TOWARDS UNDERSTANDING SPECIALIZATION IN MOE: A MEAN-FIELD PERSPECTIVE ON ROUTING–EXPERT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture of Experts (MoE) models allow the number of parameters to greatly increase while keeping the amount of computation for a given sample unchanged. Their benefits depend on a stable specialization among experts, which is shaped by the coupled dynamics of routing and expert modules. However, this specialization can become unstable or collapse, while existing theories struggle to capture this bidirectional feedback. We study a sparsely gated MoE model through mean-field theory, retaining each expert as a finite interacting unit while its overparameterized parameters evolve as a probability measure. This yields a hybrid finite dimensional and Wasserstein dynamical system. We provide a general decomposition of the expert and router dynamics and derive the evolution of comparative advantage. Our analyses further establish sufficient conditions for linear growth of specialization over a finite interval and routing collapse as an invariant set of the joint flow. Finally, we conduct experiments to verify our theoretical results. Our analysis offers a unified macroscopic account of specialization in MoE and provides guidance for scaling sparse models without sacrificing stability.

1 INTRODUCTION

Scaling up computational resources has been empirically proven to consistently improve the AI model, with the final performance following a power-law relationship to the invested resources (Kaplan et al., 2020). However, this computationally intensive approach is now limited by hardware efficiency limits, while the core challenge is to continue expanding the capacity and intelligence of the model without proportionally increasing the computational cost (Cai et al., 2025; Liu et al., 2026b). This has led to a shift from dense architectures to conditional computation, where internal units respond selectively to each input sample to improve training efficiency (Mu & Lin, 2025).

The Mixture of Experts (MoE) model has emerged as the most prominent instantiation of conditional computation (Dai et al., 2024; Wang et al., 2025; Cheng et al., 2026a; Huang et al., 2026). An MoE model replaces the dense layers of a conventional neural network with a set of expert modules and a trainable routing network, which activates only a sparse subset of experts for each input sample (Masoudnia & Ebrahimpour, 2014; Cai et al., 2024). Since this design decouples the total parameter count from the computational cost per inference, MoE has become a promising architecture for large-scale AI systems, allowing a capacity far exceeding that of dense models while keeping the inference budget nearly constant (Mu & Lin, 2025; Gan et al., 2025; Tan et al., 2026).

However, this advantage depends on whether different experts establish a stable division of labor, often referred to as specialization (Yuksel et al., 2012; Cai et al., 2025). When such a division emerges, each expert captures a distinct region of the input distribution, allowing the router to compose a small set of relevant operations for each sample. This can improve parameter efficiency (Li et al., 2025b; Cheng et al., 2026b) and make decisions made by individual experts more interpretable (Liu et al., 2026a). Since specialization is not automatically guaranteed by the MoE objective or by random initialization, understanding the conditions under which it emerges, stabilizes, or collapses is important for scaling sparse models (Liu et al., 2024; MAYAKI, 2026; Cheng et al., 2026a).

Despite a growing body of empirical work (Oldfield et al., 2024; Bai et al., 2025; He et al., 2026; Luo et al., 2026; Guo et al., 2026), the theoretical understanding of this phenomenon remains limited.

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.

Part of the previous theoretical work on MoE relies on the classic dynamical analysis of individual parameters under gradient descent (Chen et al., 2022; Li et al., 2025a), while another part is carried out from the perspective of Bayesian inference (Nguyen et al., 2023; Obando-Ceron et al., 2024; Thai et al., 2025). Although tractable under controlled assumptions, these approaches struggle to model the bidirectional interactions between the routing network and the expert modules.

In this paper, we adopt the Mean-Field Theory (MFT) (Mei et al., 2018; 2019) to study the dynamics underlying the specialization. We retain a finite set of interacting experts and a trainable router, while taking the mean-field limit within each overparameterized expert. This yields a hybrid finite-dimensional/Wasserstein dynamical system in which the router evolves in Euclidean parameter space and each expert evolves as a probability measure. From this macroscopic perspective, we can analyze the coupled routing–expert dynamics that tracking with finitely many parameters cannot easily capture, and extend the analysis to more general nonlinear expert models. We do not attempt a complete theory of specialization and center around a central dynamical question: how does the bidirectional feedback between routing and expert optimization generate, sustain, or destroy comparative advantages among experts? Therefore, the specialization is viewed as the macroscopic regime in which such comparative advantages grow and align with routing allocation.

To this end, we study an MoE model with sparse gating and analyze the dynamics. On the expert side, the dynamics decompose into a specialization force driven by tasks and a fluctuation term induced by routing noise. On the router side, they are split into a term for pairwise comparison of performance and a residual term. From this decomposition, we derive the exact evolution of the comparative advantage of one expert over another on a given task, showing that its growth is governed by task allocation and geometric task structure. We establish a sufficient condition for at least linear growth over a finite interval, which provides a dynamical signature of sustained specialization. In contrast, we characterize routing collapse as an invariant set of the joint flow in which the routing distribution degenerates to a single expert. Finally, we show how a standard regularizer for load balancing can destabilize this collapsed set, with a quantitative threshold above which the system may return to a specialization regime.

Our major contributions are as follows.

- We formulate the bidirectional feedback between routing and experts in a mean-field MoE, and establish two complementary quasi-static limits. We also construct a radial Gaussian model as a solvable case that instantiates these structures, which yields more accurate theoretical results.
- We derive the evolution equation for comparative advantage and show that the evolution is primarily governed by task allocation and task geometry. Building on this, we contrast and characterize two opposing states, namely sustained specialization and routing collapse. An MFT explanation is provided for how a regularization method can induce the transition between these two states.
- We conduct extensive numerical experiments to verify our theoretical predictions. The results demonstrate how routing-expert feedback amplifies comparative advantage into specialization. By contrast, highly concentrated routing suppresses this feedback and drives the system toward collapse. The regularization method can then reactivate the dynamics by destabilizing this state.

2 Related Work

2.1 Mixture of Experts

Expert specialization in MoE has been studied from both empirical and theoretical perspectives. Empirically, specialization is often promoted through routing and architectural design. A routing strategy named expert choice routing improves load allocation by allowing experts to select tokens (Zhou et al., 2022), while task-expert matching encourages modular specialization in multitask learning (Chen et al., 2023). More recent architectures further enhance specialization through fine-grained and shared experts (Dai et al., 2024), factorized expert parameterizations (Oldfield et al., 2024), cluster-aware initialization (Chu et al., 2026), or routing aligned with expert representation spaces (Cheng et al., 2026a). Theoretical studies can be broadly divided into optimization dynamics and statistical perspectives. Analyses grounded in traditional gradient descent dynamics characterize how routing partitions the data and how routing-expert interactions lead to expert differentiation depending on tasks (Chen et al., 2022; Li et al., 2025a). Complementary statistical studies investi-

gate parameter estimation, routing-induced local aggregation, and approximation estimation trade-offs (Nguyen et al., 2023; He et al., 2026). Despite these advances, existing theory typically focuses on particular training regimes or architectural mechanisms, leaving the coupled evolution of routing and expert distributions less understood.

2.2 Mean-Field Theory

Mean-field theory (MFT) provides a macroscopic description of neural networks by representing a large collection of interacting parameters through probability measures. Existing work establishes this representation and its asymptotic justification. For wide two-layer networks, the training dynamics can be described by the evolution of a parameter distribution (Mei et al., 2018), while particle-system formulations, laws of large numbers, and propagation-of-chaos results connect finite networks to their mean-field limits (Rotskoff & Vanden-Eijnden, 2018; Sirignano & Spiliopoulos, 2020). A complementary research direction focuses on optimization in the measure space. Gradient-based learning can be formulated as a Wasserstein gradient flow, allowing analyses of global convergence and the optimization landscape (Chizat & Bach, 2018). Related results provide dimension-independent approximation bounds, clarify connections with kernel regimes (Mei et al., 2019), and develop feature-based descriptions of neural gradient flows (Fang et al., 2021). These ideas have also been extended beyond shallow networks to multilayer architectures (Sirignano & Spiliopoulos, 2022). More recently, mean-field formulations have been adapted to Transformer architectures to characterize token or parameter dynamics in large-scale limits (Castin et al., 2025; Barboni et al., 2026). Building on these developments, we use MFT as a dynamical description of MoE expert populations and study their interaction with routing, with the goal of characterizing specialization, collapse, and load-balancing behavior within a unified framework.

3 Problem Setting and Preliminaries

This section establishes a general mean-field framework for analyzing the specialization dynamics of MoE. We consider M tasks, N experts, and a differentiable routing network, while abstracting away specific forms of data distributions, expert architectures, and routing parameterizations. A solvable special case with $M = 2$ and $N = 2$ under a radial Gaussian model is presented in Appendix A, which instantiates and verifies the general framework.

3.1 Task Distribution and MoE Model

Let there be $M \in \mathbb{N}$ tasks indexed by $i \in \{1, \dots, M\}$. Each task i is associated with a data distribution $\mathbb{P}_i$ over the sample space $\mathcal{X} \times \mathcal{Y}$, and a prior weight $k_i > 0$ with $\sum_{i=1}^{M} k_i = 1$. The overall data distribution is

$$\mathbb{P} = \sum_{i=1}^{M} k_i \, \mathbb{P}_i. \tag{1}$$

The MoE contains N experts $E_1, \dots, E_N$. For each expert j, let the parameter space be $\mathcal{W}_j \subseteq \mathbb{R}^{d_j}$. Under the infinite-width limit, the parameters of expert j are described by a probability measure $\rho_j \in \mathcal{P}_2(\mathcal{W}_j)$ with finite second moments. The output of expert j is

$$f_{\rho_j}(x) = \int_{\mathcal{W}_j} \phi_j(x; r) \, \rho_j(\mathrm{d}r), \tag{2}$$

where $\phi_j(x; r)$ denotes the output of a single neuron (or basis function) with a local parameter $r \in \mathcal{W}_j$. A finite-width model corresponds to the special case where ρ_j is an empirical measure. For a given sample (x, y), if it is processed by expert j, we denote the loss on this sample by $\ell_j(x, y; \rho_j)$, and refer to it as the sample loss of expert j. Here the dependence on ρ_j enters through the expert output. The task risk of expert j on task i is

$$R_i(\rho_j) := \mathbb{E}_{(x,y)\sim\mathbb{P}_i}\big[\ell_j(x, y; \rho_j)\big]. \tag{3}$$

Let the routing parameters be $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$. Given x, the router produces a probability vector

$$\pi(x; \theta) = \big(\pi_1(x; \theta), \dots, \pi_N(x; \theta)\big) \in \Delta^{N-1}, \tag{4}$$

where Δ^{N-1} denotes the simplex: $\pi_j(x; \theta) \geq 0$ and $\sum_{j=1}^{N} \pi_j(x; \theta) = 1$. The quantity π_j may represent either a soft routing weight or the conditional selection probability under Top-k routing.

3.2 JOINT RISK AND MEAN-FIELD DYNAMICS

Given the router and experts, the overall risk of the MoE is

$$\mathcal{R}(\theta,\rho)=\mathcal{R}(\theta,\rho_1,\dots,\rho_N)=\sum_{i=1}^{M}k_i\,\mathbb{E}_{\mathbb{P}_i}\Big[\sum_{j=1}^{N}\pi_j(x;\theta)\,\ell_j(x,y;\rho_j)\Big]. \tag{5}$$

To expose the coupling between tasks and experts, define the marginal assignment probability of task i to expert j:

$$P_{ij}(\theta):=\mathbb{E}_{\mathbb{P}_i}\big[\pi_j(x;\theta)\big],\sum_{j=1}^{N}P_{ij}(\theta)=1. \tag{6}$$

Further, define the sample selection covariance within tasks:

$$\mathcal{C}_{ij}(\theta,\rho_j):=\mathrm{Cov}_{\mathbb{P}_i}\big(\pi_j(x;\theta),\,\ell_j(x,y;\rho_j)\big)=\mathbb{E}_{\mathbb{P}_i}\big[\pi_j(x;\theta)\ell_j(x,y;\rho_j)\big]-P_{ij}(\theta)R_i(\rho_j). \tag{7}$$

Then the joint risk admits an exact decomposition into a term on a task level and a covariance residual:

$$\mathcal{R}(\theta,\rho)=\sum_{i=1}^{M}\sum_{j=1}^{N}k_iP_{ij}(\theta)\,R_i(\rho_j)+\sum_{i=1}^{M}\sum_{j=1}^{N}k_i\,\mathcal{C}_{ij}(\theta,\rho_j). \tag{8}$$

For each expert j, define the normalized load and conditional task composition as

$$m_j(\theta):=\sum_{i=1}^{M}k_iP_{ij}(\theta),\quad \zeta_{ij}(\theta):=\frac{k_iP_{ij}(\theta)}{m_j(\theta)}(m_j>0), \tag{9}$$

so that $\sum_i\zeta_{ij}=1$. These quantities will be used in the decomposition of the expert dynamics.

The routing parameters evolve by Euclidean gradient flow, while the expert measures evolve by Wasserstein gradient flow:

$$\dot\theta=-\eta_\theta\nabla_\theta\mathcal{R},\partial_t\rho_j=\eta_j\,\nabla_r\cdot\left(\rho_j\,\nabla_r\frac{\delta\mathcal{R}}{\delta\rho_j}\right),j=1,\dots,N. \tag{10}$$

To reveal how expert performance feeds back into the router, define for each task i the pairwise comparative advantage of expert j over expert k as

$$\mathcal{A}_i^{j\succ k}(\rho):=R_i(\rho_k)-R_i(\rho_j). \tag{11}$$

A positive value indicates that expert j outperforms expert k on task i.

For subsequent analysis, we introduce several auxiliary mean-field quantities. For each expert j, define the task gradient field for task i and the Wasserstein velocity of expert j as

$$g_i(\rho_j;r):=\nabla_r\frac{\delta R_i}{\delta\rho_j}(r),\quad v_j(\rho_j;r):=-\eta_j\nabla_r\frac{\delta\mathcal{R}}{\delta\rho_j}. \tag{12}$$

The Gram matrix of these task gradients under the $L^2(\rho_j)$ inner product is

$$G_{iu}^{(j)}(\rho_j):=\int_{\mathcal{W}_j}g_i(\rho_j;r)\cdot g_u(\rho_j;r)\,\rho_j(\mathrm{d}r)=\big\langle g_i(\rho_j),\,g_u(\rho_j)\big\rangle_{L^2(\rho_j)}. \tag{13}$$

This matrix captures the geometric alignment between different task gradients within the parameter space of expert j. Additionally, define the gradient field of the sample selection covariance as

$$\mathcal{C}_{uj}^{(g)}(\theta,\rho_j;r):=\nabla_r\frac{\delta\mathcal{C}_{uj}}{\delta\rho_j}(\theta,\rho_j;r), \tag{14}$$

which can be equivalently written as

$$\mathcal{C}_{uj}^{(g)}(\theta,\rho_j;r)=\mathbb{E}_{\mathbb{P}_u}\Big[\pi_j(x;\theta)\,\nabla_r\frac{\delta\ell_j}{\delta\rho_j}(x,y;\rho_j;r)\Big]-P_{uj}(\theta)\,g_u(\rho_j;r). \tag{15}$$

The $L^2(\rho_j)$ norm of this covariance gradient satisfies

$$\|\mathcal{C}_{uj}^{(g)}\|_{L^2(\rho_j)}\le\sigma_{uj}^{\pi}\,\sigma_{uj}^{g}, \tag{16}$$

where

$$\sigma_{uj}^{g}:=\big(\mathbb{E}_{\mathbb{P}_u}\Big\|\nabla_r\frac{\delta\ell_j}{\delta\rho_j}-g_u(\rho_j)\Big\|_{L^2(\rho_j)}^2\big)^{1/2},\sigma_{uj}^{\pi}:=\big(\mathbb{E}_{\mathbb{P}_u}|\pi_j(x;\theta)-P_{uj}(\theta)|^2\big)^{1/2}. \tag{17}$$

These bounds quantify the interaction between routing variability and gradient fluctuations.

4 MAIN RESULTS

This section presents the main analytical results for the general framework with M tasks and N experts illustrated in Section 3. We first establish the fundamental energy dissipation law that governs the coupled system and derive the key dynamical equations that link the evolution of expert performance to the task allocation matrix. Then, compare the two opposing states of sustained specialization and routing collapse, subsequently elaborating on the sufficient conditions for their formation and their impacts. Finally, we give a mean-field explanation of how the regularization strategy can mitigate collapse and promote specialization in this process.

Before formal discussion, we note that while the mean-field approach yields only qualitative conclusions when the coupling is strong, the model can become more tractable under certain conditions during the training phase. In particular, when the optimization rate of either the routing mechanism or the expert modules substantially outpaces that of the counterpart, the system reduces to a quasi-static scenario (Tkachenko, 2025; Zhao et al., 2026), which makes it easier to solve and allows explicit error estimates (see the special case in Appendix A for details). We state the following propositions, and the proofs are detailed in Appendices C.10 and C.11.

Proposition 4.1. *Consider a singularly perturbed system*

$$\epsilon\dot{\theta} = -\nabla_\theta \mathcal{R}(\theta, \rho), \partial_t \rho_j + \nabla_r \cdot (\rho_j v_j(\theta, \rho)) = 0. \tag{18}$$

Assume that 1. there exist $\theta_*(\rho)$ within the compact region and $\kappa_\theta > 0$ such that $\nabla_\theta \mathcal{R}(\theta_*(\rho), \rho) = 0, \langle \theta - \theta_*(\rho), \nabla_\theta \mathcal{R}(\theta, \rho)\rangle \geq \kappa_\theta \|\theta - \theta_*(\rho)\|_2^2$ holds within the unified local attractor domain of the branch. 2. $t \mapsto \theta_*(\rho(t))$ is absolutely continuous and $\|\frac{d}{dt}\theta_*(\rho(t))\|_2 \leq C_\theta$ 3. the expert velocity field is uniformly Lipschitz with respect to θ, then

$$\|\theta(t) - \theta_*(\rho(t))\|_2 \leq e^{-\kappa_\theta t/\epsilon}\|\theta(0) - \theta_*(\rho(0))\|_2 + \frac{C_\theta}{\kappa_\theta}\epsilon. \tag{19}$$

Proposition 4.2. *Consider a singularly perturbed system*

$$\dot{\theta} = -\eta_\theta \nabla_\theta \mathcal{R}(\theta, \rho), \epsilon \partial_t \rho_j = \eta_j \nabla_r \cdot \left(\rho_j \nabla_r \frac{\delta \mathcal{R}}{\delta \rho_j}\right). \tag{20}$$

Assume that 1. there exists a stable expert equilibrium branch $\rho_*(\theta)$ such that after freezing any θ, the expert flow is exponentially attracted to $\rho_*(\theta)$ within the unified attractor domain at a rate of at least κ_ρ below $d_{\mathcal{W}}$. 2. this branch is uniformly Lipschitz with respect to θ, $\|\dot{\theta}\|_2 \leq K_\theta$. 3. $\nabla_\theta \mathcal{R}(\theta, \rho)$ is Lipschitz with respect to ρ under $d_{\mathcal{W}}$, where the Wasserstein distance is defined as

$$d_{\mathcal{W}}(\rho, \widetilde{\rho}) := \big(\sum_{j=1}^{N} W_2^2(\rho_j, \widetilde{\rho}_j)\big)^{1/2}. \tag{21}$$

Then, the slow routing dynamics can be reduced to

$$\dot{\theta} = -\eta_\theta \nabla_\theta \mathcal{R}(\theta, \rho_*(\theta)) + O(\epsilon). \tag{22}$$

4.1 ENERGY DISSIPATION AND COUPLED DECOMPOSITION

When the above special conditions are not met, we have the following general results. We begin by establishing the basic property of the joint gradient flow that the total risk is a non-increasing Lyapunov function. This result confirms the mathematical consistency of the dynamics and clarifies the setting in which subsequent specialization or collapse occurs.

Proposition 4.3. *Suppose that the joint gradient flow in Eq. (10) satisfies the regularity and no-flux boundary conditions. Then*

$$\frac{d}{dt}\mathcal{R} = -\eta_\theta \|\nabla_\theta \mathcal{R}\|_2^2 - \sum_{j=1}^{N} \eta_j \int_{\mathcal{W}_j} \left\| \nabla_r \frac{\delta \mathcal{R}}{\delta \rho_j}(r) \right\|_2^2 \rho_j(dr) \leq 0. \tag{23}$$

The proof of Proposition 4.3 is detailed in Appendix C.12, which confirms that the training risk will never increase. However, a decreasing risk alone does not guarantee that experts will specialize in ways that produce an effective division of labor. Next, we decompose the gradients of the coupled system to reveal what drives the evolution of experts and routers.

Proposition 4.4. *For the joint risk in Eq. (5) and each expert with $m_j > 0$, the functional derivative of the risk with respect to the expert measure is given by*

$$\frac{\delta \mathcal{R}}{\delta \rho_j} = m_j \frac{\delta}{\delta \rho_j} \left[\sum_{i=1}^{M} \zeta_{ij} R_i(\rho_j) \right] + \sum_{i=1}^{M} k_i \frac{\delta \mathcal{C}_{ij}}{\delta \rho_j}. \tag{24}$$

For each task i, choose an arbitrary expert b_i, then the routing dynamics are governed by

$$\dot{\theta} = \eta_\theta \sum_{i=1}^{M} k_i \sum_{j \neq b_i} \nabla_\theta P_{ij}(\theta)\, \mathcal{A}_i^{j \succ b_i}(\rho) - \eta_\theta \sum_{i=1}^{M} \sum_{j=1}^{N} k_i\, \nabla_\theta \mathcal{C}_{ij}(\theta, \rho_j). \tag{25}$$

The proof of Proposition 4.4 is detailed in Appendix C.13. The first term in Eq. (24) describes how task allocation shapes expert learning, while the second term captures fluctuations within tasks. Similarly, Eq. (25) shows that routing is driven by the relative performance of experts and sample-dependent effects. This decomposition offers an intuitive way to interpret MoE training. A decrease in overall risk alone does not necessarily mean that experts will specialize, and the emergence of a meaningful division of labor depends on whether routing preferences and expert performance reinforce each other. The subsequent sections will make the feedback mechanism more explicit.

4.2 Dynamics of Comparative Advantage

Building on the decomposition above, we introduce the comparative advantage to measure how much better one expert is than another at specializing in a given task. To study the evolution in a setting where experts are comparable, we assume that all experts have the same parameter space $\mathcal{W}$ and learning rate η_E. This allows defining a unified task gradient field and its associated Gram matrix, which encodes the geometric relationship between different tasks. The following result provides the exact evolution equation for the comparative advantage, exposing the role of the assignment matrix P_{uj} and the geometry of the task gradient.

Proposition 4.5. *Assume that the expert measures evolve along the Wasserstein gradient flow and $\theta(t)$ follows the routing dynamics in Eq. (25). Then $\mathcal{A}_i^{j \succ k}$ satisfies the evolution equation*

$$\dot{\mathcal{A}}_i^{j \succ k} = \eta_E\, \mathcal{D}_i^{j \succ k} + \mathcal{X}_i^{j \succ k}, \tag{26}$$

where $\mathcal{D}_i^{j \succ k} := \sum_{u=1}^{M} k_u [P_{uj}(\theta) G_{iu}(\rho_j) - P_{uk}(\theta) G_{iu}(\rho_k)]$, and the residual term from sample selection is $\mathcal{X}_i^{j \succ k} := \eta_E \sum_{u=1}^{M} k_u [\langle g_i(\rho_j), \mathcal{C}_{uj}^{(g)}(\theta, \rho_j) \rangle_{L^2(\rho_j)} - \langle g_i(\rho_k), \mathcal{C}_{uk}^{(g)}(\theta, \rho_k) \rangle_{L^2(\rho_k)}]$.

Moreover, the residual term satisfies the upper bound $|\mathcal{X}_i^{j \succ k}| \leq \eta_E\, \mathcal{B}_i^{j \succ k}$, with $\mathcal{B}_i^{j \succ k}$ defined as

$$\mathcal{B}_i^{j \succ k} := \sum_{u=1}^{M} k_u [\|g_i(\rho_j)\|_{L^2(\rho_j)} \sigma_{uj}^{\pi} \sigma_{uj}^{g} + \|g_i(\rho_k)\|_{L^2(\rho_k)} \sigma_{uk}^{\pi} \sigma_{uk}^{g}]. \tag{27}$$

The proof of Proposition 4.5 is given in Appendix C.14. This decomposition separates the evolution of comparative advantage into a term determined by routing allocation and task gradient geometry, and a bounded residual induced by sample selection within a task. More importantly, the former can be bounded by quantities that have clear interpretations in terms of routing and gradient geometry, which yields the following sufficient condition for sustained specialization.

Proposition 4.6. *Fix a task i, a pair of experts $\{j, k\}$, and an interval $I = [t_0, t_1]$. Suppose that there exist constants $a, g_{\min}, g_{\max} > 0$ and $\gamma \geq 0$ such that for all $t \in I$ and $l \in \{j, k\}$,*

$$P_{ij}(\theta) - P_{ik}(\theta) \geq a, G_{ii}^{(l)} \geq g_{\min}, G_{uu}^{(l)} \leq g_{\max}$$

and for each $u \neq i$, $|G_{iu}^{(l)}| \leq \gamma g_{\max}$. Then

$$\mathcal{D}_i^{j \succ k} \geq k_i a g_{\min} - k_i P_{ik}(g_{\max} - g_{\min}) - \gamma g_{\max} \sum_{u \neq i} k_u (P_{uj} + P_{uk}). \tag{28}$$

Consequently, if there exists $\nu > 0$ such that for all $t \in I$,

$$k_i a g_{\min} - k_i P_{ik}(g_{\max} - g_{\min}) - \gamma g_{\max} \sum_{u \neq i} k_u (P_{uj} + P_{uk}) - \mathcal{B}_i^{j \succ k} \geq \nu, \tag{29}$$

then

$$\dot{\mathcal{A}}_i^{j \succ k} \geq \eta_E \nu > 0, \mathcal{A}_i^{j \succ k}(t) \geq \mathcal{A}_i^{j \succ k}(t_0) + \eta_E \nu (t - t_0). \tag{30}$$

The proof of Proposition 4.6 is given in Appendix C.15. This result demonstrates that an initial routing preference can be amplified into a persistent performance advantage when the task-specific learning signal dominates the mismatch between tasks and experts, the interference between tasks, and the fluctuations within samples. In practical terms, specialization is therefore not determined by routing alone. It also depends on the extent to which the routed tasks have compatible learning directions. This offers a simple explanation for why the same routing mechanism may lead to a clear specialization for some task structures while yielding weak differentiation for others.

4.3 Routing Collapse and Regularization

The preceding section describes a sufficient condition for sustained specialization. However, Proposition 4.4 also points to a possible failure mode. If routing becomes highly concentrated on a single expert, the positive feedback loop that drives specialization can be broken, which could lead to the phenomenon known as routing collapse. We next characterize this mode and then analyze the most common regularization method.

Proposition 4.7. *Suppose that there exist an expert j^* and $\epsilon \in [0, 1)$ such that, for each task i and $j \neq j^*$, $P_{ij^*}(\theta) \geq 1 - \epsilon$ and $P_{ij}(\theta) \leq \epsilon$. If there exist constants $C_\pi, B_\ell, B_g > 0$ such that $\forall i, j$,*

$$\|\nabla_\theta \pi_j(x;\theta)\| \leq C_\pi \epsilon, |\ell_j(x, y; \rho_j)| \leq B_\ell, \|\nabla_r \frac{\delta \ell_j}{\delta \rho_j}(x, y; \rho_j; \cdot)\|_{L^2(\rho_j)} \leq B_g, \ a.s. \tag{31}$$

then the routing velocity is bounded by $\|\dot{\theta}\| \leq \eta_\theta N C_\pi B_\ell \epsilon$. Moreover, we have

$$\|v_{j^*} + \eta_{j^*} \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i}\big[\nabla_r \frac{\delta \ell_{j^*}}{\delta \rho_{j^*}}\big]\|_{L^2(\rho_{j^*})} \leq \eta_{j^*} B_g \epsilon, \tag{32}$$

and $\max_{j \neq j^*} \|v_j\|_{L^2(\rho_j)} \leq \eta_{\max} B_g \epsilon, \eta_{\max} := \max_{j \neq j^*} \eta_j$.

The proof of Proposition 4.7 is detailed in Appendix C.16. For small ϵ, the system remains for a long time in a regime where one single expert dominates. This expert continues to learn from the mixture, which may alter its comparative advantage. However, this change has negligible impact on the router, with the other experts nearly stopping the update. Consequently, a strong specialization will not emerge quickly. Instead, the system remains in a collapsed routing regime with a highly imbalanced workload. When $\epsilon = 0$, the proposition reduces to Lemma D.3, which indicates that the trajectory will remain on the collapsed invariant manifold. A standard remedy is to incorporate a regularizer for load balancing. The following proposition shows that the regularizer can destabilize the collapse boundary, thereby making the collapsed state unstable to perturbations and potentially allowing the system to return to a regime where specialization re-emerges.

Proposition 4.8. *Define the load imbalance measure $A(\theta) := \frac{1}{2}\sum_{j=1}^{N}(m_j(\theta) - 1/N)^2$ and consider the regularized risk $\mathcal{R}_\lambda = \mathcal{R} + \lambda A$. If $\nabla_\theta A(\theta_c) \neq 0$ at a collapsed fixed point θ_c of the original system, then this point is no longer a fixed point after load balancing is introduced. If instead $\nabla_\theta A(\theta_c) = 0$ and there exists a unit direction u such that $u^\top \nabla^2_\theta A(\theta_c) u < 0$, then for*

$$\lambda > \frac{\max\{u^\top \nabla^2_\theta \mathcal{R}(\theta_c, \rho) u, \, 0\}}{-u^\top \nabla^2_\theta A(\theta_c) u}, \tag{33}$$

the regularized risk has negative curvature in direction u. Consequently, this collapsed fixed point is no longer a locally stable minimum of the gradient flow.

The proof of Proposition 4.8 is provided in Appendix C.17. The result demonstrates that load balancing can alter the local stability of a collapsed routing state. It can either move the system away from the original fixed point or, when regularization is sufficiently strong, introduce an unstable direction. Once this state has been destabilized, the router is no longer trapped near a highly concentrated routing pattern and can again respond to performance differences among experts. This confers on load balancing a dynamical interpretation that extends beyond simply balancing expert usage. In particular, load balancing can restore the routing-expert feedback suppressed near collapse, thereby enabling specialization to re-emerge.

5 NUMERICAL EXPERIMENTS

We design a series of experiments, each testing a theoretical prediction in Section 4. Due to space limitations, we verify part of the main results in this section, while the remaining results are deferred to Appendix B.1.

5.1 EXPERIMENTAL SETUP

Synthetic Tasks and MoE Model. Unless otherwise specified, we consider $M = N = 4$ tasks and experts with equal task prior $k_i = 1/4$. The input dimension is $d = 8$. For task i, samples are generated according to $x \sim \mathcal{N}(\mu_i, \sigma_x^2 I), \mu_i = \delta e_i$, where δ controls the separation between task means. The regression target is $y = 1.3 \tanh(v_i^\top x) + \xi, \xi \sim \mathcal{N}(0, 0.03^2)$, with $v_i = \sqrt{1-s}\, e_i + \sqrt{s}\, e_1$, where $s \in [0, 1]$ controls the similarity between the task directions. Hence, increasing s introduces a larger shared component among the task predictors and changes the cross-task gradient geometry appearing in Proposition 4.6. Each expert is a two-layer network of the form $f_j(x) = \frac{1}{H} \sum_{h=1}^{H} a_{jh} \tanh(w_{jh}^\top x)$, which can be viewed as a particle approximation of the expert measure ρ_j in the mean-field formulation. The router is a standard linear softmax model

$$\pi_j(x; \theta) = \frac{\exp((u_j^\top x + b_j)/\tau)}{\sum_{k=1}^{N} \exp((u_k^\top x + b_k)/\tau)}, \tag{34}$$

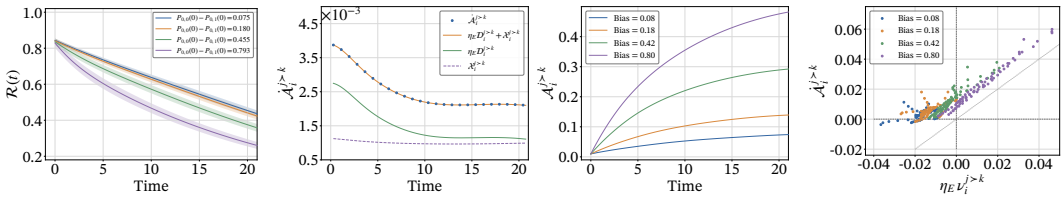
with τ denoting the routing temperature and b_j denoting the routing bias. The coupled dynamics are simulated by an explicit Euler discretization of the gradient flows. Unless otherwise stated, the step size is $\Delta t = 0.035$, the learning-rate coefficient for the expert is $\eta_E = 0.16$, the learning-rate coefficient for the router is $\eta_\theta = 0.06$, and each experiment is run for 600 update steps. We use the standard squared loss, and the statistics are reported on five random seeds.

Experimental Protocols. To verify Proposition 4.3, we record the empirical training risk along the coupled optimization trajectory under four different initial routing biases. Importantly, the risk is evaluated on the same empirical distribution used for gradient updates, so that the experiment directly corresponds to the Lyapunov function in Eq. (23). Next, we verify the exact dynamics of comparative advantage in Proposition 4.5. We numerically estimate $\dot{\mathcal{A}}_i^{j \succ k}$ using centered finite differences and compare it with $\eta_E \mathcal{D}_i^{j \succ k} + \mathcal{X}_i^{j \succ k}$. Finally, we study the macroscopic specialization pattern produced by the routing-expert feedback, which corresponds to Proposition 4.6. We vary the initial routing bias and jointly train all experts and the router. At each recorded time point, we evaluate the empirical counterpart of the lower bound, and then examine both the final task-expert assignment matrix P_{ij} and the task risk matrix $R_i(\rho_j)$.

5.2 EXPERIMENTAL RESULTS

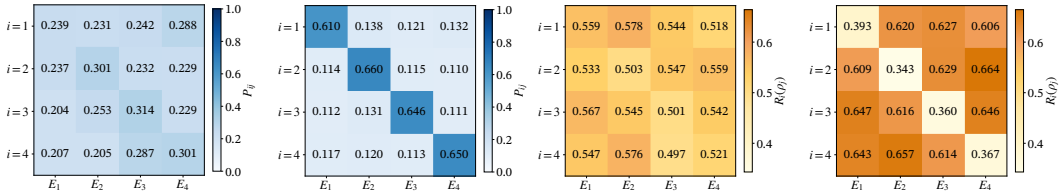
Energy dissipation and comparative advantage. We first verify the basic property of the joint system. Figure 1(a) reports the empirical training risk for four initial routing biases. In all cases, the risk decreases monotonically throughout optimization. This behavior is consistent with Proposition 4.3, which establishes $\mathcal{R}$ as a Lyapunov function of the joint routing-expert gradient flow. Next, we examine Proposition 4.5, which decomposes the evolution of comparative advantage into a term $\eta_E \mathcal{D}_i^{j \succ k}$ and a residual $\mathcal{X}_i^{j \succ k}$. Figure 1(b) compares the numerically evaluated derivative of $\mathcal{A}_i^{j \succ k}$ with the theoretical result on the right hand side, which shows that $\eta_E \mathcal{D}_i^{j \succ k} + \mathcal{X}_i^{j \succ k}$ closely follows the numerical derivative.

Sufficient condition for sustained specialization. Having verified the exact evolution equation, we next test whether the lower bound in Proposition 4.6 correctly identifies the regimes in which comparative advantage is guaranteed to increase. Figure 1(c) shows the evolution of $\mathcal{A}_i^{j \succ k}$ under different initial routing biases. A stronger task-specific routing bias produces a more pronounced separation between the expert favored by the routing bias and its competitor, consistent with the mechanism predicted by Proposition 4.6. Figure 1(d) compares the empirical derivative $\dot{\mathcal{A}}_i^{j \succ k}$ with the theoretical lower bound. Whenever the lower bound is positive, the measured derivative remains above the corresponding lower bound value. This is the finite-sample analogue of $\dot{\mathcal{A}}_i^{j \succ k} \geq \eta_E \nu > 0$.



(a) (b) (c) (d)

Figure 1: (a) Energy dissipation curves under different initial biases, where the shaded region indicates variation across random seeds; (b) Numerical estimates of $\dot{\mathcal{A}}_i^{j\succ k}$ are compared with $\eta_E \mathcal{D}_i^{j\succ k} + \mathcal{X}_i^{j\succ k}$ and the two components; (c) Evolution of comparative advantage; (d) Empirical $\dot{\mathcal{A}}_i^{j\succ k}$ versus the theoretical lower bound.



(a) (b) (c) (d)

Figure 2: Macroscopic task-expert structure after joint training. (a)(b): final assignment matrices P_{ij} under weak and strong initial routing bias. (c)(d): corresponding task risk matrices $R_i(\rho_j)$.

Emergence of task-expert specialization. Figure 2 compares the final assignment and task risk matrices for weak and strong initial routing biases. A weak initial bias leaves the assignment less structured and weakens the task-expert separation, whereas a stronger bias yields a clearer diagonal structure in P_{ij}. Experts that are assigned more samples from task i achieve lower risk on that task than their competitors. Diagonal routing alone does not constitute specialization. The joint emergence of routing preference and task-specific advantage is consistent with the mechanism in Propositions 4.4, 4.5 and 4.6. Routing shapes an effective task composition for each expert, and performance differences feed back into routing.

Summary. The numerical results support the main findings in Section 4. The joint objective first induces shared dissipative dynamics, but energy descent alone cannot determine the structure of routing. When the allocation specific to each task and the performance of experts strengthen each other, a comparative advantage follows the decomposition in Proposition 4.5 and can enter the sustained specialization regime described in Proposition 4.6. In contrast, once the routing distribution becomes highly concentrated on a single expert, both the adaptation and the learning of other experts are suppressed on a scale $O(\epsilon)$. This can lead to persistent proximity to collapse in Proposition 4.7. Load balancing addresses this by destabilizing the collapsed routing state, so that the system can regain routing flexibility and may return to the specialization regime.

6 Conclusion

We develop a mean-field theory of specialization in an MoE model with sparse gating. By coupling a finite trainable router with experts represented by measures, we obtained a hybrid Euclidean–Wasserstein flow that exposes the bidirectional feedback between routing and expert optimization. We decompose the expert and router dynamics and derive the evolution of pairwise comparative advantage in terms of task allocation and task gradient geometry. This yields a sufficient condition for sustained specialization, characterizes routing collapse as an invariant set, and shows that load balancing can destabilize it above a quantitative threshold. Extensive numerical experiments are conducted to support the theory. From this theoretical perspective, we conclude that specialization therefore emerges as a macroscopic dynamical regime of coupled allocation and performance feedback, rather than an automatic consequence of sparsity.

AI USE STATEMENT

In this work, generative AI tools were utilized exclusively for assisting with initial language proofreading and LaTeX template formatting. All core conceptual ideas, theoretical derivations, algorithm designs, code implementations, and experimental formulations were independently conceived, verified, and authored by the researchers. The authors take full responsibility for the entirety of the content, claims, and formulations presented in this paper.

ETHICS STATEMENT

This research focuses on foundational machine learning algorithms. All datasets used are standard, publicly available academic benchmarks. The proposed method does not generate harmful content, involve sensitive personal data, or introduce societal bias.

REFERENCES

Jun Bai, Minghao Tong, Yang Liu, Zixia Jia, and Zilong Zheng. Understanding and leveraging the expert specialization of context faithfulness in mixture-of-experts llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21938–21953, 2025.

Raphaël Barboni, Maarten V de Hoop, Takashi Furuya, and Gabriel Peyré. Training infinitely deep and wide transformers. *arXiv preprint arXiv:2605.17660*, 2026.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. 2024.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):3896–3915, 2025.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G Learned-Miller, and Chuang Gan. Mod-squad: Designing mixtures of experts as modular multi-task learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11828–11837, 2023.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062, 2022.

Anzhe Cheng, Shukai Duan, Shixuan Li, Chenzhong Yin, Mingxi Cheng, Heng Ping, Tamoghna Chattopadhyay, Sophia I Thomopoulos, Shahin Nazarian, Paul Thompson, et al. Ermoe: Eigen-reparameterized mixture-of-experts for stable routing and interpretable specialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12997–13006, 2026a.

Xin Cheng, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Han Zhang, Yu-Kun Li, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. Conditional memory via scalable lookup: A new axis of sparsity for large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4968–4990, San Diego, California, United States, July 2026b. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.226. URL https://aclanthology.org/2026.acl-long.226/.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Sanghyeok Chu, Pyunghwan Ahn, Gwangmo Song, Seung Hwan Kim, Honglak Lee, and Bohyung Han. Enhancing mixture-of-experts specialization via cluster-aware upcycling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11283–11292, 2026.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

Cong Fang, Jason Lee, Pengkun Yang, and Tong Zhang. Modeling from features: a mean-field framework for over-parameterized deep neural networks. In *Conference on learning theory*, pp. 1887–1936. PMLR, 2021.

Wensheng Gan, Zhenyao Ning, Zhenlian Qi, and Philip S Yu. Mixture of experts (moe): A big data perspective. *Information Fusion*, pp. 103664, 2025.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, et al. Advancing expert specialization for better moe. *Advances in Neural Information Processing Systems*, 38:48767–48809, 2026.

Siyuan He, Bokai Yang, Jie Hu, Ziwen Gao, and Yuhong Yang. Towards a statistical understanding of mixture-of-experts. *arXiv preprint arXiv:2609.03501*, 2026.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, Rongrong Ji, and Liujuan Cao. Discovering important experts for mixture-of-experts models pruning through a theoretical perspective. *Advances in Neural Information Processing Systems*, 38:135973–136003, 2026.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness Shroff. Theory on mixture-of-experts in continual learning. In *International Conference on Learning Representations*, volume 2025, pp. 8169–8206, 2025a.

Junzhuo Li, Bo Wang, Xiuze Zhou, Peijie Jiang, Jia Liu, and Xuming Hu. Decoding knowledge attribution in mixture-of-experts: A framework of basic-refinement collaboration and efficiency analysis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22431–22446, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1093. URL https://aclanthology.org/2025.acl-long.1093/.

Bo Liu, Di Dai, Jingwei Liu, Jiarui Jin, Xiaocheng Fang, Guangkun Nie, Hongyan Li, and Shenda Hong. Causalmoe: A billion-scale multimodal foundation model for granger causal discovery with pattern-routed heterogeneous experts. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11422–11431, 2026a.

Jiacheng Liu, Peng Tang, Wenfeng Wang, Yuhang Ren, Xiaofeng Hou, Pheng Ann Heng, Minyi Guo, and Chao Li. A survey on inference optimization techniques for mixture of experts models. *ACM Computing Surveys*, 58(10):1–37, 2026b.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024.

Lirui Luo, Guoxi Zhang, Hongming Xu, Cong Fang, and Qing Li. Sphere: Mitigating the loss of spectral plasticity in mixture-of-experts for deep reinforcement learning. *arXiv preprint arXiv:2605.04712*, 2026.

Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014.

Mansour MAYAKI. Generalization and scaling laws for mixture-of-expertstransformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WxmL5UjtNm.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pp. 2388–2464. PMLR, 2019.

Siyuan Mu and Sen Lin. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*, 2025.

Huy Nguyen, Pedram Akbarian, TrungTin Nguyen, and Nhat Ho. A general theory for softmax gating multinomial logistic mixture of experts. *arXiv preprint arXiv:2310.14188*, 2023.

Johan Obando-Ceron, Ghada Sokar, Timon Willi, Clare Lyle, Jesse Farebrother, Jakob Foerster, Gintare Karolina Dziugaite, Doina Precup, and Pablo Samuel Castro. Mixtures of experts unlock parameter scaling for deep rl. *arXiv preprint arXiv:2402.08609*, 2024.

James Oldfield, Markos Georgopoulos, Grigorios G Chrysos, Christos Tzelepis, Yannis Panagakis, Mihalis A Nicolaou, Jiankang Deng, and Ioannis Patras. Multilinear mixture of experts: Scalable expert specialization through factorization. *Advances in Neural Information Processing Systems*, 37:53022–53063, 2024.

Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Advances in neural information processing systems*, 31, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

Haoru Tan, Sitong Wu, Yanfeng Chen, Jun Xia, Ruobing Xie, Bin Xia, Samm Sun, and Xiaojuan Qi. Fast data mixture optimization via gradient descent. In *International Conference on Learning Representations*, volume 2026, pp. 702–715, 2026.

Tuan Thai, TrungTin Nguyen, Dat Do, Nhat Ho, and Christopher Drovandi. Model selection for gaussian-gated gaussian mixture of experts using dendrograms of mixing measures. *arXiv preprint arXiv:2505.13052*, 2025.

Alexei V Tkachenko. Thermodynamic bounds on energy use in quasi-static deep neural networks. *arXiv preprint arXiv:2503.09980*, 2025.

An Wang, Xingwu Sun, Ruobing Xie, Shuaipeng Li, Jiaqi Zhu, Zhen Yang, Pinxue Zhao, Weidong Han, Zhanhui Kang, Di Wang, et al. Hmoe: Heterogeneous mixture of experts for language modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21954–21968, 2025.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Di Zhao, Jin Lin, and Bin Song. Llmmr: A large language model-guided mixture-of-experts framework for intent-aware routing in leo satellite networks. *IEEE Transactions on Cognitive Communications and Networking*, 2026.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

A EXAMPLE: RADIAL GAUSSIAN DISTRIBUTION

This section presents a special case of the general framework in Section 3 under $M = 2, N = 2$. For convenience, we consider a randomized router with the same marginal distribution probabilities but that is independent of the sample conditional on the task identity. Therefore, this model corresponds to the limit where $\mathcal{C}_{ij} = 0$ in the general framework. In this example,

$$k_1 = k_2 = \frac{1}{2}, P_{11} = p_1, P_{12} = 1 - p_1, P_{21} = 1 - p_2, P_{22} = p_2.$$

In the static slice described in Appendix A.3, we employ an auxiliary router with identical marginal probabilities but independent of the samples given the task. If the construction is intended to be identical to that in the main text, simply add the covariance term to the corresponding proposition.

A.1 PROBLEM SETTING AND PRELIMINARIES

A.1.1 TASK SETTING AND MOE MODEL

We assume that the distributions $\mathbb{P}_1, \mathbb{P}_2$ of two sets of data $(x_1, y_1), (x_2, y_2)$ follow:

with probability 1/2: $y = +1, x \sim \mathcal{N}(0, (1 + \Delta_i)^2)\boldsymbol{I}_{2d}$,

with probability 1/2: $y = -1, x \sim \mathcal{N}(0, (1 - \Delta_i)^2)\boldsymbol{I}_{2d}$,

where $i = 1, 2$ corresponds to task 1 and task 2, $\Delta_i \in [0, 1], \Delta_1 \neq \Delta_2$. Let

$$\Delta_3 = \zeta_1\Delta_1 + (1 - \zeta_1)\Delta_2, \Delta_4 = \zeta_2\Delta_1 + (1 - \zeta_2)\Delta_2.$$

Here, Δ_3, Δ_4 denote the effective approximation parameters corresponding to the mixture distribution, and their approximation errors are given by Proposition A.12. We denote the general data distribution as $\mathbb{P} = \frac{1}{2}\mathbb{P}_1 + \frac{1}{2}\mathbb{P}_2$.

In the above setup, the two tasks share the same label prior (both equal to $1/2$) but possess different parameters Δ_i. Therefore, the discriminability between tasks is determined by the difference between Δ_1 and Δ_2. The effective parameters Δ_3, Δ_4 of the mixture distribution will be used in subsequent analyses to approximate the impact of the mixed data on experts.

Next, we introduce the toy MoE model in this case: the model consists of one routing layer and two experts E_1, E_2, where each expert is a two-layer MLP. For an input $x \in \mathbb{R}^{2d}$, the output form of expert E_i is

$$\hat{y}_i(x; r) = \frac{1}{H}\sum_{h=1}^{H} \sigma(x; r_{i,h}). \tag{35}$$

Since the samples exhibit a radial structure, routing also adopts a radial form. For the input $x \in \mathbb{R}^{2d}$, the routing output is defined as

$$u(x; s) = \mathrm{sigm}\big(s^2 - \|x\|_2^2; c\big), \tag{36}$$

where $\mathrm{sigm}(z; c) = \frac{1}{1+e^{-cz}}$. Let $\pi_1(x; s) = \mathrm{sigm}\big(s^2 - \|x\|_2^2; c\big)$, $\pi_2(x; s) = 1 - \pi_1(x; s)$ denote the weights assigned to expert E_1 and E_2 for sample x, respectively. It degenerates into a hard routing as $c \to \infty$. Lemma D.2 will show that linear routing is not suitable for this task.

For the routing parameter s, define the task-sample allocation probability:

$$p_1(s) = \mathbb{E}_{(x,y)\sim\mathbb{P}_1}[\pi_1(x; s)] = \int_{[0,+\infty)} \mathrm{sigm}(s^2 - z; c)\, \mu_1(dz), \tag{37}$$

$$p_2(s) = \mathbb{E}_{(x,y)\sim\mathbb{P}_2}[\pi_2(x; s)] = 1 - \int_{[0,+\infty)} \mathrm{sigm}(s^2 - z; c)\, \mu_2(dz). \tag{38}$$

Here, μ_1, μ_2 denote the marginal push-forward measures of $\|\boldsymbol{x}\|_2^2$ under each task, whose rigorous definition is given in Eq. (48). The above equation implies that the data from task 1 are assigned to experts E_1 and E_2 with probabilities $p_1(s)$ and $1 - p_1(s)$, respectively, while the data from task 2 are assigned to experts E_1 and E_2 with probabilities $1 - p_2(s)$ and $p_2(s)$, respectively.

A.1.2 MEAN-FIELD PRELIMINARIES

When the number of neurons $H \to \infty$, the discrete set of neuron weights is described by the probability measure $\rho_j(dr)$ of weight magnitudes. In addition, perform statistical averaging over random input directions to eliminate high-dimensional angular degrees of freedom, and define $q(r) = \mathbb{E}[\sigma(r\mathcal{G})]$, where $\mathcal{G} \sim \mathcal{N}(0,1)$. The function $q(r)$ shall satisfy the following regularity conditions:

#1. $q(r) \in C^2[0,+\infty], \sup_{r\in[0,+\infty]} |q''(r)| < \infty$; #2. $\forall r \in (0,+\infty), 0 < q'(r) < \infty, \lim_{r\to 0} q'(r) = \lim_{r\to+\infty} q'(r) = 0$; #3. $-\infty < q(0+) < -1, 1 < q(+\infty) < +\infty$; #4. Let $Z(r) = \frac{q'((1-\Delta)r)}{q'((1+\Delta)r)}$. $\forall \Delta \in (0,1), r \in (0,+\infty), Z'(r) > 0$; #5. $\sigma(\cdot)$ is a non-decreasing, bounded, and Lipschitz-continuous activation function.

Define the family of auxiliary functions:

$$q^+(r;\Delta) \triangleq q\big((1+\Delta)r\big), q^-(r;\Delta) \triangleq q\big((1-\Delta)r\big), \tag{39}$$

$$q_i^+(r) = q\big((1+\Delta_i)r\big), q_i^-(r) = q\big((1-\Delta_i)r\big), i = 1,2. \tag{40}$$

$$q_3^+(r) = \zeta_1 q_1^+(r) + (1-\zeta_1) q_2^+(r), q_3^-(r) = \zeta_1 q_1^-(r) + (1-\zeta_1) q_2^-(r) \tag{41}$$

$$q_4^+(r) = \zeta_2 q_1^+(r) + (1-\zeta_2) q_2^+(r), q_4^-(r) = \zeta_2 q_1^-(r) + (1-\zeta_2) q_2^-(r). \tag{42}$$

For $i,j \in \{1,2,3,4\}$ and $\rho_j \in \mathscr{P}([0,\infty])$, the infinite-width risk and auxiliary coefficients are defined as

$$R_i^\infty(\rho_j) = \frac{1}{2}\big(1 - \int q_i^+(r)\rho_j(dr)\big)^2 + \frac{1}{2}\big(1 + \int q_i^-(r)\rho_j(dr)\big)^2 \tag{43}$$

$$\psi_i^\infty(r;\rho_j) = \lambda_i^+(\rho_j)\cdot q_i^+(r) + \lambda_i^-(\rho_j)\cdot q_i^-(r), \tag{44}$$

where

$$\lambda_i^+(\rho_j) = \frac{1}{2}\big(\int q_i^+(r)\rho_j(dr) - 1\big), \lambda_i^-(\rho_j) = \frac{1}{2}\big(\int q_i^-(r)\rho_j(dr) + 1\big). \tag{45}$$

For convenience, for any $\rho \in \mathscr{P}([0,\infty])$, let

$$R^\infty(\rho;\Delta) = \frac{1}{2}\big(1 - \int q^+(r;\Delta)\rho(dr)\big)^2 + \frac{1}{2}\big(1 + \int q^-(r;\Delta)\rho(dr)\big)^2, \tag{46}$$

where

$$\lambda^+(\rho;\Delta) = \frac{1}{2}\big(\int q^+(r;\Delta)\rho(dr) - 1\big), \lambda^-(\rho;\Delta) = \frac{1}{2}\big(\int q^-(r;\Delta)\rho(dr) + 1\big). \tag{47}$$

Consider the mapping $\mathcal{Z}(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 \colon \mathbb{R}^{2d} \to [0,+\infty)$, define the push-forward Borel probability measure

$$\mu_i^+ := \mathcal{Z}_\sharp \mathcal{N}\big(0, (1+\Delta_i)^2 I_{2d}\big), \mu_i^- := \mathcal{Z}_\sharp \mathcal{N}\big(0, (1-\Delta_i)^2 I_{2d}\big), \tag{48}$$

$$\mu_i := \frac{1}{2}\mu_i^+ + \frac{1}{2}\mu_i^- \in \mathscr{P}([0,+\infty]), \tag{49}$$

where μ_i^+ (resp. μ_i^-) denotes the distribution of $\|\boldsymbol{x}\|_2^2$ under the distribution $\mathbb{P}_i$ conditioned on $y = +1$ resp. $y = -1$). μ_i is the marginal distribution of $\|\boldsymbol{x}\|_2^2$ under $\mathbb{P}_i$. For $\Delta_i < 1$, the push forward measures admit the following densities. When $\Delta_i = 1$, the $y = -1$ component degenerates to an atom at the origin and is treated separately. The density of this measure with respect to the Lebesgue measure is

$$f(z; 1\pm\Delta_i) = \frac{z^{d-1}\exp\big(-z/\big[2(1\pm\Delta_i)^2\big]\big)}{\big(2(1\pm\Delta_i)^2\big)^d \Gamma(d)}, z \geq 0,$$

This density is only used for numerical calculations, and the measure form is directly adopted in theoretical analysis. The infinite-width output of experts under the mean-field limit is

$$f_{\rho_j}(\boldsymbol{x}) = \int \sigma(\boldsymbol{x};r)\rho_j(dr).$$

Proposition A.1. *For any fixed finite $c > 0$, the functions $s \mapsto p_1(s;c), p_2(s;c)$ satisfy:*

1. *$p_1, p_2 \in C^2([0,+\infty))$ and $\sup_{s\geq 0} |p_i''(s;c)| < +\infty$;*
2. *$0 < p_1(0;c) < \frac{1}{2}, \frac{1}{2} < p_2(0;c) < 1$, $\lim_{s\to+\infty} p_1(s;c) = 1, \lim_{s\to+\infty} p_2(s;c) = 0$;*
3. *For all $s > 0$, $p_1'(s;c) > 0, p_2'(s;c) < 0$.*

A.1.3 Co-evolution of Routing and Experts

We assume that $\sigma(x; r)$ is differentiable with uniformly bounded derivatives, and all integrands satisfy the uniform integrability condition. To preserve analytical tractability, we do not directly analyze the sample-dependent radial gate $\pi_1(x; s)$ in the risk. Instead, we employ the randomized router introduced in Section A. Specifically, let $J \in \{1, 2\}$ denote the index of the selected expert. Conditional on the task identity $T = i$, the auxiliary router J is defined by the marginal assignment probabilities

$$\mathbb{P}(J = 1 \mid T = 1) = p_1(s), \mathbb{P}(J = 2 \mid T = 1) = 1 - p_1(s), \tag{50}$$

$$\mathbb{P}(J = 1 \mid T = 2) = 1 - p_2(s), \mathbb{P}(J = 2 \mid T = 2) = p_2(s), \tag{51}$$

with the crucial conditional independence assumption $J \perp (X, Y) \mid T$. Therefore, the router and the expert follow the coupled dynamics

$$\dot{s} = -\eta_s \partial_s \mathcal{R},\ \partial_t \rho_j + \partial_r(\rho_j v_j) = 0,\ v_j := -\eta_j \partial_r \frac{\delta \mathcal{R}}{\delta \rho_j}, j = 1, 2 \tag{52}$$

Note $R_{ij} := R_i(\rho_j), i, j \in \{1, 2\}$. Then the joint risk is

$$\mathcal{R} = \frac{1}{2}\left[p_1 R_{11} + (1 - p_1)R_{12} + (1 - p_2)R_{21} + p_2 R_{22}\right]. \tag{53}$$

Consider the total traffic received by the two experts

$$m_1 = \frac{1}{2}(p_1 + 1 - p_2), m_2 = \frac{1}{2}(1 - p_1 + p_2), \tag{54}$$

as well as its conditional task weights $\zeta_1 = \frac{p_1}{2m_1}, \zeta_2 = \frac{1-p_1}{2m_2}$. Based on the sign properties of $p_1'(s) > 0, p_2'(s) < 0$ in Proposition A.1, we further define the comparative advantages

$$\mathcal{A}_1 := \mathcal{A}_1^{1 \succ 2} = R_{12} - R_{11}, \mathcal{A}_2 := \mathcal{A}_2^{2 \succ 1} = R_{21} - R_{22}.$$

A.2 Specialization of the General Results

Propositions A.2 - A.4 are corollaries of Propositions 4.3 - 4.5.

Proposition A.2. *Under the no-flux boundary condition, we have*

$$\frac{d}{dt}\mathcal{R} = -\eta_s |\partial_s \mathcal{R}|^2 - \sum_{j=1}^{2} \eta_j \int \left|\partial_r \frac{\delta \mathcal{R}}{\delta \rho_j}(r)\right|^2 \rho_j(dr) \le 0. \tag{55}$$

Proposition A.3. *Within the zero-covariance slice introduced above, consider the hard threshold limit $c \to \infty$, then the routing dynamics satisfies*

$$\dot{s} = \frac{\eta_s}{2}\left[p_1'(s)\mathcal{A}_1 + p_2'(s)\mathcal{A}_2\right] = \frac{\eta_s}{2}\left[p_1'(s)\mathcal{A}_1 - |p_2'(s)|\mathcal{A}_2\right]. \tag{56}$$

Proposition A.4. *For tasks $i \in \{1, 2\}$ and experts $j, k \in \{1, 2\}$, suppose that the routing variable $\theta(t)$ follows the dynamics (25). Then the comparative advantage satisfies the evolution equation*

$$\dot{\mathcal{A}}_i^{j \succ k} = \eta_E \, \mathcal{D}_i^{j \succ k} + \mathcal{X}_i^{j \succ k}.$$

where

$$\mathcal{D}_i^{j \succ k} := \sum_{u=1}^{2} \frac{1}{2}\Big[P_{uj}(\theta) G_{iu}(\rho_j) - P_{uk}(\theta) G_{iu}(\rho_k)\Big], \tag{57}$$

in which $G_{iu}(\rho) := \big\langle g_i(\rho), g_u(\rho)\big\rangle_{L^2(\rho)}$ denotes the Gram matrix of the task gradient. $P_{uj}(\theta) = \mathbb{E}_{\mathbb{P}_u}\big[\pi_j(x; \theta)\big]$ is the marginal assignment probability assigned to expert j under task u.

Proposition A.5. *As $c \to \infty$, we have $p_1(0) = 0, p_2(0) = 1, m_1(0) = 0, m_2(0) = 1$. The set $\{(s, \rho_1, \rho_2) : s = 0\}$ is invariant under the joint gradient flow. More specifically, $\dot{s} = 0, \partial_t \rho_1 = 0$, and E_2 is trained on the mixture distribution $\frac{1}{2}\mathbb{P}_1 + \frac{1}{2}\mathbb{P}_2$. Hence $s = 0$ yields a collapsed fixed point. Similarly, we show that $s = \infty$ gives another collapsed fixed point.*

The proof of Proposition A.5 is detailed in Appendix C.2. Consider the hard routing limit $c \to \infty$ and define $\Lambda(\rho_1, \rho_2) := \kappa_1 \mathcal{A}_1(\rho_1, \rho_2) - \kappa_2 \mathcal{A}_2(\rho_1, \rho_2)$. Let $\varphi_i(z)$ denote the density of the measure μ_i with respect to the Lebesgue measure on $[0, \infty)$. Under the Gaussian setup, $\Delta_i < 1$ implies $\mu_i(\{0\}) = 0$, and there exists a constant $\kappa_i > 0$ such that

$$\varphi_i(z) = \kappa_i z^{d-1} + o(z^{d-1}), z \to 0^+. \tag{58}$$

Set the regularization term as $\mathcal{B}_\lambda(s) = \frac{\lambda}{2} h(s)^2$ with $h(s) = p_1(s) - p_2(s)$, then the joint objective becomes $\mathcal{R}_\lambda = \mathcal{R} + \mathcal{B}_\lambda$.

To rigorously characterize the stability of the collapsed fixed point $s = 0$, we invoke the quasi-static reduction enabled by the fast expert response. Under this reduction, the expert measures are slaved to the routing variable via $\rho_j = \rho_{*,j}(s)$, which yields the effective coefficient

$$\Lambda_*(s) := \Lambda\big(\rho_{*,1}(s), \rho_{*,2}(s)\big), C_\lambda^* := \Lambda_*(0) + 2\lambda(\kappa_1 + \kappa_2). \tag{59}$$

Proposition A.6. *If $C_\lambda^* < 0$, there exist constants $s_0 > 0$ and $\gamma > 0$ such that for all $0 < s \le s_0$,*

$$\Lambda_*(s) + 2\lambda(\kappa_1 + \kappa_2) \le -\gamma.$$

$s = 0$ is locally asymptotically stable for internal perturbations.

If $C_\lambda^* > 0$, then there exist constants $s_0 > 0$ and $\gamma > 0$ such that for all $0 < s \le s_0$,

$$\Lambda_*(s) + 2\lambda(\kappa_1 + \kappa_2) \ge \gamma.$$

For every sufficiently small $s(0) = \varepsilon > 0$, $s = 0$ is locally repelling for internal perturbations.

In the borderline case $C_\lambda^* = 0$, the higher-order terms in the expansion determine stability, which is not addressed here.

The proof of Proposition A.6 is detailed in Appendix C.3. This formulation clarifies the direction of the flow for small perturbations, while respecting the constraint $s \ge 0$, and explicitly distinguishes the equilibrium property of the boundary point from its stability under perturbations.

A.3 Static Slices

Proposition A.7. *Under the assumptions of Proposition 4.1, in this special case we have*

$$|s(t) - s_*(\rho(t))| \le e^{-\kappa_s t/\epsilon} |s(0) - s_*(\rho(0))| + \frac{C_s}{\kappa_s}\epsilon. \tag{60}$$

The dynamics of the expert measure can be reduced in the sense of velocity fields on the Wasserstein tangent space as

$$\partial_t \rho_i = \eta_i \partial_r \left(\rho_i \partial_r \frac{\delta \mathcal{R}}{\delta \rho_i}\big(s_*(\rho), \rho\big) \right) + O(\epsilon), i = 1, 2. \tag{61}$$

Proposition A.8. *Under the assumptions of Proposition 4.2, let*

$$d(t) = d_{\mathcal{W}}\big(\rho(t), \rho_*(s(t))\big). \tag{62}$$

Then

$$d(t) \le e^{-\kappa_\rho t/\epsilon} d(0) + \frac{L_* M_s}{\kappa_\rho}\epsilon. \tag{63}$$

The slow-route dynamics can be reduced to

$$\dot{s}(t) = -\eta_s \partial_s \mathcal{R}\big(s(t), \rho_*(s(t))\big) + O(\epsilon). \tag{64}$$

Propositions A.7 and A.8 are special cases of Lemmas 4.1 and 4.2, respectively. The quasi-static reduction via singular perturbation established in these propositions formalizes the connection between the fully coupled MoE dynamics and the two static slices defined in Appendix A.3. Specifically, in the fast-routing limit ($\epsilon \to 0$), the expert measures are effectively frozen on the routing time scale; consequently, the reduced dynamics corresponds to the static expert partition problem, in which the router optimizes the threshold s given fixed experts. In the fast-expert limit, the routing

parameter is effectively frozen on the expert time scale; consequently, the reduced dynamics corresponds to the static routing partition problem, in which the expert measures optimize their empirical risks given fixed routing assignments.

The two static slices can be viewed as complementary response maps that form a closed feedback loop. Given mixed weights (p_1, p_2), each expert converges to its final distribution ρ_j^*, giving the corresponding comparative advantage $\mathcal{A}_i^*$. Given comparative advantages $\mathcal{A}_1, \mathcal{A}_2$, the router converges to its final threshold s_*. Combining these response maps with a slow-variation approximation (ignoring fast changes), the high-dimensional coupled dynamics simplifies, during mid-to-late training, to a one-dimensional effective ordinary differential equation for the router parameter $s(t)$.

Reduced Dynamics and Complementary Static Slices Under the condition of Proposition A.8, expert measures track the equilibrium branch up to $O(\epsilon)$: $\rho(t) \approx \rho_*(s(t)) + O(\epsilon)$. Substituting this into the routing ordinary differential equation yields the one-dimensional effective dynamics in the mid-to-late training phase:

$$\dot{s}(t) = F(s) \triangleq \frac{\eta_s}{2}\Big[p_1'(s)\mathcal{A}_1^*(s) + p_2'(s)\mathcal{A}_2^*(s)\Big] + O(\epsilon). \tag{65}$$

Here, $\mathcal{A}_i^*(s)$ denotes the comparative advantage obtained from the expert response mapping under the routing condition induced by s.

This reduced ordinary differential equation serves as the core tool for analyzing the dynamics of specialization. We can investigate fixed points, the stability of fixed points, conditions for the emergence of specialization equilibria, and the existence conditions for collapsing fixed points ($s \to 0$ or $s \to +\infty$). The analysis of the static slice is no longer treated as an independent research object. Instead, the two static slices provide complementary response maps that close the reduced dynamics. The slice of static router yields the expert response map $\mathcal{A}_i^*(s)$, while the slice of static experts yields the routing response map $s_*(\mathcal{A}_1, \mathcal{A}_2)$, equivalently the fixed-point relation

$$p_1'(s_*)\mathcal{A}_1 = |p_2'(s_*)|\mathcal{A}_2.$$

Slice of Static Router We assume that the routing is performed at the task-level: a given sample from $\mathbb{P}_1$ is assigned to E_1 with probability p_1 and to E_2 with probability $(1 - p_1)$, while a given sample from $\mathbb{P}_2$, it is assigned to E_2 with probability p_2 and to E_1 with probability $(1 - p_2)$. Under this assumption, the routing random variable is conditionally independent of the specific sample (x, y) given the task. Consequently, the risk can be decomposed into the product of assignment probabilities and task risks. Consider the joint risk functional $\mathcal{R} : \mathcal{I} \times \mathcal{P}_2(\mathbb{R}_+)^2 \to \mathbb{R}$, where $\mathcal{I} \subset \mathbb{R}$ is a compact interval, and $\rho = (\rho_1, \rho_2) \in \mathcal{P}_2(\mathbb{R}_+)^2$, together with the singularly perturbed system

$$\epsilon\dot{s} = -\partial_s\mathcal{R}(s, \rho),\ \partial_t\rho_i + \partial_r(\rho_i v_i) = 0,\ v_i := -\eta_i\partial_r\frac{\delta R}{\delta\rho_i}(s, \rho), i = 1, 2. \tag{66}$$

Then the expert measure satisfies the continuity equation $\partial_t\rho_i + \partial_r(\rho_i v_i) = 0$. Suppose that there exists a compact set $\mathcal{K} \subset \mathcal{P}_2(\mathbb{R}_+)^2$ and an equilibrium branch $s_* : \mathcal{K} \to \mathcal{I}$, and assume that the trajectories under consideration always lie within some uniform tubular neighborhood of this branch.

Given routing assignment probabilities (p_1, p_2), compute traffic and conditional task weights m_j, ζ_j; under frozen routing, experts converge to the equilibrium measure ρ_j^*, which further yields the equilibrium comparative advantage. The expert response map is therefore defined.

We perform measure normalization

$$\zeta_1 = \frac{p_1}{p_1 - p_2 + 1}, \zeta_2 = \frac{1 - p_1}{1 - p_1 + p_2}. \tag{67}$$

This also implies that the actual inputs for E_1, E_2 are mixture distributions

$$\mathbb{P}_{1,0} = \zeta_1\mathbb{P}_1 + (1 - \zeta_1)\mathbb{P}_2, \mathbb{P}_{2,0} = \zeta_2\mathbb{P}_1 + (1 - \zeta_2)\mathbb{P}_2. \tag{68}$$

Due to the spherical symmetry of the data distribution, assume without loss of generality that $\Delta_1 < \Delta_2$. We first consider the case of expert E_1.

Proposition A.9. *Let* $\mathcal{D} = \{\Delta \in [0,1] : \exists r \in (0,+\infty)$ *s.t.* $q^+(r;\Delta) \geq 1, q^-(r;\Delta) \leq -1\}$*, and define* $\Delta^\infty := \inf \mathcal{D}$*. Set* $\widetilde{\Delta}^\infty := \inf\{\Delta \in [0,1] : \exists \rho \in \mathscr{P}([0,+\infty])$*s.t.* $\int q^+(r;\Delta)\, d\rho = 1, \int q^-(r;\Delta)\, d\rho = -1\}$*. Then* $\widetilde{\Delta}^\infty = \Delta^\infty$*.*

Proposition A.10. *Suppose* $\Delta > \Delta^\infty$*. Then there exist uncountably many probability measures* ρ^* *with finite support such that* $R^\infty(\rho^*;\Delta) = 0$*. Moreover, let* $\rho \in \mathscr{P}([0,+\infty])$ *and* $\{0,+\infty\} \cap \operatorname{supp}(\rho) = \varnothing$*. Then there exists a probability measure supported by at most two points* $\rho' = a\delta_{r_1} + (1-a)\delta_{r_2}, 0 \leq a \leq 1, 0 < r_1 \leq r_2 < +\infty$*, such that* $\int q^+(r;\Delta)\, d\rho' = \int q^+(r;\Delta)\, d\rho, \int q^-(r;\Delta)\, d\rho' = \int q^-(r;\Delta)\, d\rho$*. In particular,* $R^\infty(\rho';\Delta) = R^\infty(\rho;\Delta)$*. If the above moment vector does not lie on the curve* $\Gamma(\Delta) = \big\{\big(q^+(r;\Delta), q^-(r;\Delta)\big) : r \in (0,+\infty)\big\}$*, then we may further take* $r_1 < r_2, 0 < a < 1$*.*

According to Proposition A.10, we can set $\rho_3 = a_1\delta_{r_{31}} + (1-a_1)\delta_{r_{32}},\ r_{31} < r_{32}$. For any $\rho \in \mathscr{P}([0,+\infty))$, define vector

$$\boldsymbol{\alpha}_i(\rho) := \big(\alpha_i^+(\rho), \alpha_i^-(\rho)\big)^\top = \Big(\int q_i^+(r)\rho(dr) - 1, \int q_i^-(r)\rho(dr) + 1\Big)^\top \tag{69}$$

Let $\boldsymbol{\beta}_i = (\beta_i^+, \beta_i^-),\ i \in \{1,2\},\ \boldsymbol{\gamma}_i = (\gamma_i^+, \gamma_i^-)$, where

$$\beta_i^+(r_1,r_2;a) = \alpha_i^+(r_1,r_2;a) - q_i^+(r_i^*) + 1, \tag{70}$$

$$\beta_i^-(r_1,r_2;a) = \alpha_i^-(r_1,r_2;a) - q_i^-(r_i^*) - 1, \tag{71}$$

$$\gamma_i^+ = q_i^+(r_{i+2}^*) - q_i^+(r_1^*), \gamma_i^- = q_i^-(r_{i+2}^*) - q_i^-(r_1^*) \tag{72}$$

The meaning of r_i^* is given in Lemma D.1, where $r_i^* = \arg\min_r R^\infty(\delta_r;\Delta_i)$.

If $\Delta_1 < \Delta^\infty < \Delta_3$, we generally cannot determine the variation of $\|\boldsymbol{\beta}_1(r_{31},r_{32};a_1)\|_2^2$. In fact, the variation of the vector inner product $\langle \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_1 - \boldsymbol{\beta}_1\rangle$ depends on the specific choices of the activation functions $q(r)$ and p, which directly leads to the fluctuation of $\boldsymbol{\beta}_1$. The case $\Delta_3 < \Delta^\infty$ is analogous, since we cannot judge the monotonicity of $|r_3^* - r_1^*|$. This is also verified in our numerical simulations.

Similarly, for expert E_2, let $\rho_4 = a_2\delta_{r_{41}} + (1-a_2)\delta_{r_{42}}$. If $\Delta_4 < \Delta^\infty < \Delta_2$, then $\|\boldsymbol{\beta}_2(r_{41},r_{42};a_2)\|_2^2$ still fluctuates with Δ_1 and generally does not possess monotonicity.

Proposition A.11. *The general risk of specialization under distribution* $\mathbb{P}$ *is*

$$R(\rho_3;\rho_4) := \frac{1}{2}p_1R_1^\infty(\rho_3) + \frac{1}{2}(1-p_1)R_1^\infty(\rho_4) + \frac{1}{2}p_2R_2^\infty(\rho_4) + \frac{1}{2}(1-p_2)R_2^\infty(\rho_3) \tag{73}$$

where for $i \in \{1,2\}, j \in \{3,4\}$ *we have*

$$2R_i^\infty(\rho_j) = \begin{cases} \|\boldsymbol{\alpha}_i(r_{j1},r_{j2};a_{j-2})\|_2^2, \Delta_i \geq \Delta^\infty \\ \|\boldsymbol{\beta}_i(r_{j1},r_{j2};a_{j-2})\|_2^2, \Delta_i < \Delta^\infty \leq \Delta_j \\ \|\boldsymbol{\gamma}_i\|_2^2, \Delta_j < \Delta^\infty \end{cases} \tag{74}$$

We emphasize that ρ_3, ρ_4 *are related to the fixed probabilities* p_1, p_2*, i.e., in fact,* $\rho_3 = \rho_3(p_1,p_2), \rho_4 = \rho_4(p_2,p_1)$*.*

Finally, we present the error estimate for this effective-parameter approximation.

Proposition A.12. *Let* $U = \sup_{z\geq 0}|q''(z)| < \infty$*, and* $\|q\|_\infty \leq B$*. Given a weight* $w \in [0,1]$*, the mixture distribution* $P_w = w\mathbb{P}_1 + (1-w)\mathbb{P}_2$*, define the effective parameter and mixture variance as* $\overline{\Delta}_w = w\Delta_1 + (1-w)\Delta_2,\ V_w = w(1-w)(\Delta_2-\Delta_1)^2$*. Denote the mixed response by* $Q_w^\pm(r) = wq_1^\pm(r) + (1-w)q_2^\pm(r)$*. For all* $r \geq 0$*, the pointwise bound holds:*

$$\big|wq((1\pm\Delta_1)r) + (1-w)q((1\pm\Delta_2)r) - q((1\pm\overline{\Delta}_w)r)\big| \leq \frac{U}{2}r^2V_w. \tag{75}$$

If a probability measure $\rho \in \mathscr{P}([0,+\infty])$ *has* $m_2(\rho) := \int r^2\rho(dr) < +\infty$*, then*

$$\left|\int Q_w^\pm(r)\rho(dr) - \int q((1\pm\overline{\Delta}_w)r)\rho(dr)\right| \leq \frac{U}{2}m_2(\rho)V_w. \tag{76}$$

Define the asymptotic risk under mixture distribution:

$$R_w^{\mathrm{mix}}(\rho) = \frac{1}{2}\left(1 - \int Q_w^+(r)\rho(dr)\right)^2 + \frac{1}{2}\left(1 + \int Q_w^-(r)\rho(dr)\right)^2,$$

and the risk $R^{\infty}_{\bar{\Delta}_w}(\rho)$ corresponding to the effective parameter is given by Eq. (46), then the risk error bound satisfies

$$\left|R_w^{\mathrm{mix}}(\rho) - R^{\infty}_{\bar{\Delta}_w}(\rho)\right| \le (B+1)Um_2(\rho)V_w. \tag{77}$$

The proofs of Proposition A.9-Proposition A.12 are given in Appendix C.4-Appendix C.7. The propositions provide the response map on the expert side for the slice of static router. They express the risk through the assignment probabilities and bound the error of the effective-Δ approximation. This can help turn the original problem involving a mixture of tasks into a tractable risk calculus in low dimensions and support the subsequent analysis of the reduced dynamics.

Slice of Static Experts Consider the joint risk functional $\mathcal{R} : \mathcal{I} \times \mathcal{P}_2(\mathbb{R}_+)^2 \to \mathbb{R}$, where the routing parameter $s \in \mathcal{I}$ is a variable that varies slowly, and the expert measure $\rho = (\rho_1, \rho_2)$ is the opposite.

Consider the singularly perturbed system

$$\dot{s} = -\partial_s \mathcal{R}(s,\rho),\ \epsilon\partial_t\rho_i + \partial_r(\rho_i v_i) = 0,\ v_i := -\eta_i\partial_r\frac{\delta\mathcal{R}}{\delta\rho_i}(s,\rho), i = 1,2. \tag{78}$$

Define the product Wasserstein distance

$$d_{\mathcal{W}}(\rho,\widetilde{\rho}) := \left(W_2^2(\rho_1,\widetilde{\rho}_1) + W_2^2(\rho_2,\widetilde{\rho}_2)\right)^{1/2}. \tag{79}$$

Suppose that there exists a stable equilibrium branch $s \mapsto \rho_*(s) = (\rho_{*1}(s), \rho_{*2}(s))$ along with its uniform local basin of attraction, and assume that the trajectories under consideration always remain within this basin of attraction.

Given the fixed risk matrix $\{R_{ij}\}$, we obtain the comparative advantages $\mathcal{A}_1, \mathcal{A}_2$. The routing gradient flow is

$$\dot{s} = \frac{\eta_s}{2}\left[p_1'(s)\mathcal{A}_1 + p_2'(s)\mathcal{A}_2\right]. \tag{80}$$

Define the routing response map: $(\mathcal{A}_1, \mathcal{A}_2) \mapsto s_*$, where the fixed point s_* satisfies the condition of stationary point:

$$p_1'(s_*)\mathcal{A}_1 = |p_2'(s_*)|\mathcal{A}_2. \tag{81}$$

To distinguish from $\rho_3(p_1,p_2), \rho_4(p_1,p_2)$, we now denote the fixed measures as $\bar{\rho}_3, \bar{\rho}_4$. For samples from $\mathbb{P}_1$, E_1 is fed with probability $p_1(s)$ with risk $R_1^{\infty}(\bar{\rho}_3)$, while E_2 is fed with probability $1 - p_1(s)$ with risk $R_1^{\infty}(\bar{\rho}_4)$. For samples from $\mathbb{P}_2$, E_1 is fed with probability $\mathbb{E}_{\mathbb{P}_2}[\pi_1(x)] = 1 - p_2(s)$ with risk $R_2^{\infty}(\bar{\rho}_3)$, while E_2 is fed with probability $p_2(s)$ with risk $R_2^{\infty}(\bar{\rho}_4)$.

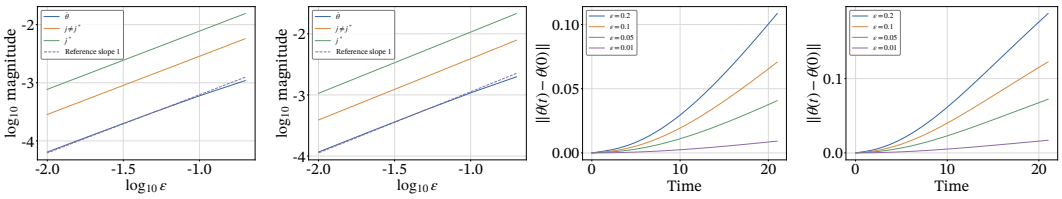
The general data distribution is $\mathbb{P} = \frac{1}{2}\mathbb{P}_1 + \frac{1}{2}\mathbb{P}_2$, and we consider

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_1}\left[\frac{1}{2}\left(y - \pi_1\hat{y}_1 - \pi_2\hat{y}_2\right)^2\right] = \mathbb{E}_{\mathbb{P}_1}\Big[&\pi_1(x)\cdot\frac{1}{2}(y-\hat{y}_1(x))^2 \\ &+ \pi_2(x)\cdot\frac{1}{2}(y-\hat{y}_2(x))^2 - \pi_1(x)\pi_2(x)\cdot\left(\hat{y}_1(x) - \hat{y}_2(x)\right)^2\Big].\end{aligned} \tag{82}$$

In the hard-router limit $c \to \infty$, either π_1 or π_2 equals zero, and the cross-term vanishes. Combining the definitions of $p_i(s)$ and $R_i^{\infty}(\rho)$ in Eq. (37) and Eq. (73), we obtain

Proposition A.13. *Under the mixture distribution $\mathbb{P}$, the specialization risk of static MoE is*

$$\begin{aligned}R_{\mathrm{r}}(\rho_{\mathrm{r}}) = \frac{1}{2}\int\Big[&p_1(s)R_1^{\infty}(\bar{\rho}_3) + (1-p_1(s))R_1^{\infty}(\bar{\rho}_4) \\ &+ p_2(s)R_2^{\infty}(\bar{\rho}_4) + (1-p_2(s))R_2^{\infty}(\bar{\rho}_3)\Big]\rho_{\mathrm{r}}(ds)\end{aligned} \tag{83}$$



(a) (b) (c) (d)

Figure 3: Approximate routing collapse. (a) and (b): scaling of the router velocity $\dot{\theta}$, velocity of expert $j \neq j^*$, and mixture approximation error of expert j^* as $\epsilon \to 0$. The dashed line with a slope of 1 is used as the reference line. (c) and (d): displacement of the router from the near-collapsed initialization for different values of ϵ. These results support Proposition 4.7.

Suppose $\rho_{\mathrm{r}} = \delta_s$, then by the Dominated Convergence Theorem, we can differentiate the risk as a function of the routing parameter s, and use the properties from Proposition A.1 that $p_1'(s) > 0,\ p_2'(s) < 0$:

$$\begin{aligned} R_{\mathrm{r}}'(s) = \frac{1}{2}\,[&p_1'(s)\,R_1^{\infty}(\bar{\rho}_3) - p_1'(s) R_1^{\infty}(\bar{\rho}_4) \\ &+ p_2'(s) R_2^{\infty}(\bar{\rho}_4) - p_2'(s)\,R_2^{\infty}(\bar{\rho}_3)]. \end{aligned} \tag{84}$$

Proposition A.14. *Consider the hard routing limit $c \to \infty$, and minimize the functional $R_{\mathrm{r}}(\rho_r)$ over the measure space $\mathscr{P}([0, +\infty])$. This is a linear functional with respect to ρ_r, and its minimum can also be attained at a point measure $\rho_r^* = \delta_{s^*}, s^* = \arg\min_{s \geq 0} h(s)$. This means that optimal routing corresponds to selecting a deterministic threshold parameter s^*.*

The proofs of Propositions A.13 and A.14 is given in Appendix C.9. They give an explicit expression for the routing risk as a functional of the routing measure, which helps analyze the behavior of the reduced one-dimensional routing dynamics.

B Additional Experiments

B.1 Further Verification

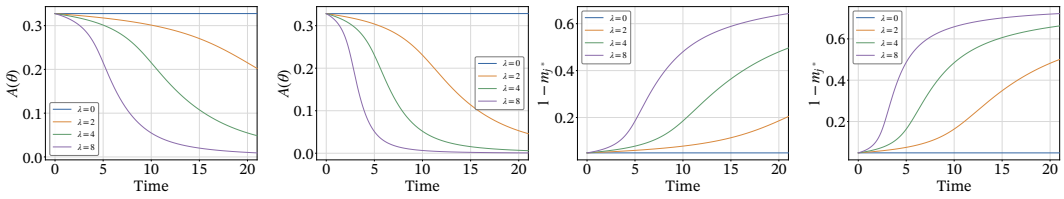
We test the prediction in Proposition 4.7. The router is initialized so that a single expert j^* receives at least $1 - \epsilon$ of every task, with $\epsilon \in \{0.20, 0.10, 0.05, 0.02, 0.01\}$. We measure the router velocity, the largest velocity among experts, and the discrepancy between the velocity of the dominant expert and the training approximation. The experiment is repeated for task separations $\delta \in \{2.0, 2.6, 3.2, 3.8\}$. We also track the displacement of the routing parameters from their near-collapsed initialization.

We also examine the load balancing mechanism in Proposition 4.8. Starting from $\epsilon = 0.05$ near-collapsed router, we optimize the regularized objective

$$\mathcal{R}_\lambda = \mathcal{R} + \lambda A,\, A(\theta) = \frac{1}{2}\sum_{j=1}^{N}\left(m_j - \frac{1}{N}\right)^2, \tag{85}$$

using $\lambda \in \{0, 2, 4, 8\}$. To isolate the destabilizing effect of the regularizer on routing, the experts are frozen and only the router is updated. We record the load imbalance $A(\theta)$, the total probability $1 - m_{j^*}$ assigned to non-dominant experts, and the initial regularized routing-gradient norm $\|\nabla_\theta \mathcal{R}_\lambda(\theta(0))\|$. We adopt $\delta = 2.0$ and $\delta = 3.8$ to test whether the behavior persists under different task geometries.

Scaling near routing collapse. Figures 3(a)-(b) report the relevant quantities on the log-log axes. The router velocity $\dot{\theta}$, the velocity of non-dominant experts $j \neq j^*$, and the approximation error of the expert j^* all increase approximately linearly with ϵ in log-log coordinates, with slopes close to the reference slope one. This provides numerical evidence for the result predicted by Proposition 4.7. The same qualitative behavior is observed for both task separation values. Figures 3(c)-(d) show the corresponding finite-time variation of routing parameters. As ϵ decreases, the router moves



(a) (b) (c) (d)

Figure 4: Effect of load balancing on a near-collapsed router. (a) and (b): evolution of the imbalance measure $A(\theta) = \frac{1}{2}\sum_j (m_j - 1/N)^2$. (c) and (d): total routing probability $1 - m_{j^*}$ assigned to experts other than the initially dominant expert. Increasing λ destabilizes the near-collapsed routing state, consistent with Proposition 4.8.

Table 1: Differential identity

τ	MSE ($\times 10^{-3}$)	Max residual
0.75	5.0 ± 3.7	1.5237 ± 1.1264
2.0	0.7 ± 0.3	1.8234 ± 3.5966
4.0	0.7 ± 0.5	1.0517 ± 0.1220
8.0	0.8 ± 0.6	1.0872 ± 0.1865

increasingly slowly away from its initial state. Hence, the local suppression of the routing gradient translates into a collapsed regime. Although the point is not necessarily an exact invariant state for $\epsilon > 0$, the routing variables can remain close to it during a substantial optimization interval.

Load balancing destabilizes the collapsed regime. Figure 4 starts with the same collapsed condition and varies the strength λ of regularization. Without load balancing ($\lambda = 0$), the routing distribution remains strongly concentrated on the dominant expert and the imbalance measure $A(\theta)$ remains large. Introducing the regularizer changes this behavior systematically. Figures 4(a)-(b) show that increasing λ drives the load-imbalance measure downward. At the same time, Figures 4(c)-(d) show an increase in $1 - m_{j^*}$, meaning that probability leaves the initially dominant expert and is redistributed to the remaining experts. The effect is observed for both task separation settings.

Robustness Test Table 1 reports the normalized differential error of the identity in Proposition 4.5. Table 2 reports the fraction of trajectory points for which the sufficient condition is satisfied, the mean positive lower bound, and the change in comparative advantage. Table 3 reports the slopes of double logarithmic fitting for the three quantities over four task separation values δ. Slopes close to 1 provide the quantitative counterpart of Figure 3. Table 4 reports the final diagonal assignment p_{ii} and the task risk gap for all four initial routing biases, extending the two representative matrices shown in Figure 2. Table 5 reports the initial and final load imbalance, the final non-dominant routing, and the initial regularized routing-gradient norm for each λ and task separation setting. The results complement the representative trajectories in Figures 1- 2 and confirm the robustness.

B.2 Experiments on the solvable example

In this section, we verify the results of the special case constructed in Appendix A. We numerically validate the main theoretical predictions for the radial Gaussian example with two tasks, using representative parameters $\Delta_1 = 0.25$ and $\Delta_2 = 0.65$. The experiments focus on four phenomena, including routing collapse versus specialization under regularization, convergence of the comparative advantage, the validity of the zero-covariance slice, and the effective-Δ approximation error.

Routing collapse and regularization. Figure 5(a) shows the routing trajectory $s(t)$ for the representative initial value $s_0 = 0.10$ and several regularization strengths λ. Although the critical value λ_c is not drawn explicitly, the curves exhibit the predicted threshold behavior. For sufficiently

Table 2: Sufficient condition

Similarity	Bias	$S > 0$ proportion	$\eta_E \nu$	$\Delta \mathcal{A}_i^{j \succ k}$
0.0	0.18	0.0163 ± 0.0365	1.04 ± 0.01	129.43 ± 40.64
0.0	0.42	0.2707 ± 0.1244	7.42 ± 5.16	282.47 ± 40.53
0.0	0.80	0.8272 ± 0.1814	15.24 ± 1.41	472.15 ± 77.72
0.75	0.18	0.0707 ± 0.0998	1.26 ± 0.75	133.12 ± 38.43
0.75	0.42	0.4299 ± 0.0967	7.82 ± 4.56	291.03 ± 37.03
0.75	0.80	0.8816 ± 0.1729	15.61 ± 0.98	479.78 ± 73.22

Table 3: Finite-sample collapse scaling

δ	Quantity	Slope	Slope SE ($\times 10^{-3}$)	Magnitude ($\times 10^{-3}$)
2.0	$\|\dot{\theta}\|$	0.9598 ± 0.007	27.92 ± 4.42	0.36 ± 0.10
2.0	$\|\tilde{v}_j\|$	1.001 ± 0.005	0.74 ± 0.39	1.46 ± 0.38
2.0	$\|\bar{v}_{j^*}\|$	1.001 ± 0.005	0.64 ± 0.20	3.93 ± 0.81
2.6	$\|\dot{\theta}\|$	0.9495 ± 0.006	28.13 ± 4.06	0.42 ± 0.16
2.6	$\|\tilde{v}_j\|$	1.002 ± 0.007	0.98 ± 0.38	1.64 ± 0.34
2.6	$\|\bar{v}_{j^*}\|$	1.002 ± 0.006	0.85 ± 0.39	4.44 ± 0.82
3.2	$\|\dot{\theta}\|$	0.9496 ± 0.006	28.13 ± 3.76	0.57 ± 0.18
3.2	$\|\tilde{v}_j\|$	1.002 ± 0.008	1.06 ± 0.45	1.80 ± 0.32
3.2	$\|\bar{v}_{j^*}\|$	1.002 ± 0.007	1.06 ± 0.32	4.94 ± 0.88
3.8	$\|\dot{\theta}\|$	0.9500 ± 0.006	27.80 ± 3.76	0.66 ± 0.22
3.8	$\|\tilde{v}_j\|$	1.002 ± 0.010	1.22 ± 0.53	1.92 ± 0.39
3.8	$\|\bar{v}_{j^*}\|$	1.002 ± 0.008	1.11 ± 0.48	5.32 ± 0.87

Here $\|\tilde{v}_j\| := \max_{j \neq j^*} \|v_j\|$ and $\|\bar{v}_{j^*}\| := \|v_{j^*} + \eta_{j^*} \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i}\big[\nabla_r \frac{\delta \ell_{j^*}}{\delta \rho_{j^*}}\big]\|_{L^2(\rho_{j^*})}$.

small λ, the routing parameter decays to zero, while for larger λ, it converges to a nonzero plateau. The same qualitative behavior is observed for the other initial values s_0. This is consistent with Proposition A.6: the regularized objective changes the sign of

$$C_\lambda^* = \Lambda_*(0) + 2\lambda(\kappa_1 + \kappa_2),$$

so that $s = 0$ is locally asymptotically stable in the small λ regime and locally repelling in the large λ regime, where a specialization equilibrium emerges.

Finite-width convergence of the comparative advantage. Figure 5(b) reports the evolution of the comparative advantage $\mathcal{A}_1(t)$ for finite widths $H = 16, 32, 64, 128, 256$, together with a reference curve at $H = 512$. For each seed, the results are qualitatively identical, so the figure displays a representative seed. The comparative advantage starts near zero, first decreases, and then increases toward a plateau. As H increases, the finite-width curves approach the $H = 512$ reference monotonically, and the largest width nearly coincides with the reference. This supports the mean-field picture underlying Proposition A.4, where the comparative advantage evolves according to

$$\dot{\mathcal{A}}_i^{j \succ k} = \eta_E \, \mathcal{D}_i^{j \succ k} + \mathcal{X}_i^{j \succ k},$$

and where finite-width fluctuations are expected to diminish as $H \to \infty$.

Covariance slice and the zero-covariance approximation. Figure 5(c) displays the covariance contribution $\mathcal{X}_i^{k \succ j}(t)$ as a function of time. The slice $\mathcal{C} = 0$ is identically zero and serves as a reference line. The curves based on samples are nonzero and show a clear ordering with respect to η_E. A larger η_E produces a more pronounced covariance contribution. These curves relax monotonically to the zero plateau as t increases, indicating that the effect of sample covariance weakens during training. This provides numerical support for the zero-covariance modeling choice adopted in the static slice: in the late time regime, the covariance term becomes small, so the decomposition

$$\dot{\mathcal{A}}_i^{j \succ k} = \eta_E \, \mathcal{D}_i^{j \succ k} + \mathcal{X}_i^{j \succ k}$$

Table 4: $M = N = 4$ structure ($\times 10^{-3}$)

Bias	P_{ii}^0	P_{ii}	$\mathcal{A}_i^{i \succ k}$
0.03	19.88 ± 5.62	288.82 ± 22.54	25.20 ± 32.51
0.1	73.34 ± 9.58	359.84 ± 28.65	76.35 ± 34.30
0.2	155.36 ± 17.54	464.90 ± 38.58	146.14 ± 38.12
0.4	329.92 ± 33.16	641.6 ± 44.08	265.76 ± 46.11

Here $P_{ii}^0 := \frac{1}{M} \sum_{i=1}^{M} (P_{ii}(0) - \max_{k \neq i} P_{ik}(0))$.

Table 5: Load-balancing collapse experiment($\times 10^{-3}$)

δ	λ	Initial $A(\theta)$	Final $A(\theta)$	$1 - m_{j^*}$	$\|\nabla_\theta \mathcal{R}_\lambda\|$
2.0	0	327.5 ± 0.4	327.58 ± 0.62	49.10 ± 0.77	5.24 ± 1.27
2.0	2	327.5 ± 0.4	202.14 ± 58.95	204.36 ± 82.17	456.13 ± 101.96
2.0	4	327.5 ± 0.4	48.63 ± 35.52	496.28 ± 101.9	416.25 ± 150.98
2.0	8	327.5 ± 0.4	9.28 ± 7.72	643.13 ± 54.19	243.25 ± 101.33
3.8	0	328.2 ± 0.4	328.45 ± 1.15	48.16 ± 1.17	9.13 ± 2.51
3.8	2	328.2 ± 0.4	46.05 ± 29.48	500.55 ± 88.92	312.65 ± 108.72
3.8	4	328.2 ± 0.4	6.08 ± 4.49	662.70 ± 38.84	159.36 ± 52.18
3.8	8	328.2 ± 0.4	0.83 ± 1.17	722.47 ± 22.52	89.58 ± 37.97

is well approximated by its zero-covariance counterpart.

Effective-Δ approximation error. Figures 6(a)-(d) compare the maximum absolute response error with the theoretical bound from Proposition A.12 as a function of the mixture variance V_w. Both the response error and the theoretical bound grow approximately linearly with V_w, and the slope of the theoretical bound scales approximately proportionally with U. For $U \geq 4.8$, all response-error points lie below the theoretical bound. For $U = 2.4$, the plotted curve is slightly below a few numerical points; this is expected because $U = 2.4$ is below the actual value $\sup_{r \geq 0} |q''(r)| = 4.8$ for the chosen activation, so the assumption of Proposition A.12 is not satisfied and the curve is not guaranteed to be an upper bound. Overall, the experiment confirms the error estimate

$$\left| R_w^{\mathrm{mix}}(\rho) - R_{\overline{\Delta}_w}^{\infty}(\rho) \right| \leq (B+1) U m_2(\rho) V_w,$$

and illustrates the necessity of using a valid upper bound on $|q''|$.

Summary. Regularization controls the stability of the collapsed routing fixed point. The comparative advantage converges as the width increases, in agreement with the mean-field advantage dynamics. The contribution from sample covariance weakens over training, justifying the zero-covariance slice as a useful approximation. Finally, the effective-Δ approximation error obeys the predicted scaling with V_w whenever the theoretical upper bound on $|q''|$ is respected.

C Proofs

C.1 Proof of Proposition A.1

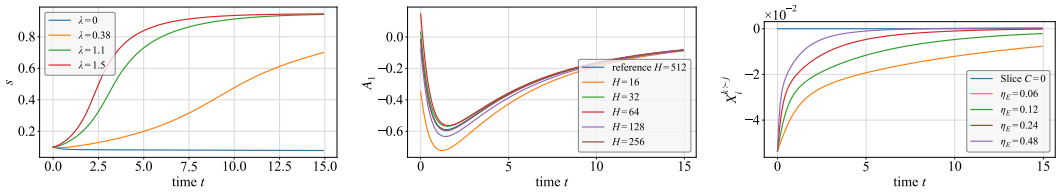
Proof. Since $\mathrm{sigm}(z; c) = \frac{1}{1+\exp(-cz)}$, we have

$$\mathrm{sigm}'(z; c) = c\,\mathrm{sigm}(z; c)\big(1 - \mathrm{sigm}(z; c)\big) > 0.$$

Because $\mathrm{sigm} \in C^\infty(\mathbb{R})$ and all its derivatives are bounded on any finite interval

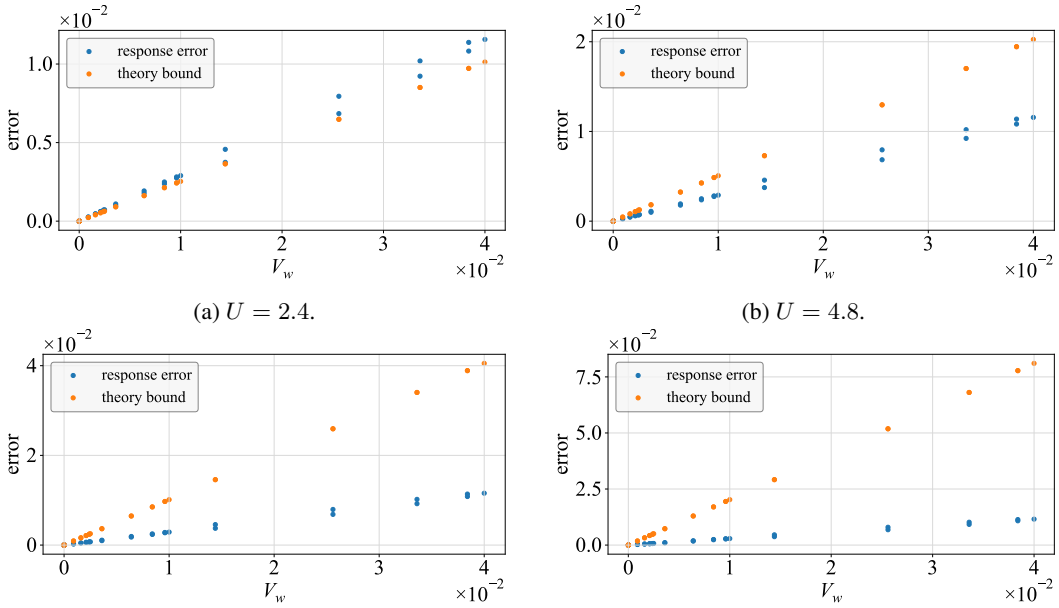
$$p_1''(s; c) = 2 \int \mathrm{sigm}'(s^2 - z; c)\, \mu_1(dz) + 4s^2 \int \mathrm{sigm}''(s^2 - z; c)\, \mu_1(dz). \tag{86}$$

We compute $p_2''(s; c)$ analogously, which gives $p_1, p_2 \in C^\infty([0, +\infty))$.



(a) (b) (c)

Figure 5: (a): Routing trajectories $s(t)$ for $s_0 = 0.10$ and different regularization strengths λ. Small λ leads to collapse to $s = 0$, while large λ leads to convergence to a nonzero specialization plateau, in agreement with Proposition A.6. (b): Comparative advantage $\mathcal{A}_1(t)$ for several finite widths H and a reference width $H = 512$. Increasing H yields a monotone approach to the reference curve, and the largest finite width nearly coincides with the mean-field reference. This is consistent with the advantage dynamics in Proposition A.4. (c): Covariance contribution $\mathcal{X}_i^{k\succ j}(t)$. The $\mathcal{C} = 0$ slice is identically zero. The curves of sample covariance are nonzero, ordered by η_E, and relax monotonically to the zero plateau, justifying the zero-covariance approximation in the regime.



(a) $U = 2.4$. (b) $U = 4.8$.

(c) $U = 9.6$. (d) $U = 19.2$.

Figure 6: Effective-Δ approximation error versus V_w. The response error and the theoretical bound grow approximately linearly with V_w. For $U \geq 4.8$, all numerical points lie below the bound. For $U = 2.4$, the bound is not guaranteed because it is below the true $\sup |q''|$, and the slight violation is therefore consistent with the theory.

When $s^2 - z \geq s^2/2$, we have $|\mathrm{sigm}'(s^2 - z; c)| \leq C_c \exp(-cs^2/2)$, so

$$s^2 \int_{z \leq s^2/2} |\mathrm{sigm}''(s^2 - z; c)| \, \mu_1(dz) \leq C_c s^2 e^{-cs^2/2}. \tag{87}$$

When $s^2 - z \leq s^2/2$, from the density representation of measure μ_1, there exist constants C, a_1 such that

$$\mu_1([s^2/2, \infty)) \leq C(1 + s^{2d-2}) \exp(-a_1 s^2). \tag{88}$$

Hence $\sup_{s \geq 0} |p_i''(s; c)| < \infty$.

At $s = 0$, $p_1(0; c) = \int \mathrm{sigm}(-z; c) \, \mu_1(dz)$, so $0 < p_1(0; c) < 1/2$. Similarly $1/2 < p_2(0; c) < 1$. By the dominated convergence theorem, $p_1(+\infty; c) = 1$ and $p_2(+\infty; c) = 0$.

For $s > 0$, $\mathrm{sigm}'(z; c) > 0$, so

$$p_1'(s; c) = 2s \int \mathrm{sigm}'(s^2 - z; c)\, \mu_1(dz) > 0, p_2'(s; c) = -2s \int \mathrm{sigm}'(s^2 - z; c)\, \mu_2(dz) < 0. \quad (89)$$

As $c \to \infty$, $\mathrm{sigm}(s^2 - z; c) \to \mathbf{1}_{\{z<s^2\}}$. Therefore $p_1(s; \infty) = \mu_1([0, s^2))$, $p_2(s; \infty) = 1 - \mu_2([0, s^2))$.

Consequently, $p_1(0+; \infty) = 0$, $p_2(0+; \infty) = 1$, $p_1(+\infty; \infty) = 1$, $p_2(+\infty; \infty) = 0$. □

C.2 Proof of Proposition A.5

Proof. At $s = 0$, all samples (up to a null set) are assigned to E_2, so $p_1 = 0$, $p_2 = 1$. The total traffic of E_1 is $m_1 = p_1 + 1 - p_2 = 0$, and its mean-field velocity vanishes.

On the other hand, the routing equation reads

$$\dot{s} = \frac{\eta_s}{2} \left[p_1'(s) \mathcal{A}_1 + p_2'(s) \mathcal{A}_2 \right]. \quad (90)$$

Since $p_1'(0) = p_2'(0) = 0$, we get $\dot{s} = 0$ regardless of expert risks. Thus, this set is invariant under the dynamics.

We next show that $s \to \infty$ corresponds to the other collapsing fixed point. Introduce the change-of-variable $u = \frac{s}{1+s}$, mapping $s \in [0, +\infty) \mapsto u \in [0, 1)$. Then $s \to \infty$ corresponds to $u \to 1^-$. Since

$$\dot{u} = \frac{1}{(1+s)^2} \dot{s}, \quad (91)$$

we have $(1 + s)^{-2} \to 0$ and $\dot{u} \to 0$ as $s \to \infty$. Consider the set $\{(u, \rho_1, \rho_2)\}$ on the compactified phase-space boundary. On this boundary $m_2 = 0$, so $\partial_t \rho_2 = 0$. Hence $u = 1$, corresponding to $s \to \infty$, forms another collapsing fixed point where all samples are assigned to E_1. □

C.3 Proof of Proposition A.6

Proof. As $c \to \infty$, $\mathrm{sigm}(s^2 - z; c) \to \mathbf{1}_{\{z<s^2\}}$ holds μ_i-almost everywhere ($\mu_i(\{0\}) = 0$). By dominated convergence,

$$p_1(s) = \int_0^\infty \mathbf{1}_{\{z<s^2\}}\, \mu_1(dz) = \mu_1([0, s^2)), p_2(s) = 1 - \mu_2([0, s^2)). \quad (92)$$

Differentiate for $s > 0$, we obtain

$$p_1'(s) = 2s\, \varphi_1(s^2), p_2'(s) = -2s\, \varphi_2(s^2). \quad (93)$$

From Eq. (48), $\mu_i = \frac{1}{2}\mu_i^+ + \frac{1}{2}\mu_i^-$, with density

$$\varphi_i(z) = \frac{1}{2} f(z; 1 + \Delta_i) + \frac{1}{2} f(z; 1 - \Delta_i), \quad (94)$$

where $f(z; \sigma) = \frac{z^{d-1} e^{-z/(2\sigma^2)}}{(2\sigma^2)^d \Gamma(d)}$. As $z \to 0^+$, $e^{-z/(2\sigma^2)} = 1 + o(1)$, so

$$\varphi_i(z) = \kappa_i z^{d-1} + o(z^{d-1}), \kappa_i = \frac{1}{2\Gamma(d)} \left[\frac{1}{(2(1 + \Delta_i)^2)^d} + \frac{1}{(2(1 - \Delta_i)^2)^d} \right] > 0, \quad (95)$$

which is exactly Eq. (58).

By definition $h(s) = p_1(s) - p_2(s) = \int_0^{s^2} \big(\varphi_1(z) + \varphi_2(z)\big)\, dz - 1$. Substitute Eq. (58) and integrate:

$$h(s) = -1 + \frac{\kappa_1 + \kappa_2}{d} s^{2d} + o(s^{2d}), \quad (96)$$

$$h'(s) = 2s\big(\varphi_1(s^2) + \varphi_2(s^2)\big) = 2(\kappa_1 + \kappa_2) s^{2d-1} + o(s^{2d-1}). \quad (97)$$

Similarly

$$p_1'(s) = 2\kappa_1 s^{2d-1} + o(s^{2d-1}), p_2'(s) = -2\kappa_2 s^{2d-1} + o(s^{2d-1}). \quad (98)$$

The regularized routing gradient flow is

$$\dot{s} = -\eta_s \partial_s \mathcal{R}_\lambda = \frac{\eta_s}{2}\big[p_1'(s)\mathcal{A}_1 + p_2'(s)\mathcal{A}_2\big] - \eta_s \lambda\, h(s)h'(s). \tag{99}$$

Insert the expansions above:

$$\frac{\eta_s}{2}\big[p_1'\mathcal{A}_1 + p_2'\mathcal{A}_2\big] = \eta_s(\kappa_1\mathcal{A}_1 - \kappa_2\mathcal{A}_2)s^{2d-1} + o(s^{2d-1}), \tag{100}$$

$$-\eta_s\lambda h(s)h'(s) = -\eta_s\lambda\big[-1+O(s^{2d})\big]\big[2(\kappa_1+\kappa_2)s^{2d-1}+o(s^{2d-1})\big] = 2\eta_s\lambda(\kappa_1+\kappa_2)s^{2d-1}+o(s^{2d-1}). \tag{101}$$

Combining terms yields the pointwise expansion

$$\dot{s} = \eta_s\big[\Lambda(\rho_1,\rho_2) + 2\lambda(\kappa_1+\kappa_2)\big]s^{2d-1} + o(s^{2d-1}), \tag{102}$$

where $\Lambda(\rho_1,\rho_2) = \kappa_1\mathcal{A}_1 - \kappa_2\mathcal{A}_2$.

The coefficient Λ depends on the expert measures ρ_1, ρ_2, which evolve together with s. Hence, the sign of $\dot{s}$ at a single instant cannot determine the long-term behavior. To obtain a rigorous stability statement, we invoke the quasi-static reduction established in the fast-expert-response regime (see Proposition 4.2). In that case, the experts equilibrate on a much faster timescale than the router, so that $\rho_j(t) = \rho_{*,j}(s(t)) + o(1)$ uniformly for s in a compact neighborhood of 0. Substituting this slaved relation into Eq. (102) yields the *effective* dynamics

$$\dot{s} = \eta_s\big[\Lambda_*(s) + 2\lambda(\kappa_1+\kappa_2)\big]s^{2d-1} + o(s^{2d-1}), \tag{103}$$

where

$$\Lambda_*(s) := \Lambda\big(\rho_{*,1}(s),\rho_{*,2}(s)\big), C_\lambda^* := \Lambda_*(0) + 2\lambda(\kappa_1+\kappa_2). \tag{104}$$

The remainder $o(s^{2d-1})$ in Eq. (103) is uniform with respect to the expert states due to the uniform convergence of $\rho_j(t)$ to $\rho_{*,j}(s(t))$ in a small neighborhood of $s=0$. Assume $C_\lambda^* < 0$. Since $\Lambda_*(s)$ is continuous at $s=0$ (by the continuity of the mean-field map $\rho \mapsto \mathcal{A}_i$ and the quasi-static branch), there exist constants $s_0 > 0$ and $\gamma > 0$ such that for all $0 < s \le s_0$,

$$\Lambda_*(s) + 2\lambda(\kappa_1+\kappa_2) \le -\gamma. \tag{105}$$

Choosing s_0 small enough also ensures that the $o(s^{2d-1})$ term in Eq. (103) is bounded in absolute value by $\frac{\eta_s\gamma}{2}s^{2d-1}$ for $0 < s \le s_0$. Consequently,

$$\dot{s} \le -\frac{\eta_s\gamma}{2}s^{2d-1}, 0 < s \le s_0. \tag{106}$$

This differential inequality implies that for any initial $0 < \varepsilon < s_0$, the solution satisfies

$$s(t)^{2-2d} \ge \varepsilon^{2-2d} + (d-1)\eta_s\gamma\, t(d>1), \tag{107}$$

and therefore $s(t) \to 0$ as $t \to \infty$. For $d=1$, the inequality becomes $\dot{s} \le -(\eta_s\gamma/2)s$, which also gives exponential decay. Hence, the collapsed fixed point $s=0$ is locally asymptotically stable for internal perturbations.

Assume $C_\lambda^* > 0$. By continuity, there exist $s_0 > 0, \gamma > 0$ such that

$$\Lambda_*(s) + 2\lambda(\kappa_1+\kappa_2) \ge \gamma, 0 < s \le s_0. \tag{108}$$

For sufficiently small $\varepsilon > 0$, we have $\dot{s}(0^+) > 0$, and the trajectory satisfies $s(t) > \varepsilon$ for all sufficiently small $t > 0$; indeed, the flow is strictly outward until it leaves the neighborhood. Thus, $s=0$ is locally repelling.

At $s=0$, $p_1'(0) = p_2'(0) = h'(0) = 0$ (since $s^{2d-1}\big|_{s=0} = 0$), so $\dot{s}(0) = 0$. Moreover, from the structure of the gradient flow, the entire set $\{s=0\}$ is invariant: if $s(0)=0$, then $s(t) \equiv 0$ for all $t \ge 0$. This confirms that $s=0$ is a fixed point, whose stability is determined by the sign of C_λ^* as above.

If $C_\lambda^* = 0$, the sign of $\dot{s}$ depends on the next-order terms in the expansion of $\Lambda_*(s)$ or on the remainder; this case is not addressed here.

Considering that $h(s) = -1 + \frac{\kappa_1+\kappa_2}{d}s^{2d} + o(s^{2d})$,

$$\mathcal{B}_\lambda(s) = \frac{\lambda}{2}h(s)^2 = \frac{\lambda}{2} - \frac{\lambda(\kappa_1+\kappa_2)}{d}s^{2d} + o(s^{2d}). \tag{109}$$

For $d > 1$, $2d > 2$, hence $\mathcal{B}_\lambda''(0) = 0$. The regularization does not alter the leading-order curvature at $s=0$. The attraction is carried by the s^{2d-1} term in the gradient. When $d=1$, $\mathcal{B}_\lambda''(0) = -2\lambda(\kappa_1+\kappa_2) < 0$, and the leading-order transverse dynamics is linear, corresponding to an ordinary linear instability/attraction in the routing direction. □

C.4 Proof of Proposition A.9

Proof. Define

$$\mathcal{F} = \big\{\Delta \in [0,1] : \exists \rho \in \mathscr{P}([0,+\infty]) \text{ s.t. } \int q^+(r;\Delta)\, d\rho = 1,\ \int q^-(r;\Delta)\, d\rho = -1\big\}. \quad (110)$$

It suffices to show $\mathcal{F} = \mathcal{D}$.

First, prove $\mathcal{F} \subseteq \mathcal{D}$. Suppose in contradiction that $\Delta \in \mathcal{F}$ but $\Delta \notin \mathcal{D}$. Since $q^-(\cdot;\Delta)$ is strictly increasing, there exists a unique r_- with $q^-(r_-;\Delta) = -1$. From $\Delta \notin \mathcal{D}$ we get $q^+(r_-;\Delta) < 1$. Set $\lambda_* = \frac{\partial_r q^+(r_-;\Delta)}{\partial_r q^-(r_-;\Delta)} > 0$, $h(r) = q^+(r;\Delta) - \lambda_* q^-(r;\Delta)$.

By assumption, $r \mapsto \frac{\partial_r q^-(r;\Delta)}{\partial_r q^+(r;\Delta)}$ is strictly increasing, so h is strictly increasing before r_- and strictly decreasing after r_-. Thus, r_- is the unique maximizer: $h(r) \le h(r_-)$ for all $r \ge 0$.

If ρ satisfies $\int q^-(r;\Delta)\, d\rho = -1$, integrate the inequality:

$$\int q^+(r;\Delta)\, d\rho - \lambda_* \int q^-(r;\Delta)\, d\rho \le q^+(r_-;\Delta) - \lambda_* q^-(r_-;\Delta).$$

Using $\int q^-(r;\Delta)\, d\rho = q^-(r_-;\Delta) = -1$, we obtain $\int q^+(r;\Delta)\, d\rho \le q^+(r_-;\Delta) < 1$, contradicting $\Delta \in \mathcal{F}$. Hence $\mathcal{F} \subseteq \mathcal{D}$.

Now prove $\mathcal{D} \subseteq \mathcal{F}$. Take any $\Delta \in \mathcal{D}$. There exists $r_0 \in (0,+\infty)$ such that $X := q^+(r_0;\Delta) \ge 1, Y := q^-(r_0;\Delta) \le -1$. If $(X,Y) = (1,-1)$, take $\rho = \delta_{r_0}$. Otherwise, let $\gamma(r;\Delta) = \big(q^+(r;\Delta), q^-(r;\Delta)\big)$. The line through (X,Y) and $(1,-1)$ intersects the diagonal $x = y$ at some point (z,z). Explicitly, $(1,-1) = \lambda(X,Y) + (1-\lambda)(z,z), \lambda = \frac{2}{X-Y} \in (0,1]$.

From $X \ge 1$, $Y \le -1$ and $q(0) < -1 < 1 < q(+\infty)$, we get $z \in \big(q(0), q(+\infty)\big)$. Therefore, (z,z) lies in the relative interior of the diagonal segment $\operatorname{conv}\big\{\gamma(0;\Delta), \gamma(+\infty;\Delta)\big\}$, and $(1,-1) \in \operatorname{conv}\big\{\gamma(0;\Delta), \gamma(r_0;\Delta), \gamma(+\infty;\Delta)\big\}$.

By continuity of $\gamma(\cdot;\Delta)$, replace the endpoints 0 and $+\infty$ by sufficiently small and sufficiently large finite points. Thus, there exist finite $r_1, r_2, r_3 \in [0,+\infty)$ and weights $a_j \ge 0$, $\sum_{j=1}^3 a_j = 1$, such that $\sum_{j=1}^3 a_j \gamma(r_j;\Delta) = (1,-1)$. Set $\rho = \sum_{j=1}^3 a_j \delta_{r_j}$. Then $\int q^+(r;\Delta)\, d\rho = 1$, $\int q^-(r;\Delta)\, d\rho = -1$. Hence $\Delta \in \mathcal{F}$, so $\mathcal{D} \subseteq \mathcal{F}$.

In summary, $\mathcal{D} = \mathcal{F}$. Taking the lower bound on both sides gives $\widetilde{\Delta}^\infty = \inf \mathcal{F} = \inf \mathcal{D} = \Delta^\infty$. □

C.5 Proof of Proposition A.10

Proof. Let $\gamma(r;\Delta) = \big(q^+(r;\Delta), q^-(r;\Delta)\big)$.

We first show that there are uncountable zero-risk measures. Since $\Delta > \Delta^\infty$, choose $\widetilde{\Delta} \in (\Delta^\infty, \Delta)$. Because $\widetilde{\Delta} \in \mathcal{D}$, there exists $r_0 > 0$ such that $q^+(r_0;\widetilde{\Delta}) \ge 1, q^-(r_0;\widetilde{\Delta}) \le -1$. Since $q' > 0$, increasing $\widetilde{\Delta}$ to Δ yields strict inequalities: $q^+(r_0;\Delta) > 1, q^-(r_0;\Delta) < -1$.

Consider the line connecting (X,Y) to $(1,-1)$; it intersects the diagonal $y = x$ at (z,z). From the two-point formula,

$$z = -\frac{X+Y}{X-Y-2}.$$

Since $X > 1$, $Y < -1$, we have $X - Y - 2 > 0$. Direct verification gives

$$z < 1 \iff -\frac{X+Y}{X-Y-2} < 1 \iff X > 1,$$

$$z > -1 \iff -\frac{X+Y}{X-Y-2} > -1 \iff Y < -1.$$

Thus $-1 < z < 1$. Recalling $q(0) < -1 < 0 < 1 < q(+\infty)$, we get $q(0) < z < q(+\infty)$. Hence (z,z) lies in the interior of the segment joining $\gamma(0;\Delta) = (q(0), q(0))$ and $\gamma(+\infty;\Delta) =$

$(q(+\infty), q(+\infty))$. Moreover, $(1,-1)$ lies on the segment between (z,z) and $\gamma(r_0;\Delta) = (X,Y)$, with $(1,-1) \neq (z,z)$ and $(1,-1) \neq (X,Y)$. Therefore

$$(1,-1) \in \operatorname{int}\operatorname{conv}\{\gamma(0;\Delta), \gamma(r_0;\Delta), \gamma(+\infty;\Delta)\}.$$

By the continuity of $\gamma(\cdot;\Delta)$, there exist finite points $0 < r_1 < r_0 < r_3 < +\infty$ such that

$$(1,-1) \in \operatorname{int}\operatorname{conv}\{\gamma(r_1;\Delta), \gamma(r_0;\Delta), \gamma(r_3;\Delta)\}.$$

Points inside a triangle are stable under perturbations of vertices. Hence, there exists a non-degenerate open neighborhood $I \subset (0, r_0)$ of r_1 such that for every $s \in I$,

$$(1,-1) \in \operatorname{int}\operatorname{conv}\{\gamma(s;\Delta), \gamma(r_0;\Delta), \gamma(r_3;\Delta)\}.$$

Thus, for each $s \in I$, there exist strictly positive weights $a_1(s), a_2(s), a_3(s) > 0$ with $\sum_{j=1}^{3} a_j(s) = 1$, satisfying $a_1(s)\gamma(s;\Delta) + a_2(s)\gamma(r_0;\Delta) + a_3(s)\gamma(r_3;\Delta) = (1,-1)$. Define $\rho_s = a_1(s)\delta_s + a_2(s)\delta_{r_0} + a_3(s)\delta_{r_3}$. Then $R^\infty(\rho_s;\Delta) = 0$. Different $s \in I$ produce different supports $\operatorname{supp}(\rho_s) = \{s, r_0, r_3\}$. Since I is uncountable, there exist uncountably many measures of finite support and zero risk.

We now prove the two-point representation statement. From $\{0, +\infty\} \cap \operatorname{supp}(\rho) = \varnothing$ and the closedness of the support set, there exists $0 < a \leq b < +\infty$ such that $\operatorname{supp}(\rho) \subseteq [a,b]$. Let the moment vector of ρ be $m_\rho = \int \gamma(r;\Delta)\,\rho(dr)$. Then $m_\rho \in \operatorname{conv}\gamma([a,b];\Delta)$. Since $[a,b]$ is connected and $\gamma(\cdot;\Delta)$ continuous, $\gamma([a,b];\Delta)$ is a compact connected subset of $\mathbb{R}^2$. By the Fenchel–Eggleston convex-hull theorem, any point inside the convex hull of a compact connected planar set can be written as a convex combination of at most two points from the set. Hence, there exist $r_1, r_2 \in [a,b],\ \alpha \in [0,1]$, such that $m_\rho = \alpha\gamma(r_1;\Delta) + (1-\alpha)\gamma(r_2;\Delta)$. After swapping indices, we may assume $r_1 \leq r_2$. Set $\rho' = \alpha\delta_{r_1} + (1-\alpha)\delta_{r_2}$. Then $\int \gamma(r;\Delta)\,\rho'(dr) = \int \gamma(r;\Delta)\,\rho(dr)$, i.e.

$$\int q^+(r;\Delta)\,d\rho' = \int q^+(r;\Delta)\,d\rho, \int q^-(r;\Delta)\,d\rho' = \int q^-(r;\Delta)\,d\rho. \tag{111}$$

Since $R^\infty(\cdot;\Delta)$ depends only on these two moments, we immediately obtain $R^\infty(\rho';\Delta) = R^\infty(\rho;\Delta)$. Finally, if $m_\rho \notin \Gamma(\Delta)$, the representation cannot degenerate to a single point, so we can take $r_1 < r_2, 0 < \alpha < 1$. □

C.6 Proof of Proposition A.11

Proof. When $\Delta_i \geq \Delta^\infty$, task i satisfies $\Delta_i \geq \Delta^\infty$, so there exist measures to achieve zero risk under $\mathbb{P}_i$. Measure ρ_j is zero risk for the mixture distribution with parameter Δ_j, but this does not imply zero risk for task Δ_i. Therefore,

$$R_i^\infty(\rho_j) = \frac{1}{2}\|\boldsymbol{\alpha}_i(r_{j1}, r_{j2}; a_{j-2})\|_2^2. \tag{112}$$

Similarly, when $\Delta_i < \Delta^\infty \leq \Delta_j$: since $\Delta_j > \Delta^\infty$, the mixture distribution admits a two-point measure ρ_j with zero risk, while $\Delta_i < \Delta^\infty$ means that no measure can achieve zero risk for task i. Hence,

$$R_i^\infty(\rho_j) = \frac{1}{2}\|\boldsymbol{\beta}_i(r_{j1}, r_{j2}; a_{j-2})\|_2^2. \tag{113}$$

Finally, if $\Delta_j < \Delta^\infty$, even the mixture distribution does not have a zero risk measure. We use $(1/2)\gamma_i$ to characterize the activation difference evaluated at the optimal single point r_j^* of the mixture distribution versus the point r_i^* of the task. □

C.7 Proof of Proposition A.12

Proof. Fix a sign $\xi \in \{+1, -1\}$ and define for task parameter δ

$$\phi_{\xi,r}(\delta) = q((1+\xi\delta)r). \tag{114}$$

The first and second derivatives are $\phi'_{\xi,r}(\delta) = \xi r q'((1+\xi\delta)r)$, and $\phi''_{\xi,r}(\delta) = r^2 q''((1+\xi\delta)r)$. From $U = \sup_{z\geq 0}|q''(z)| < +\infty$,

$$|\phi''_{\xi,r}(\delta)| \leq Ur^2. \tag{115}$$

Write $\overline{\Delta}_w = w\Delta_1 + (1-w)\Delta_2$. Perform Taylor expansions of $\phi_{\xi,r}(\Delta_1)$ and $\phi_{\xi,r}(\Delta_2)$ about $\overline{\Delta}_w$:

$$\begin{aligned}\phi_{\xi,r}(\Delta_1) &= \phi_{\xi,r}(\overline{\Delta}_w) + \phi'_{\xi,r}(\overline{\Delta}_w)(\Delta_1 - \overline{\Delta}_w) + \mathcal{E}_1,\\ \phi_{\xi,r}(\Delta_2) &= \phi_{\xi,r}(\overline{\Delta}_w) + \phi'_{\xi,r}(\overline{\Delta}_w)(\Delta_2 - \overline{\Delta}_w) + \mathcal{E}_2,\end{aligned} \tag{116}$$

with remainder bounds $|\mathcal{E}_k| \le \frac{U}{2} r^2 (\Delta_k - \overline{\Delta}_w)^2,\ k = 1, 2.$

Multiply the first equation by w, the second by $1-w$, and sum. Since $w(\Delta_1 - \overline{\Delta}_w) + (1-w)(\Delta_2 - \overline{\Delta}_w) = 0$, all first-order terms cancel exactly, giving

$$\begin{aligned}&\left|w\phi_{\xi,r}(\Delta_1) + (1-w)\phi_{\xi,r}(\Delta_2) - \phi_{\xi,r}(\overline{\Delta}_w)\right|\\ &\le \frac{U}{2} r^2 \Big(w(\Delta_1 - \overline{\Delta}_w)^2 + (1-w)(\Delta_2 - \overline{\Delta}_w)^2\Big).\end{aligned} \tag{117}$$

Observe $\Delta_1 - \overline{\Delta}_w = -(1-w)(\Delta_2 - \Delta_1), \Delta_2 - \overline{\Delta}_w = w(\Delta_2 - \Delta_1)$. Hence

$$w(\Delta_1 - \overline{\Delta}_w)^2 + (1-w)(\Delta_2 - \overline{\Delta}_w)^2 = w(1-w)(\Delta_2 - \Delta_1)^2 =: V_w. \tag{118}$$

Substituting yields the pointwise bound Eq. (75). For measures satisfying $m_2(\rho) = \int r^2 \rho(dr) < +\infty$, apply the integral triangle inequality:

$$\begin{aligned}&\left|\int Q_w^{\pm}(r)\rho(dr) - \int q((1 \pm \overline{\Delta}_w)r)\rho(dr)\right|\\ &\le \int \left|Q_w^{\pm}(r) - q((1 \pm \overline{\Delta}_w)r)\right| \rho(dr) \le \frac{U}{2} m_2(\rho) V_w,\end{aligned} \tag{119}$$

which proves the integral bound in Eq. (76).

Next, we derive the risk error bound. Denote the exact mixed residual and the effective residual:

$$\boldsymbol{a} = \begin{pmatrix}\int Q_w^+ d\rho - 1\\ \int Q_w^- d\rho + 1\end{pmatrix}, \boldsymbol{b} = \begin{pmatrix}\int q((1+\overline{\Delta}_w)r) d\rho - 1\\ \int q((1-\overline{\Delta}_w)r) d\rho + 1\end{pmatrix}. \tag{120}$$

Let $\varepsilon_\rho = \frac{U}{2} m_2(\rho) V_w$. Then $|\boldsymbol{a}_i - \boldsymbol{b}_i| \le \varepsilon_\rho$, so $\|\boldsymbol{a} - \boldsymbol{b}\|_2 \le \sqrt{2}\,\varepsilon_\rho$. Using the difference of squares formula,

$$\left|\frac{1}{2}\|\boldsymbol{a}\|_2^2 - \frac{1}{2}\|\boldsymbol{b}\|_2^2\right| \le \frac{1}{2}\big(\|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2\big)\,\|\boldsymbol{a} - \boldsymbol{b}\|_2. \tag{121}$$

From $\|q\|_\infty \le B$, the components are bounded, so $\|\boldsymbol{a}\|_2, \|\boldsymbol{b}\|_2 = O(B+1)$. We obtain

$$\left|R_w^{\mathrm{mix}}(\rho) - R_{\overline{\Delta}_w}^{\infty}(\rho)\right| = \frac{1}{2}\left|\|\boldsymbol{a}\|_2^2 - \|\boldsymbol{b}\|_2^2\right| \le (B+1) U m_2(\rho) V_w, \tag{122}$$

which is Eq. (77). □

C.8 Proof of Proposition A.13

Proof. The proposition follows directly from the law of total probability. □

C.9 Proof of Proposition A.14

Proof. We treat exclusively the hard-routing limit $c \to \infty$. Let

$$p_1(s) := \lim_{c\to\infty} p_1(s; c), p_2(s) := \lim_{c\to\infty} p_2(s; c),$$

where the limits exist pointwise by Proposition A.1. The same proposition ensures that p_1, p_2 are continuous on $[0, +\infty)$, strictly monotone (p_1 increasing, p_2 decreasing), and satisfy

$$\lim_{s\to+\infty} p_1(s) = 1, \lim_{s\to+\infty} p_2(s) = 0.$$

Define the hard-routing cost function

$$h(s) \triangleq \frac{1}{2}\Big[p_1(s)R_1^{\infty}(\bar{\rho}_3) + (1-p_1(s))\,R_1^{\infty}(\bar{\rho}_4) + p_2(s)R_2^{\infty}(\bar{\rho}_4) + (1-p_2(s))\,R_2^{\infty}(\bar{\rho}_3)\Big]. \tag{123}$$

Since all risk terms are non-negative and p_1, p_2 are continuous, h is continuous on $[0,+\infty)$. Moreover, the above limits imply that h has a finite limit at infinity:

$$h(+\infty) := \lim_{s\to+\infty} h(s) = \frac{1}{2}\left[R_1^{\infty}(\bar{\rho}_3) + R_2^{\infty}(\bar{\rho}_3)\right]. \tag{124}$$

Thus, h admits a continuous extension to the compactification $[0,+\infty]$. As this extended domain is compact, there exists

$$s^* \in \arg\min_{s\in[0,+\infty]} h(s). \tag{125}$$

If $s^* = +\infty$, the optimal routing measure is supported at infinity, which corresponds to a degenerate limiting behavior. To obtain a finite optimal threshold, we impose the non-degeneracy condition

$$\inf_{s\in[0,+\infty)} h(s) < h(+\infty). \tag{126}$$

Under Eq. (126), the minimizer s^* lies in $[0,+\infty)$.

Now consider the risk functional

$$R_r(\rho_r) = \int_{[0,+\infty)} h(s)\,\rho_r(ds), \tag{127}$$

which is linear in ρ_r. For any $\rho_r \in \mathscr{P}([0,+\infty))$, monotonicity of integration yields

$$R_r(\rho_r) = \int h(s)\,\rho_r(ds) \geq \int h(s^*)\,\rho_r(ds) = h(s^*). \tag{128}$$

Taking the Dirac measure $\rho_r^* = \delta_{s^*}$ gives

$$R_r(\rho_r^*) = \int h(s)\,\delta_{s^*}(ds) = h(s^*), \tag{129}$$

so the lower bound is attained. Hence δ_{s^*} is an optimal routing measure.

If h is strictly convex on $[0,+\infty)$ (e.g., $h''(s) > 0$ whenever it exists), then the minimizer s^* is unique, and consequently the optimal routing measure is uniquely δ_{s^*}. □

C.10 Proof of Proposition 4.1

Proof. Let $g(t) := \theta(t) - \theta_*(\rho(t))$. By assumption, $t \mapsto \theta_*(\rho(t))$ is absolutely continuous; therefore, for t almost everywhere,

$$\dot{g}(t) = -\frac{1}{\epsilon}\nabla_\theta \mathcal{R}(\theta(t),\rho(t)) - \frac{d}{dt}\theta_*(\rho(t)). \tag{130}$$

When $g(t) \neq 0$,

$$\frac{d}{dt}\|g(t)\|_2 = \frac{\langle g(t), \dot{g}(t)\rangle}{\|g(t)\|_2}. \tag{131}$$

By the uniform strong monotonicity condition,

$$\langle g(t), \nabla_\theta \mathcal{R}(\theta(t),\rho(t))\rangle \geq \kappa_\theta \|g(t)\|_2^2, \tag{132}$$

and by the boundedness of the velocity of motion along the equilibrium branch,

$$\big\|\frac{d}{dt}\theta_*(\rho(t))\big\|_2 \leq C_\theta. \tag{133}$$

Therefore

$$\epsilon\frac{d}{dt}\|g(t)\|_2 \leq -\kappa_\theta\|g(t)\|_2 + \epsilon C_\theta. \tag{134}$$

At the point where $g(t) = 0$, using the Dini upper derivative, the above estimate still holds, so

$$\epsilon D^+ \|g(t)\|_2 \le -\kappa_\theta \|g(t)\|_2 + \epsilon C_\theta. \tag{135}$$

By Gronwall's comparison principle,

$$\|g(t)\|_2 \le e^{-\kappa_\theta t/\epsilon} \|g(0)\|_2 + \frac{C_\theta}{\kappa_\theta} \epsilon. \tag{136}$$

If the initial deviation is uniformly bounded, then after $T_\epsilon = O(\epsilon \log(1/\epsilon))$ initial layers,

$$\|\theta(t) - \theta_*(\rho(t))\|_2 = O(\epsilon). \tag{137}$$

Furthermore, by the uniform Lipschitz condition of the velocity field with respect to θ, there exists $L_\theta > 0$ such that

$$\|v_j(\theta, \rho) - v_j(\widetilde{\theta}, \rho)\|_{L^2(\rho_j)} \le L_\theta \|\theta - \widetilde{\theta}\|_2. \tag{138}$$

Let $v_j^*(\rho) := v_j(\theta_*(\rho), \rho)$, then

$$\|v_j(\theta(t), \rho(t)) - v_j^*(\rho(t))\|_{L^2(\rho_j(t))} \le L_\theta \|\theta(t) - \theta_*(\rho(t))\|_2 = O(\epsilon) \tag{139}$$

holds after the initial layer. Thus, we obtain the quasi-static reduction in the sense of the Wasserstein tangent space velocity field as stated in the proposition. □

C.11 Proof of Proposition 4.2

Proof. Let $d(t) := d_{\mathcal{W}}\big(\rho(t), \rho_*(\theta(t))\big)$. Fix a time t and temporarily freeze the routing parameters at $\theta(t)$. Let $\widehat{\rho}(\tau)$ denote the corresponding unscaled expert fast flow, with initial value $\widehat{\rho}(0) = \rho(t)$. By uniform local exponential attractivity,

$$D_\tau^+ d_{\mathcal{W}}\big(\widehat{\rho}(\tau), \rho_*(\theta(t))\big)\Big|_{\tau=0} \le -\kappa_\rho d(t). \tag{140}$$

In the actual slow-time equation, the expert velocity is amplified by $1/\epsilon$, so the instantaneous dissipative contribution of the expert state itself to $d(t)$ is $-(\kappa_\rho/\epsilon)d(t)$. On the other hand, a stable equilibrium branch satisfies

$$d_{\mathcal{W}}\big(\rho_*(\theta_1), \rho_*(\theta_2)\big) \le L_* \|\theta_1 - \theta_2\|_2. \tag{141}$$

Using the triangle inequality, we decompose $d(t + h) - d(t)$ into two parts: the motion of the expert state and the motion of the target equilibrium branch. When the joint flow possesses the local regularity required by the proposition, letting $h \downarrow 0$ yields

$$D^+ d(t) \le -\frac{\kappa_\rho}{\epsilon} d(t) + L_* \|\dot{\theta}(t)\|_2. \tag{142}$$

Since $\|\dot{\theta}(t)\|_2 \le K_\theta$,

$$D^+ d(t) \le -\frac{\kappa_\rho}{\epsilon} d(t) + L_* K_\theta. \tag{143}$$

Gronwall's comparison principle yields

$$d(t) \le e^{-\kappa_\rho t/\epsilon} d(0) + \frac{L_* K_\theta}{\kappa_\rho} \epsilon. \tag{144}$$

Therefore, after passing through an initial layer of thickness $T_\epsilon = O(\epsilon \log(1/\epsilon))$,

$$d_{\mathcal{W}}\big(\rho(t), \rho_*(\theta(t))\big) = O(\epsilon). \tag{145}$$

Since $\nabla_\theta \mathcal{R}$ is Lipschitz continuous with respect to ρ under $d_{\mathcal{W}}$, there exists an $L_\rho > 0$ such that

$$\|\nabla_\theta \mathcal{R}(\theta(t), \rho(t)) - \nabla_\theta \mathcal{R}(\theta(t), \rho_*(\theta(t)))\|_2 \le L_\rho d(t) = O(\epsilon). \tag{146}$$

Therefore,

$$\Big\|\dot{\theta}(t) + \eta_\theta \nabla_\theta \mathcal{R}\big(\theta(t), \rho_*(\theta(t))\big)\Big\|_2 \le \eta_\theta L_\rho d(t) = O(\epsilon), \tag{147}$$

where η_θ is regarded as a fixed constant independent of ϵ. Thus

$$\dot{\theta}(t) = -\eta_\theta \nabla_\theta \mathcal{R}\big(\theta(t), \rho_*(\theta(t))\big) + O(\epsilon). \tag{148}$$

□

C.12 Proof of Proposition 4.3

Proof. Write

$$\Phi_j(r) := \frac{\delta \mathcal{R}}{\delta \rho_j}(r).$$

Under the regularity assumptions of the proposition, the variational chain rule applies along the Wasserstein joint gradient flow:

$$\frac{d}{dt}\mathcal{R} = \left\langle \nabla_\theta \mathcal{R}, \dot{\theta} \right\rangle + \sum_{j=1}^{N} \left\langle \frac{\delta \mathcal{R}}{\delta \rho_j}, \partial_t \rho_j \right\rangle. \tag{149}$$

The second term is understood as a duality pairing in the distributional sense; for sufficiently smooth densities it reduces to the usual integral.

From $\dot{\theta} = -\eta_\theta \nabla_\theta \mathcal{R}$,

$$\left\langle \nabla_\theta \mathcal{R}, \dot{\theta} \right\rangle = -\eta_\theta \|\nabla_\theta \mathcal{R}\|_2^2. \tag{150}$$

On the expert side, the Wasserstein gradient flow reads $\partial_t \rho_j = \eta_j \nabla_r \cdot (\rho_j \nabla_r \Phi_j)$, so

$$\langle \Phi_j, \partial_t \rho_j \rangle = \eta_j \int_{\mathcal{W}_j} \Phi_j(r)\, \nabla_r \cdot (\rho_j \nabla_r \Phi_j)\, dr. \tag{151}$$

Under no-flux boundary conditions $\rho_j \nabla_r \Phi_j \cdot n = 0$ (and corresponding decay assumptions at infinity if $\mathcal{W}_j$ is unbounded), integration by parts yields

$$\langle \Phi_j, \partial_t \rho_j \rangle = -\eta_j \int_{\mathcal{W}_j} \|\nabla_r \Phi_j(r)\|_2^2\, \rho_j(dr). \tag{152}$$

Combining pieces:

$$\frac{d}{dt}\mathcal{R} = -\eta_\theta \|\nabla_\theta \mathcal{R}\|_2^2 - \sum_{j=1}^{N} \eta_j \int_{\mathcal{W}_j} \left\| \nabla_r \frac{\delta \mathcal{R}}{\delta \rho_j}(r) \right\|_2^2 \rho_j(dr) \le 0. \tag{153}$$

Hence $\mathcal{R}$ is a Lyapunov functional for the joint dynamics. □

C.13 Proof of Proposition 4.4

Proof. From the expert variational formula, set

$$\Gamma_{ij}(r) := \frac{\delta \ell_j}{\delta \rho_j}(x, y; \rho_j; r), \overline{\Gamma}_{ij}(r) := \mathbb{E}_{\mathbb{P}_i}[\Gamma_{ij}(r)]. \tag{154}$$

Since P_{ij} is independent of ρ_j,

$$\frac{\delta \mathcal{C}_{ij}}{\delta \rho_j}(r) = \mathbb{E}_{\mathbb{P}_i}[\pi_j \Gamma_{ij}(r)] - P_{ij} \overline{\Gamma}_{ij}(r) \tag{155}$$

$$= \mathbb{E}_{\mathbb{P}_i}\left[(\pi_j - P_{ij})(\Gamma_{ij}(r) - \overline{\Gamma}_{ij}(r)) \right]. \tag{156}$$

Therefore

$$\frac{\delta \mathcal{R}}{\delta \rho_j}(r) = \sum_i k_i P_{ij} \frac{\delta R_i(\rho_j)}{\delta \rho_j}(r) + \sum_i k_i \frac{\delta \mathcal{C}_{ij}}{\delta \rho_j}(r) \tag{157}$$

$$= m_j \frac{\delta}{\delta \rho_j}\Big[\sum_i \zeta_{ij} R_i(\rho_j) \Big] + \sum_i k_i \frac{\delta \mathcal{C}_{ij}}{\delta \rho_j}(r), \tag{158}$$

where we used $m_j \zeta_{ij} = k_i P_{ij}$. This establishes the decomposition on the expert side.

For the routing side, differentiate the decomposition in Eq. (8) with respect to θ:

$$\nabla_\theta \mathcal{R} = \sum_i k_i \sum_j \nabla_\theta P_{ij} R_i(\rho_j) + \sum_i \sum_j k_i \nabla_\theta \mathcal{C}_{ij}. \tag{159}$$

Choose a reference expert b_i for each task i. Using $\sum_j \nabla_\theta P_{ij} = 0$,

$$\sum_j \nabla_\theta P_{ij} R_i(\rho_j) = \sum_{j \neq b_i} \nabla_\theta P_{ij} \big(R_i(\rho_j) - R_i(\rho_{b_i})\big) = -\sum_{j \neq b_i} \nabla_\theta P_{ij}\, \mathcal{A}_i^{j \succ b_i}. \tag{160}$$

Thus

$$\nabla_\theta \mathcal{R} = -\sum_i k_i \sum_{j \neq b_i} \nabla_\theta P_{ij} \mathcal{A}_i^{j \succ b_i} + \sum_i \sum_j k_i \nabla_\theta \mathcal{C}_{ij}. \tag{161}$$

Substitute $\dot\theta = -\eta_\theta \nabla_\theta \mathcal{R}$ to obtain Eq. (25). □

C.14 Proof of Proposition 4.5

Proof. Define gradient fields for the sample loss:

$$h_u(x, y; \rho_j; r) := \nabla_r \frac{\delta \ell_j}{\delta \rho_j}(x, y; \rho_j; r), g_u(\rho_j; r) := \mathbb{E}_{\mathbb{P}_u}[h_u(x, y; \rho_j; r)]. \tag{162}$$

According to Eq. (14), $\mathbb{E}_{\mathbb{P}_u}[\pi_j h_u] = P_{uj} g_u(\rho_j) + \mathcal{C}_{uj}^{(g)}$. Hence, the expert velocity field can be written as

$$v_j(\theta, \rho) = -\eta_E \sum_{u=1}^M k_u \left[P_{uj} g_u(\rho_j) + \mathcal{C}_{uj}^{(g)}\right]. \tag{163}$$

Take the total time derivative of $\mathcal{A}_i^{j \succ k} = R_i(\rho_k) - R_i(\rho_j)$. By the chain rule for Wasserstein gradient flows, for any expert measure ρ_j,

$$\frac{d}{dt} R_i(\rho_j) = \big\langle v_j, g_i(\rho_j) \big\rangle_{L^2(\rho_j)}. \tag{164}$$

Therefore

$$\dot{\mathcal{A}}_i^{j \succ k} = \big\langle v_k, g_i(\rho_k) \big\rangle_{L^2(\rho_k)} - \big\langle v_j, g_i(\rho_j) \big\rangle_{L^2(\rho_j)}. \tag{165}$$

Insert the velocity expressions for both experts:

$$\begin{aligned} \dot{\mathcal{A}}_i^{j \succ k} &= \eta_E \sum_{u=1}^M k_u \Big[P_{uj} \big\langle g_i(\rho_j), g_u(\rho_j) \big\rangle_{L^2(\rho_j)} - P_{uk} \big\langle g_i(\rho_k), g_u(\rho_k) \big\rangle_{L^2(\rho_k)}\Big] \\ &\quad + \eta_E \sum_{u=1}^M k_u \Big[\big\langle g_i(\rho_j), \mathcal{C}_{uj}^{(g)} \big\rangle_{L^2(\rho_j)} - \big\langle g_i(\rho_k), \mathcal{C}_{uk}^{(g)} \big\rangle_{L^2(\rho_k)}\Big], \end{aligned} \tag{166}$$

which is Eq. (26).

Recall

$$\mathcal{C}_{uj}^{(g)} = \mathbb{E}_{\mathbb{P}_u}\left[(\pi_j - P_{uj}) \left(\nabla_r \frac{\delta \ell_j}{\delta \rho_j} - g_u(\rho_j)\right)\right]. \tag{167}$$

By the Cauchy–Schwarz inequality,

$$\|\mathcal{C}_{uj}^{(g)}\|_{L^2(\rho_j)} \le \sigma_{uj}^{\pi} \sigma_{uj}^{g}. \tag{168}$$

Apply Cauchy–Schwarz to $\mathcal{X}_i^{j \succ k}$ to obtain

$$|\mathcal{X}_i^{j \succ k}| \le \eta_E \mathcal{B}_i^{j \succ k}. \tag{169}$$

This completes the proof. □

C.15 Proof of Proposition 4.6

Proof.

$$P_{ij} G_{ii}^{(j)} - P_{ik} G_{ii}^{(k)} = (P_{ij} - P_{ik}) G_{ii}^{(j)} + P_{ik}\big(G_{ii}^{(j)} - G_{ii}^{(k)}\big). \tag{170}$$

Using the assumed bounds,

$$P_{ij} G_{ii}^{(j)} - P_{ik} G_{ii}^{(k)} \ge a g_{\min} - P_{ik}(g_{\max} - g_{\min}). \tag{171}$$

For each $u \neq i$, we have

$$P_{uj} G_{iu}^{(j)} - P_{uk} G_{iu}^{(k)} \ge -\gamma g_{\max}(P_{uj} + P_{uk}). \tag{172}$$

Summing over u yields the stated lower bound on $\mathcal{D}_i^{j \succ k}$. Combining it with $|\mathcal{X}_i^{j \succ k}| \le \eta_E \mathcal{B}_i^{j \succ k}$ in Proposition 4.5 gives $\dot{\mathcal{A}}_i^{j \succ k} \ge \eta_E \nu$. Integration over $[t_0, t]$ completes the proof. □

C.16 Proof of Proposition 4.7

Proof. Define

$$v_{j^*}^{\text{mix}} := -\eta_{j^*} \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i} \left[\nabla_r \frac{\delta \ell_{j^*}}{\delta \rho_{j^*}} \right]. \tag{173}$$

From the joint risk $\mathcal{R} = \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i} \left[\sum_{j=1}^{N} \pi_j(x;\theta) \ell_j(x,y;\rho_j) \right]$, the routing gradient is

$$\nabla_\theta \mathcal{R} = \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i} \big[\sum_{j=1}^{N} \nabla_\theta \pi_j(x;\theta) \ell_j(x,y;\rho_j) \big]. \tag{174}$$

Therefore,

$$\|\nabla_\theta \mathcal{R}\| \le \sum_{i=1}^{M} k_i \sum_{j=1}^{N} C_\pi \epsilon B_\ell = N C_\pi B_\ell \epsilon, \tag{175}$$

hence $\|\dot{\theta}\| = \eta_\theta \|\nabla_\theta \mathcal{R}\| \le \eta_\theta N C_\pi B_\ell \epsilon$.

For an expert j,

$$\nabla_r \frac{\delta \mathcal{R}}{\delta \rho_j} = \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i} \big[\pi_j(x;\theta) \nabla_r \frac{\delta \ell_j}{\delta \rho_j} \big]. \tag{176}$$

Thus, for any inactive expert $j \neq j^*$,

$$\big\| \nabla_r \frac{\delta \mathcal{R}}{\delta \rho_j} \big\|_{L^2(\rho_j)} \le B_g \sum_{i=1}^{M} k_i P_{ij}(\theta) \le B_g \epsilon, \tag{177}$$

which gives $\|v_j\|_{L^2(\rho_j)} \le \eta_j B_g \epsilon$.

For the active expert,

$$v_{j^*} = -\eta_{j^*} \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i} \big[\pi_{j^*}(x;\theta) \nabla_r \frac{\delta \ell_{j^*}}{\delta \rho_{j^*}} \big]. \tag{178}$$

Subtracting $v_{j^*}^{\text{mix}}$ gives

$$v_{j^*} - v_{j^*}^{\text{mix}} = \eta_{j^*} \sum_{i=1}^{M} k_i \mathbb{E}_{\mathbb{P}_i} \big[(1 - \pi_{j^*}(x;\theta)) \nabla_r \frac{\delta \ell_{j^*}}{\delta \rho_{j^*}} \big]. \tag{179}$$

Since $\mathbb{E}_{\mathbb{P}_i}[1 - \pi_{j^*}(x;\theta)] = 1 - P_{ij^*}(\theta) \le \epsilon$, we obtain

$$\|v_{j^*} - v_{j^*}^{\text{mix}}\|_{L^2(\rho_{j^*})} \le \eta_{j^*} B_g \sum_{i=1}^{M} k_i (1 - P_{ij^*}) \le \eta_{j^*} B_g \epsilon. \tag{180}$$

Taking the maximum over $j \neq j^*$, we obtain

$$\max_{j \neq j^*} \|v_j\|_{L^2(\rho_j)} \le \eta_{\max} B_g \epsilon, \eta_{\max} := \max_{j \neq j^*} \eta_j. \tag{181}$$

This completes the proof. □

C.17 Proof of Proposition 4.8

Proof. By definition

$$A(\theta) = \frac{1}{2} \sum_{j=1}^{N} \left(m_j(\theta) - \frac{1}{N} \right)^2, \mathcal{R}_\lambda(\theta,\rho) = \mathcal{R}(\theta,\rho) + \lambda A(\theta),$$

so $\nabla_\theta \mathcal{R}_\lambda = \nabla_\theta \mathcal{R} + \lambda \nabla_\theta A$. Thus $\dot{\theta} = -\eta_\theta \nabla_\theta \mathcal{R} - \eta_\theta \lambda \nabla_\theta A$.

Let θ_c be a routing equilibrium for fixed expert states ρ, i.e. $\nabla_\theta \mathcal{R}(\theta_c, \rho) = 0$. If $\nabla_\theta A(\theta_c) \neq 0$, then

$$\nabla_\theta \mathcal{R}_\lambda(\theta_c, \rho) = \lambda \nabla_\theta A(\theta_c) \neq 0,$$

so θ_c is no longer a stationary point after adding regularization.

Now suppose $\nabla_\theta A(\theta_c) = 0$. Then θ_c remains a stationary point in the routing direction, and

$$\nabla^2_\theta \mathcal{R}_\lambda(\theta_c, \rho) = \nabla^2_\theta \mathcal{R}(\theta_c, \rho) + \lambda \nabla^2_\theta A(\theta_c). \tag{182}$$

Suppose that there exists a unit vector u such that $b_u := u^\top \nabla^2_\theta A(\theta_c) u < 0$, and set $a_u := u^\top \nabla^2_\theta \mathcal{R}(\theta_c, \rho) u$, then

$$u^\top \nabla^2_\theta \mathcal{R}_\lambda(\theta_c, \rho) u = a_u + \lambda b_u. \tag{183}$$

Whenever

$$\lambda > \frac{\max\{a_u, 0\}}{-b_u}, \tag{184}$$

we get $a_u + \lambda b_u < 0$. Hence, the augmented objective $\mathcal{R}_\lambda(\cdot, \rho)$ has a strictly negative curvature along direction u at θ_c, so θ_c cannot be a local minimum for fixed ρ.

Further, if (θ_c, ρ) itself is a stationary point of the joint system and the joint objective satisfies the standard C^2 local linearization conditions at this point, the direction of negative curvature gives an unstable direction for the gradient flow. Consequently, this collapsing stationary point cannot be a locally asymptotically stable equilibrium of the gradient flow. □

D Auxiliary Lemmas

Lemma D.1. *Define the set $\mathcal{D} \subset [0, 1]$:*

$$\mathcal{D} = \{\Delta : \exists r \in (0, +\infty), \text{s.t. } q^+(r; \Delta) \geq 1,\ q^-(r; \Delta) \leq -1\} \tag{185}$$

Let $\Delta^\infty = \inf_{\Delta \in \mathcal{D}} \Delta$ and $\Delta^\infty \in (0, 1)$. If $\Delta < \Delta^\infty$, then for all $\rho \in \mathscr{P}([0, +\infty])$, we have $R^\infty(\rho; \Delta) > 0$, and the global minimum of $R^\infty(\rho; \Delta)$ is attained at some point measure δ_{r^*}. Otherwise, if $\Delta \geq \Delta^\infty$, there exists some ρ^* such that $R^\infty(\rho^*; \Delta) = 0$, and $\mathrm{supp}(\rho^*)$ is compact.

Proof. Let $\gamma(r; \Delta) = \big(q^+(r; \Delta), q^-(r; \Delta)\big)$, $C(\Delta) = \mathrm{conv}\{\gamma(r; \Delta) : r \in [0, +\infty]\}$. Then $R^\infty(\rho; \Delta) = \frac{1}{2} \left\| \int \gamma(r; \Delta)\, d\rho(r) - (1, -1) \right\|^2$.

First suppose $\Delta < \Delta^\infty$. Since $\Delta \notin \mathcal{D}$ and $q^-(\cdot; \Delta)$ is strictly increasing, there exists a unique r_- such that $q^-(r_-; \Delta) = -1,\ q^+(r_-; \Delta) < 1$. Define $\lambda_* = \frac{\partial_r q^+(r_-; \Delta)}{\partial_r q^-(r_-; \Delta)} > 0, h(r) = q^+(r; \Delta) - \lambda_* q^-(r; \Delta)$.

From the derivative ratio

$$\frac{\partial_r q^-(r; \Delta)}{\partial_r q^+(r; \Delta)} = \frac{1 - \Delta}{1 + \Delta} \frac{q'((1 - \Delta)r)}{q'((1 + \Delta)r)} \tag{186}$$

which is strictly increasing in r, it follows that h is strictly increasing before r_- and strictly decreasing after r_-. Hence r_- is the unique maximizer of h. Consequently, for any probability measure ρ satisfying $\int q^-(r; \Delta)\, d\rho \leq -1$, we have $\int q^+(r; \Delta)\, d\rho - 1 \leq q^+(r_-; \Delta) - 1 < 0$.

Thus no probability measure can simultaneously satisfy $\int q^+(r; \Delta)\, d\rho = 1,\ \int q^-(r; \Delta)\, d\rho = -1$. Equivalently, $(1, -1) \notin C(\Delta)$. Since $C(\Delta)$ is a compact convex set, $\inf_{\rho \in \mathscr{P}([0, +\infty])} R^\infty(\rho; \Delta) = \frac{1}{2} \mathrm{dist}^2\big((1, -1), C(\Delta)\big) > 0$.

Let ρ^* denote a global optimal measure, and set $a_+ = \int q^+(r; \Delta)\, d\rho^* - 1, a_- = \int q^-(r; \Delta)\, d\rho^* + 1$. From $(1, -1) \notin C(\Delta)$ we have $(a_+, a_-) \neq (0, 0)$. The first-order optimality condition for projection onto the moment space gives

$$\mathrm{supp}(\rho^*) \subseteq \arg\min_{r \in [0, +\infty]} \big\{a_+ q^+(r; \Delta) + a_- q^-(r; \Delta)\big\}. \tag{187}$$

Using the two diagonal endpoints of the curve together with convexity, the optimal residual must satisfy $a_+ < 0 < a_-$.

Define $\psi(r;\Delta) = a_+ q^+(r;\Delta) + a_- q^-(r;\Delta)$. Then

$$\psi'(r;\Delta) = \partial_r q^+(r;\Delta)\left[a_+ + a_- \frac{\partial_r q^-(r;\Delta)}{\partial_r q^+(r;\Delta)}\right]. \tag{188}$$

The function inside the brackets is strictly increasing in r, so $\psi'(\cdot;\Delta)$ has at most one zero. Therefore, $\psi(\cdot;\Delta)$ admits a unique minimizer r^* on $[0,+\infty]$, which yields $\rho^* = \delta_{r^*}$.

Claim. $\mathcal{D} \subset [0,1]$ is a closed and upward closed subset.

Upward closed: Clearly $1 \in \mathcal{D}$. For $\Delta < 1$, suppose $\Delta \in \mathcal{D}$, so there exists $r > 0$ with $q^+(r;\Delta) \geq 1$, $q^-(r;\Delta) \leq -1$. Since $q^+(\cdot;\cdot)$ is increasing and $q^-(\cdot;\cdot)$ is decreasing in Δ, for any $\Delta' \geq \Delta$ we have

$$q^+(r;\Delta') \geq q^+(r;\Delta) \geq 1, q^-(r;\Delta') \leq q^-(r;\Delta) \leq -1,$$

hence $\Delta' \in \mathcal{D}$. Therefore, $\mathcal{D}$ is upward closed.

Closedness: Let $\{\Delta_n\} \subset \mathcal{D}$ and $\Delta_n \to \Delta \in [0,1]$. By the definition of $\mathcal{D}$, for each n there exists $r_n \in (0,\infty)$ such that

$$q^+(r_n;\Delta_n) \geq 1, q^-(r_n;\Delta_n) \leq -1.$$

If $\Delta = 1$, then since $1 \in \mathcal{D}$, we have $\Delta \in \mathcal{D}$ and the closedness is trivial. Hence, assume $\Delta < 1$. Consider $\{r_n\}$ in the compactified space $[0,+\infty]$. Pass to a convergent subsequence $r_{n_k} \to r^* \in [0,+\infty]$. We rule out boundary cases: from the model assumptions $q(0) < 1$, $q(\infty) > -1$, continuity gives

$$q^+(0;\Delta) < 1, q^-(\infty;\Delta) > -1.$$

If $r^* = 0$, then $q^+(r_{n_k};\Delta_{n_k}) \to q^+(0;\Delta) < 1$, contradicting $q^+(r_{n_k};\Delta_{n_k}) \geq 1$. If $r^* = +\infty$, then $q^-(r_{n_k};\Delta_{n_k}) \to q^-(\infty;\Delta) > -1$, contradicting $q^-(r_{n_k};\Delta_{n_k}) \leq -1$. Consequently $r^* \in (0,+\infty)$. By joint continuity of $q^\pm(r;\Delta)$ in both arguments,

$$q^+(r^*;\Delta) = \lim_{k\to\infty} q^+(r_{n_k};\Delta_{n_k}) \geq 1, q^-(r^*;\Delta) = \lim_{k\to\infty} q^-(r_{n_k};\Delta_{n_k}) \leq -1.$$

This yields $\Delta \in \mathcal{D}$. Hence $\mathcal{D}$ is closed in $[0,1]$.

Since $\mathcal{D}$ is closed and upward closed, the infimum $\Delta^\infty = \inf_{\Delta\in\mathcal{D}}$ satisfies $\Delta \in \mathcal{D}$ for all $\Delta \geq \Delta^\infty$.

Now suppose $\Delta \geq \Delta^\infty$. Since $\mathcal{D}$ is a closed upper interval, $\Delta \in \mathcal{D}$, there exists $r_0 \in (0,+\infty)$ such that $q^+(r_0;\Delta) \geq 1,\ q^-(r_0;\Delta) \leq -1$.

If $\gamma(r_0;\Delta) = (1,-1)$, simply take $\rho^* = \delta_{r_0}$. Otherwise, the straight line connecting $\gamma(r_0;\Delta)$ to the diagonal $x = y$ passes through $(1,-1)$; denote its intersection with the diagonal by (z,z). From $q(0) < -1 < 1 < q(+\infty)$ we obtain $z \in \big(q(0), q(+\infty)\big)$. Thus, (z,z) lies in the relative interior of $\mathrm{conv}\big\{\gamma(0;\Delta), \gamma(+\infty;\Delta)\big\}$, and $(1,-1) \in \mathrm{conv}\big\{\gamma(0;\Delta), \gamma(r_0;\Delta), \gamma(+\infty;\Delta)\big\}$.

By continuity of $\gamma(\cdot;\Delta)$, replace the endpoints 0 and $+\infty$ with a sufficiently small finite point $r_1 > 0$ and a sufficiently large finite point $r_3 < +\infty$. Hence, there exist $a_j \geq 0$, $\sum_{j=1}^3 a_j = 1$, such that $\sum_{j=1}^3 a_j \gamma(r_j;\Delta) = (1,-1)$. Set $\rho^* = \sum_{j=1}^3 a_j \delta_{r_j}$. Then $R^\infty(\rho^*;\Delta) = 0$ and $\mathrm{supp}(\rho^*) \subseteq \{r_1, r_0, r_3\}$, so $\mathrm{supp}(\rho^*)$ is compact. □

Lemma D.2. *Suppose the conditional distributions $x \mid y = +1$ and $x \mid y = -1$ are both multivariate Gaussians with a mean of 0. Consider the logistic linear router:*

$$\pi(x;s) = \mathrm{sigm}\big(\langle s, x\rangle\big).$$

Then

$$\mathbb{E}\big[\pi(x;s) \mid y = +1\big] = \mathbb{E}\big[\pi(x;s) \mid y = -1\big] = \frac{1}{2}.$$

Proof. Let $Z_y = \langle s, x\rangle \mid y$. By assumption, $x \mid y$ is Gaussian with a mean of zero, so the linear transformation yields a Gaussian random variable Z_y with

$$\mathbb{E}[Z_y] = \langle s, \mathbb{E}[x \mid y]\rangle = 0, \forall y \in \{+1,-1\}.$$

Write $Z_y \sim \mathcal{N}(0,\sigma_y^2)$; the variance σ_y^2 may depend on the label y.

For any Gaussian random variable $Z \sim \mathcal{N}(0, \sigma^2)$, we have the identity

$$\mathbb{E}\big[\mathrm{sigm}(Z)\big] = \frac{1}{2},$$

independent of σ^2. Substitute $y = \pm 1$ to obtain

$$\mathbb{E}\big[\pi(x;s) \mid y = +1\big] = \mathbb{E}\big[\mathrm{sigm}(Z_+)\big] = \frac{1}{2}, \mathbb{E}\big[\pi(x;s) \mid y = -1\big] = \mathbb{E}\big[\mathrm{sigm}(Z_-)\big] = \frac{1}{2}.$$

This finishes the proof. □

Lemma D.3. *Suppose that there exist a collapse boundary state θ_c and an expert j^* such that for all tasks i, $P_{ij^*}(\theta_c) = 1,\ P_{ij}(\theta_c) = 0\ (j \neq j^*)$. If the routing degenerates to a regime of zero gradient at this boundary, i.e. $\nabla_\theta \pi_j(x;\theta_c) = 0, \mathbb{P}_i$ -a.s., $\forall i, j$, then*

$$\nabla_\theta \mathcal{R}(\theta_c, \rho) = 0. \tag{189}$$

All inactive experts $j \neq j^*$ stop updating, while the active expert j^* is effectively trained on the mixture distribution $\mathbb{P} = \sum_i k_i \mathbb{P}_i$. Therefore, this collapse boundary forms an invariant set of the joint gradient flow.

Proof. From $0 \leq \pi_j \leq 1$ and $P_{ij^*}(\theta_c) = 1$, $P_{ij}(\theta_c) = 0$, for each task i we have

$$\pi_{j^*}(x;\theta_c) = 1, \pi_j(x;\theta_c) = 0 (j \neq j^*) \tag{190}$$

$\mathbb{P}_i$-almost everywhere. Together with the pointwise condition $\nabla_\theta \pi_j(x;\theta_c) = 0$, the routing gradient satisfies

$$\nabla_\theta \mathcal{R} = \sum_{i=1}^{M} k_i\, \mathbb{E}_{\mathbb{P}_i}\big[\sum_{j=1}^{N} \nabla_\theta \pi_j(x;\theta) \ell_j(x,y;\rho_j)\big] = 0, \tag{191}$$

Hence $\dot{\theta}(t) = 0$ whenever $\theta(t) = \theta_c$. For inactive experts $j \neq j^*$, we obtain

$$\frac{\delta \mathcal{R}}{\delta \rho_j}(r) = \sum_i k_i \mathbb{E}_{\mathbb{P}_i}\big[\pi_j(x;\theta_c) \frac{\delta \ell_j}{\delta \rho_j}(r)\big] = 0, \tag{192}$$

so their Wasserstein velocity field vanishes and $\partial_t \rho_j = 0$.

For the active expert j^*, $\pi_{j^*} = 1$ almost everywhere, so its variational derivative becomes

$$\frac{\delta \mathcal{R}}{\delta \rho_{j^*}}(r) = \sum_i k_i \mathbb{E}_{\mathbb{P}_i}\big[\frac{\delta \ell_{j^*}}{\delta \rho_{j^*}}(r)\big], \tag{193}$$

which is exactly the expert risk gradient for the mixture distribution $\mathbb{P} = \sum_i k_i \mathbb{P}_i$. Therefore, the active expert is effectively trained on the full mixed dataset.

In summary, this collapse boundary is invariant under the joint gradient flow. □