_n^2}{\sigma_n^4} \operatorname{sech}^2\left(\frac{\sqrt{1-\sigma_n^2}(x-\mu)}{\sigma_n^2}\right).$$

Thus $I_{\sigma_n} = -\mathbb{E}[\frac{\partial^2}{\partial \mu^2} \log f_\mu(X) \mid \mu]$, proving (40).

Substituting $u = \sqrt{1-\sigma_n^2}(x-\mu)/\sigma_n^2$ gives

$$\mathbb{E}\left[\operatorname{sech}^2\left(\frac{\sqrt{1-\sigma_n^2}(X-\mu)}{\sigma_n^2}\right) \mid \mu\right] = \frac{\sigma_n}{\sqrt{1-\sigma_n^2}\sqrt{2\pi}} e^{-(1-\sigma_n^2)/(2\sigma_n^2)} \int_{\mathbb{R}} e^{-\sigma_n^2 u^2/(2(1-\sigma_n^2))} \operatorname{sech} u \, du$$
$$= \sqrt{\frac{\pi}{2}} \frac{\sigma_n}{\sqrt{1-\sigma_n^2}} e^{-(1-\sigma_n^2)/(2\sigma_n^2)} \left(1 + \mathcal{O}(\sigma_n^2)\right).$$

The last equality follows from $\int_{\mathbb{R}} \operatorname{sech} u \, du = \pi$, $\int_{\mathbb{R}} u^2 \operatorname{sech} u \, du < \infty$, and $0 \le 1 - e^{-t} \le t$ for $t \ge 0$. Thus, we obtain (41).

For the last bound, reveal the mixture component labels and subtract the corresponding offsets, leaving n independent $\mathcal{N}(\mu, \sigma_n^2)$ observations conditional on μ. Revealing the labels cannot increase the location-minimax risk:

$$R^\star_{\mathrm{M},n} \ge \inf_{\widetilde{\mu}} \sup_{\mu \in \mathbb{R}} \mathbb{E}_{\mathbf{X} \sim \mathcal{N}(\mu, \sigma_n^2)^{\otimes n}}[(\widetilde{\mu}(\mathbf{X}) - \mu)^2] = \frac{\sigma_n^2}{n}.$$

See Proposition H.1. □

I ADDITIONAL EXPERIMENTS

Joint-training setup. We provide further training details for Figures 2 and 3. The joint-training experiments below use the same setup unless stated otherwise. We train the architecture of Definition 2.2 with both softmax MoE and GLU gates, updating θ and λ jointly while keeping $\boldsymbol{\omega}$ fixed. We use $m = 8000$ fixed training tasks with contexts of length $n = 150$ from Definition 2.1, with $\sigma_\mu^2 = 0.5$, $K = 2$, and $d = 10$. Each gradient-descent step uses all m training tasks, with constant step sizes 20 for θ and 100 for λ, for 600,000 steps. We initialize $\theta_k(0) \sim \mathcal{N}(0, 0.25^2 I_d)$ and sample the fixed gating queries $\omega_r \sim \mathcal{N}(0,1)$ independently. We draw the initial expert parameters independently and uniformly from $[0, 1.9]$, rejecting draws with separation below 0.6 and ordering them by magnitude. Each joint-training experiment uses ten independent runs and 200 validation tasks per family in each run. In the joint-training plots, experts are relabeled so that expert 1 is assigned to Gaussian contexts at the end of training. Curves show medians, and shaded regions span the minimum and maximum across runs.

I.1 JOINT TRAINING, GAUSSIAN-MIXTURE ALTERNATIVE

Figures 4 and 5 repeat Figures 2 and 3 with the Gaussian-mixture alternative and $\sigma_n^2 = 0.05$. The gates move toward $\bar{w} \approx I_2$ and $|\lambda_1|$ decreases to zero, as with the uniform alternative. Unlike with the uniform alternative, $|\lambda_2|$ settles near 2, above the stopping target $\log(1/\sigma_n^2)/(4\sqrt{1-\sigma_n^2}) \approx 0.77$ of Lemma 3.4. On Gaussian-mixture contexts, both architectures improve on the sample mean but remain above the component-mean error and the lower bound of Lemma H.2. The component mean assigns observations to components according to their position relative to the sample mean, subtracts the corresponding offset $\pm\sqrt{1-\sigma_n^2}$, and averages the resulting values.

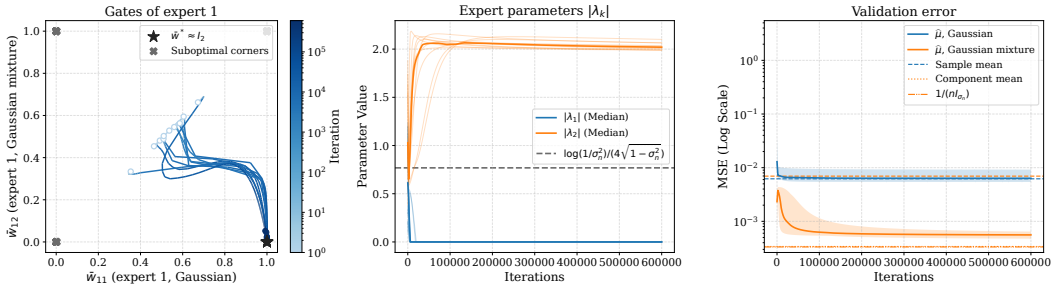

Figure 4: Softmax MoE, joint training with the Gaussian-mixture alternative, as in Figure 2. The dashed line in the middle panel marks $\log(1/\sigma_n^2)/(4\sqrt{1-\sigma_n^2})$.

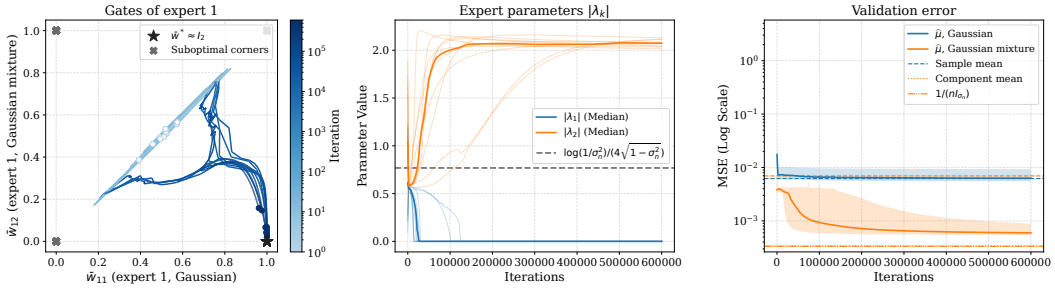

Figure 5: GLU, joint training with the Gaussian-mixture alternative, as in Figure 3. The dashed line in the middle panel marks $\log(1/\sigma_n^2)/(4\sqrt{1-\sigma_n^2})$.

I.2 Joint training: gating dynamics

Figures 6 and 7 show the GLU gates and normalization error during joint training. After an initial transient, the normalization error falls to about 10^{-2} before substantial family-specific selection develops. The gates then move toward $\bar{w} \approx I_2$, with less complete selection and greater variation across runs for the uniform alternative.

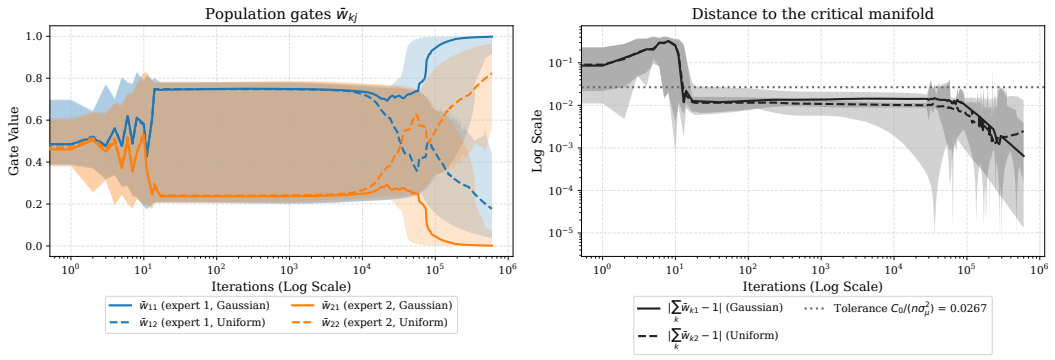

Figure 6: GLU, joint training with the uniform alternative: population gates $\bar{w}_{kj}$ (left) and normalization error $|\bar{w}_{1j} + \bar{w}_{2j} - 1|$ (right). Curves show medians, and shading spans all ten runs.

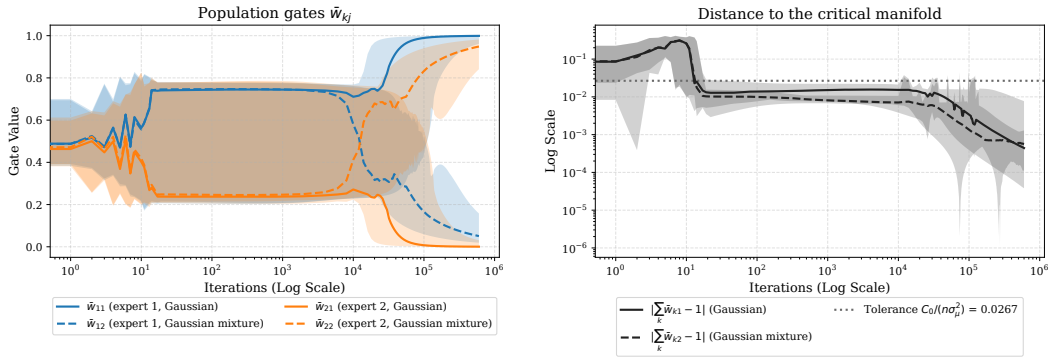

Figure 7: GLU, joint training with the Gaussian-mixture alternative, with panels and conventions as in Figure 6.

I.3 Stagewise training

We train the GLU with Definition 2.3, using the uniform alternative with $\sigma_\mu^2 = 0.05$ and $n = 500$ The values of m, K, d, the initialization, and the numbers of runs and validation tasks are as in the joint-training setup above. In the first stage, we minimize $\hat{\mathcal{L}}_m$ over θ using 20,000,000 gradient-descent steps with step size 100. In the second stage, we minimize $\hat{\mathcal{L}}_m$ over λ using 300,000 gradient-descent steps with step size 1000.

First stage: critical manifold. Figure 8 shows $(\bar{w}_{1j}, \bar{w}_{2j})$ during the first stage for Gaussian ($j = \mathrm{G}$) and uniform ($j = \mathrm{U}$) contexts. The population gates first approach $\bar{w}_{1j} + \bar{w}_{2j} \approx 1$, before substantial specialization toward e_1 and e_2 for Gaussian and uniform contexts, respectively.

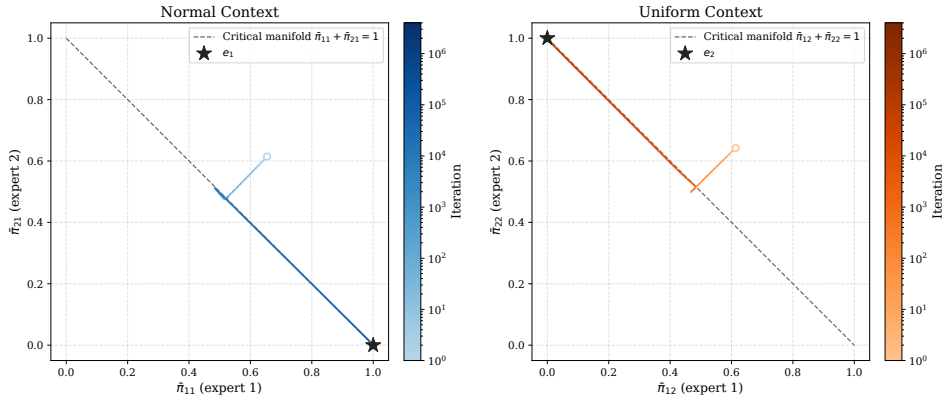

Figure 8: First stage with the uniform alternative: $(\bar{w}_{1j}, \bar{w}_{2j})$ for Gaussian (left) and uniform (right) contexts across runs, colored by iteration. The dashed line is the critical manifold $\bar{w}_{1j} + \bar{w}_{2j} = 1$.

Per-expert accuracy. Figure 9 shows $\mathbb{E}_j[(g_k - \mu)^2]$ for both experts and $\mathbb{E}_j[(\widehat{\mu} - \mu)^2]$ on validation tasks. In the first stage, λ is fixed, so the expert errors are constant and only the gating changes the error of $\widehat{\mu}$. On Gaussian contexts, expert 1 and $\widehat{\mu}$ stay near the sample-mean error. On uniform contexts, the errors of expert 2 and $\widehat{\mu}$ decrease further in the second stage toward the sample mid-range error. Although expert 2 becomes worse on Gaussian contexts during this stage, the error of $\widehat{\mu}$ remains approximately unchanged since $\bar{w}_{2\mathrm{G}} \approx 0$ after the first stage. The Gaussian-mixture results are in Section I.4.

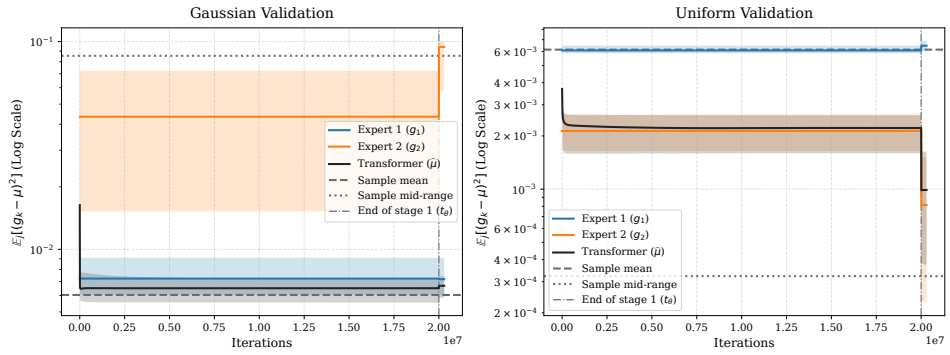

Figure 9: Uniform alternative: validation error of each expert g_k and of $\widehat{\mu}$ on Gaussian (left) and uniform (right) contexts, with the sample mean and sample mid-range as references. The vertical line marks t_θ.

I.4 Stagewise training, Gaussian-mixture alternative

We use the same Gaussian-mixture alternative as in the joint-training experiments, with the stagewise setup of Section I.3. Figure 10 repeats the uniform errors for comparison with Figure 12. In both cases, training the experts in the second stage reduces the alternative-family error while leaving the Gaussian error approximately unchanged. For the mixture, the error remains above the component-mean error and the lower bound of Lemma H.2. Figure 11 shows the first-stage gates moving toward the appropriate expert for each family, with a residual normalization error as in Figure 8.

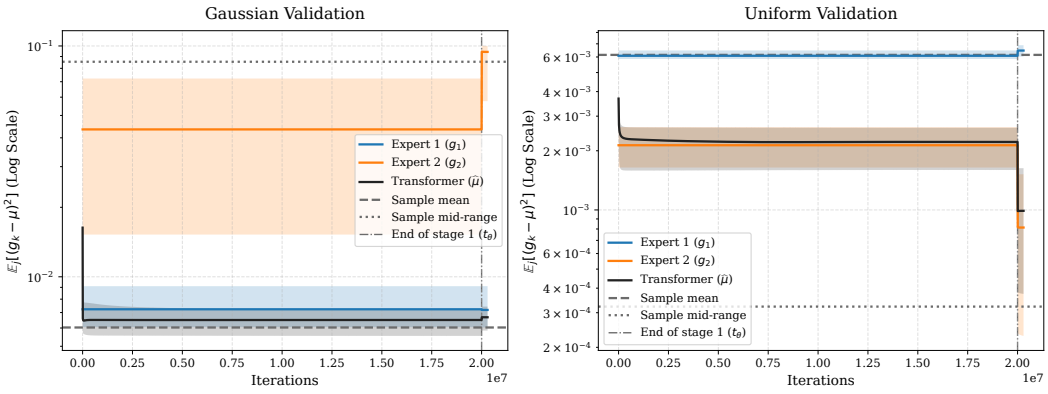

Figure 10: Uniform alternative: validation error of each expert g_k and of $\widehat{\mu}$ on Gaussian (left) and uniform (right) contexts, repeated for comparison with Figure 12. The vertical line marks t_θ.

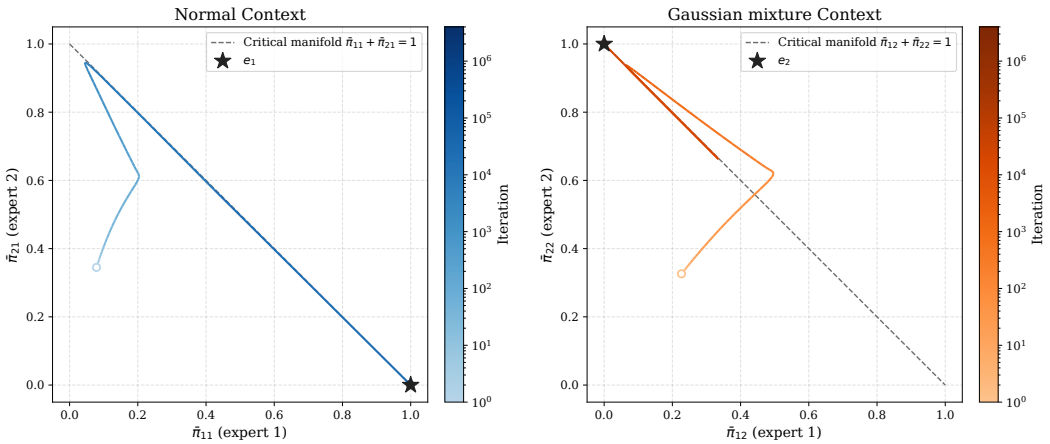

Figure 11: First stage with the Gaussian-mixture alternative: $(\bar{w}_{1j}, \bar{w}_{2j})$ for Gaussian (left) and Gaussian-mixture (right) contexts across runs, colored by iteration. The dashed line is the critical manifold $\bar{w}_{1j} + \bar{w}_{2j} = 1$.

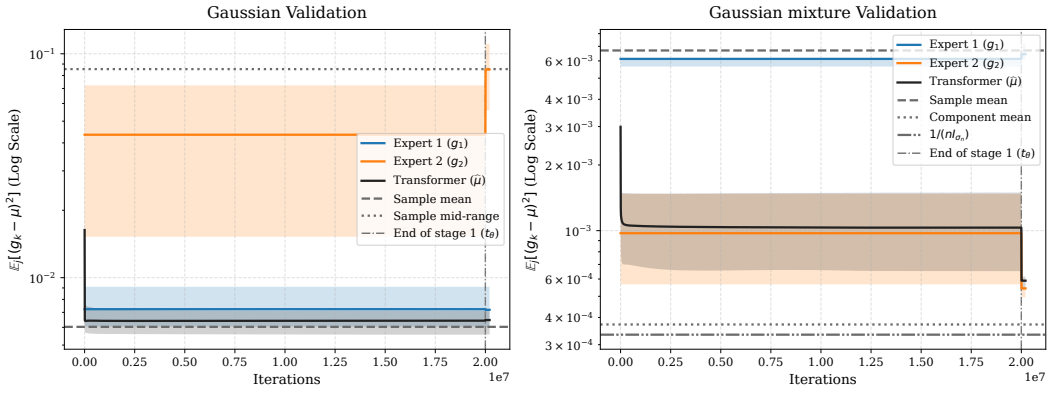

Figure 12: Gaussian-mixture alternative: validation error of each expert g_k and of $\widehat{\mu}$ on Gaussian (left) and Gaussian-mixture (right) contexts. The vertical line marks t_θ.

I.5 Relaxed architecture

We train ten attention parameters, whose attention outputs serve both as gating features and as inputs to the experts. The attention outputs are not centered or normalized. Each expert and each gate logit is a learned linear function of the ten attention outputs. We use the uniform alternative and train with SGD with $n = 200$, $m = 8{,}000$, learning rate 10, $K = 10$, and $\sigma_\mu^2 = 0.1$, and train all parameters jointly for 15,000 epochs. Each experiment uses five independent runs and 2,000 held-out validation tasks. Figure 13 shows end-to-end training of this less constrained architecture. The Gaussian validation error stays near the sample-mean error, while the uniform validation error falls well below it toward the sample mid-range error.

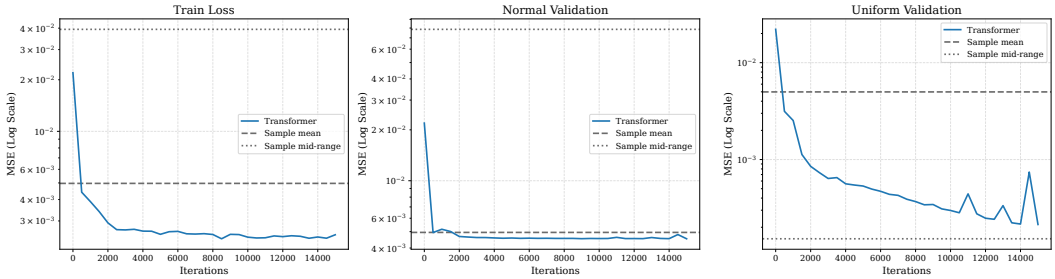

Figure 13: Less constrained architecture, end-to-end training with the uniform alternative. Left: training loss. Middle and right: validation error on Gaussian and uniform contexts, with the sample mean and sample mid-range as references.

J Role of Layer-Norm

LayerNorm enforces the condition $Z_{\mathrm{GA}} < \min\{Z_{\mathrm{GG}}, Z_{\mathrm{AA}}\}$ used in Lemma D.8. In our experiments, convergence is also observed without LayerNorm. Figure 14 illustrates the reduced flow studied in that lemma with and without LayerNorm. For the convergence argument, it would suffice to establish convergence to the optimal assignment from a region containing the reduced initial states at t_{crit}, with a margin for the approximation errors. Since the gates are initialized near $1/2$, a neighborhood of $(\bar{w}_{1\mathrm{G}}, \bar{w}_{1\mathrm{A}}) = (1/2, 1/2)$ containing these states could suffice. The stream plot suggests convergence from such a central region without LayerNorm. Under the strict expert hierarchy and $Z_{\mathrm{GA}} > Z_{\mathrm{AA}} > 0$, the reduced flow nevertheless has a region converging to the suboptimal corner $(0, 0)$: sufficiently close to this corner, when $\bar{w}_{1\mathrm{G}}/\bar{w}_{1\mathrm{A}}$ is sufficiently small, both coordinates and their ratio decrease, keeping the trajectory in this region and forcing both coordinates to zero. This does not preclude convergence from the central region.

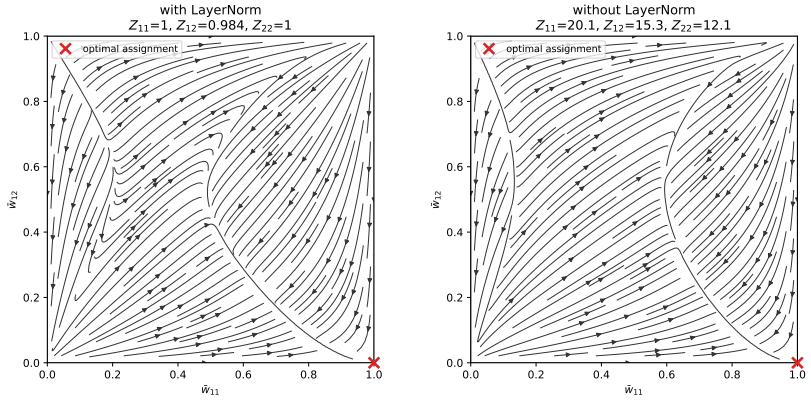

Figure 14: Stream plots of the reduced flow studied in Lemma D.8. Left: population features with LayerNorm. Right: uncentered population features at $\mu = 0$ without LayerNorm.

K Extended related work

In-context learning of statistical procedures. Prior-data fitted networks (PFNs) train transformers on synthetic tasks to approximate Bayesian prediction in context (Müller et al., 2024). Their adaptation to tabular data has produced models that perform strongly on standardized benchmarks, particularly for small and medium datasets (Hollmann et al., 2023; 2025; Erickson et al., 2025; Ye et al., 2025). Related pretrained models estimate heterogeneous treatment effects, survival curves, mutual information, and Poisson means (Balazadeh et al., 2025; Qi et al., 2026; Peyrard & Cho, 2025; Teh et al., 2025), and amortized simulation-based inference likewise maps simulated datasets to parameters (Cranmer et al., 2020). Most of these systems target prediction at a query point. These

works demonstrate that pretraining can produce effective predictors and parameter estimators. Our focus is on how gradient-based pretraining learns an estimator with rates that depend on the unknown family.

Statistical theory of pretrained predictors. Constructive and empirical analyses study which algorithms trained transformers implement in context (Akyürek et al., 2023; Chen et al., 2024b; Li et al., 2024). Bai et al. (2023) construct transformers that select among base algorithms, either by testing summary statistics of the context or by validating candidates on a held-out split, with guarantees for the empirical risk minimizer. Kim et al. (2024) show that a deep network followed by linear attention attains minimax rates for nonparametric regression over a known function class, and argue that representations fixed after pretraining limit adaptivity. Wakayama & Suzuki (2025) and Ma et al. (2025) analyze pretrained transformers as approximations to Bayesian posterior predictors (see also Nagler, 2023). In particular, Ma et al. (2025) use posterior contraction to obtain rates that adapt to unknown smoothness and effective dimension. Cannella et al. (2026) establish near-optimal regret for an idealized pretrained Bayes estimator in Poisson empirical Bayes, with concurrent guarantees for nonparametric Bayes estimators in the Gaussian case (Ignatiadis & Kankanala, 2026). These results show that the empirical risk minimizer, or an idealized Bayes predictor, can be adaptive. They do not follow the training trajectory, and they treat the adaptive map as a black box. In our model, adaptation is carried by an identifiable internal parameter, the attention scale of each expert, which moves the estimator across a family of estimators.

Training dynamics of attention. A complementary line analyzes how gradient-based training shapes attention on in-context tasks, including linear attention for linear regression (Zhang et al., 2024a;b), softmax attention (Huang et al., 2023), mean-field dynamics (Kim & Suzuki, 2024), and learning of low-dimensional structure (Oko et al., 2024). Closest to our analysis, Chen et al. (2024a) show that gradient flow on multi-head softmax attention allocates each head to one task of a multi-task linear regression, after warm-up and emergence phases. There, the task handled by each head is a fixed coordinate block determined by the task structure and the initialization. Here, the family varies across contexts and must be inferred from each sample, so specialization must be paired with learned routing. They also find that concentrated attention is suboptimal for noisy linear tasks, whereas the irregular uniform family calls for attention concentrated on the extremes. For the GLU, our gating dynamics separate into fast normalization and slow selection, which we analyze with geometric singular perturbation theory (Fenichel, 1979; Berglund, 2001).

Adaptive estimation. In statistics, an estimator is adaptive over a collection of models if, for each model, it attains the rate achievable when that model is known (Lepskii, 1991; Donoho & Johnstone, 1995; Tsybakov, 2009). Location estimation provides a classical parametric instance: for symmetric densities with finite Fisher information, the location can be estimated asymptotically as efficiently without knowing the shape as with it (van der Vaart & Wellner, 2021; Beran, 1974; Stone, 1975; Bickel, 1982; Bickel et al., 1993). Differences in rate between location families instead arise from irregularity: the uniform family has parameter-dependent support and admits the faster rate n^{-2} (Ibragimov & Has'minskii, 1981). Classical adaptive location estimators estimate the score function and apply a one-step correction, whereas the procedure learned here selects a point in a family of exponentially tilted averages.

Set architectures, gating, and mixtures of experts. Deep Sets (Zaheer et al., 2017) and Set Transformers (Lee et al., 2019) provide architectures for functions of unordered data, and our model is a small Set Transformer specialized to scalar estimation. Its second layer is either a softmax mixture of experts (Jacobs et al., 1991; Shazeer et al., 2017) or a gated linear unit (Dauphin et al., 2017; Shazeer, 2020). Comparing the two isolates the effect of enforcing a statistical constraint, namely normalization of the expert weights and hence exact translation equivariance, versus learning it. Bu et al. (2025) analyze how gradient-based training learns to retrieve task concepts from context. They show that training on question answer data enables this behavior, whereas training on their in-context demonstration distributions can fail. Their results concern factual recall, while we study how learned routing and expert specialization yield family-specific estimation rates.