SPECIALIZATION AND SLOW RECOVERY IN JOINTLY TRAINED MIXTURES OF EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Marginal load balance alone does not determine whether experts specialize; the prediction objective also matters. We study a family of squared losses that interpolates between selected-expert and averaged-output training. For two jointly trained scalar linear experts with a binary routing cue, we prove a global phase diagram under positive balancing. When $0 < q < 1$, the thresholds $k = qg^2$ and $k = g^2/q$ separate full specialization, stable partial specialization, and homogenization. Here $k = \lambda\chi$ combines balancing strength with batch-cue imbalance, and g is the cue-conditional regression signal. For every $q > 0$, expert parameters remain bounded, asymptotic starvation is excluded, and every trajectory from a finite initialization converges to a single state. At $q = 1$, each trajectory at the critical threshold selects one point on a stationary continuum. For multiple smooth experts with fixed-feature softmax routing, we derive a local Hessian criterion that separates expert-separation curvature, routing–residual coupling, and balancing curvature. At $q = 0$, the binary linear objective has no finite minimizer when the routing-contrast penalty is positive. For initial conditions in a specified invariant set, expert parameters diverge while routing approaches uniformity and cue-dependent prediction persists. For nonlinear experts, curvature can alter stability at the $q = 0$ boundary. For selected-expert training, we also prove matching population load-recovery bounds of order $e^{M/T}$ from an initial logit gap M. For simultaneous no-baseline categorical updates, a corrected potential controls shared router–expert noise and yields a recovery-time second-moment bound under explicit bounded-data and step-size conditions.

1 INTRODUCTION

Sparse mixture-of-experts (MoE) models increase capacity through conditional computation (Shazeer et al., 2017; Fedus et al., 2022). Balanced utilization alone does not imply specialization: similar marginal loads can coexist with different expert functions (Guo et al., 2025; Li et al., 2026). Auxiliary balancing losses act on batch routing statistics, so batch composition can change the cost of input-dependent routing even at identical population loads (Fedus et al., 2022; Qiu et al., 2025; Chen et al., 2026). Specialization therefore depends on prediction, routing, and balancing together.

Selected-expert and averaged-output squared losses differ by an expert-output variance term (Jacobs et al., 1991; Krogh & Vedelsby, 1994). We ask how varying this term, with routing and balancing fixed, changes limiting states, stability, and finite-parameter attainability under joint router–expert training. Long-run non-starvation does not determine recovery time: under softmax routing, an almost unused expert can remain unlikely to be selected (Li et al., 2023; Klein et al., 2024; Mei & Osband, 2026). Joint training couples recovery to expert motion and shared update noise.

We study two scalar linear experts with a softmax router observing a noisy binary cue. The objective $\mathcal{R}_q = q\mathcal{R}_{\text{sel}} + (1-q)\mathcal{R}_{\text{avg}}$, $q \in [0,1]$, varies the prediction loss at fixed routing and balancing. For $q > 0$ and positive balancing, we classify every finite joint-flow trajectory. With $k = \lambda\chi$ as the effective routing-contrast penalty, the thresholds qg^2 and g^2/q separate almost-everywhere full specialization, stable partial specialization, and global homogenization for $0 < q < 1$. Experts remain bounded and marginal loads approach $1/2$; at $q = 1$, each critical trajectory selects one point on a stationary continuum.

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.

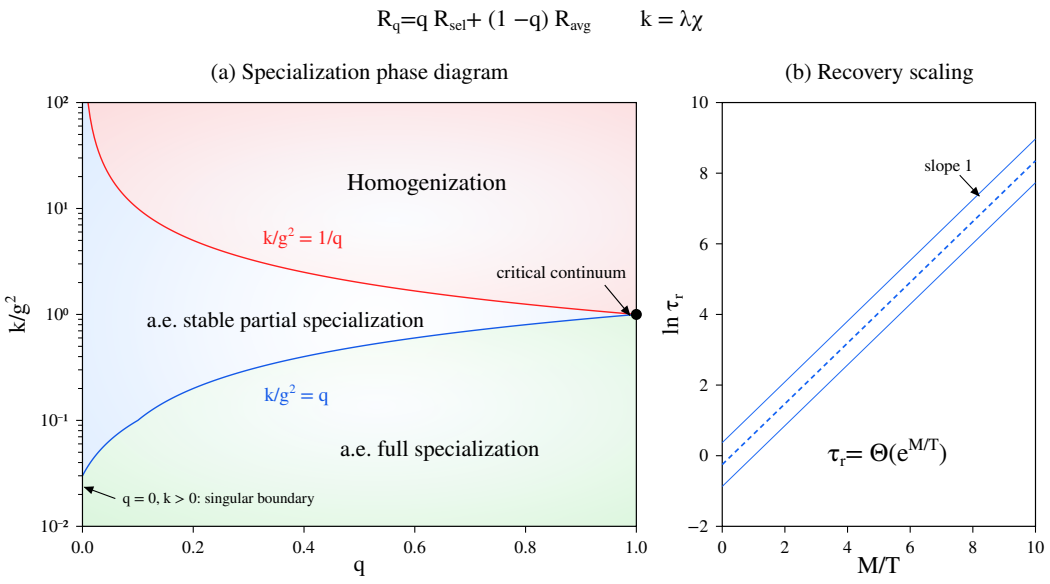


Figure 1: **Specialization regimes and recovery scale.** **(a)** In the binary model, the prediction parameter q and effective balancing strength $k = \lambda\chi$ select full specialization, partial specialization, or homogenization across the boundaries $k/g^2 = q$ and $k/g^2 = 1/q$. The point $(q, k/g^2) = (1, 1)$ has a critical continuum; $q = 0$, $k > 0$ is a singular non-attainment boundary. **(b)** Under selected-expert training, recovery from an initial logit gap M to a fixed positive load takes $\Theta(e^{M/T})$ population time. The parallel slope-one guides depict constant-factor bounds schematically, not empirical fits.

Contributions. Our contributions are twofold.

- **Global specialization dynamics.** For $q > 0$ with positive balancing, we classify full, partial, and homogeneous limits of the binary joint flow, including critical point selection and exclusion of expert divergence and asymptotic starvation. We also derive a local multi-expert curvature–coupling criterion and characterize the singular linear $q = 0$ boundary.
- **Slow recovery from softmax saturation.** Under selected-expert training, recovery from a logit gap M takes $\Theta(e^{M/T})$ population time. For simultaneous no-baseline categorical updates, a corrected potential gives a matching pathwise lower scale and a recovery-time second-moment bound despite shared router–expert noise.

1.1 Relation to prior work

Classical MoE work distinguished the two prediction losses (Jacobs et al., 1991), while ensemble ambiguity describes their squared-loss difference (Krogh & Vedelsby, 1994). Prior theory studies specialist emergence, latent clusters, soft routing, feature learning, and continual learning (Kang & Oh, 1996; Chen et al., 2022; Kawata et al., 2025; Liao & Kyrillidis, 2026; Li et al., 2025). Our global joint-dynamics result concerns the binary model; unlike Liao & Kyrillidis (2026), the local multi-expert criterion fixes routing features.

Prior work shows that balancing scope affects specialization (Qiu et al., 2025); our binary dynamics make batch composition explicit through χ and show how $k = \lambda\chi$ interacts with the prediction parameter q. Other studies examine balancing rules, routing geometry, and expert diversity (Chen et al., 2026; Wang et al., 2026; Guo et al., 2025; Li et al., 2026). Softmax analyses study saturation and estimator choice (Li et al., 2023; Garg et al., 2022; Klein et al., 2024; Mei & Osband, 2026; Kiselev, 2026); our recovery bounds treat coupled router–expert updates with no-baseline categorical dispatch.

2 JOINT TRAINING AND OBJECTIVE-DEPENDENT GLOBAL DYNAMICS

2.1 BINARY PROBLEM SETUP AND PREDICTION-DISPERSION OBJECTIVE

We study two scalar linear experts trained jointly with a softmax router. This model permits a global classification while separating cue-dependent routing, expert differentiation, and marginal load balance. The binary cue and scalar experts each reduce a contrast to one dimension; distinct router and expert inputs let us identify the two contrasts separately. The teacher construction gives an interpretable signal model, but the global result requires only (2).

For illustration, a uniform latent task $Z \in \{-1, 1\}$ generates a noisy cue $S \in \{-1, 1\}$ with $\mathbb{P}(S = Z \mid Z) = (1 + \rho)/2$, $0 < \rho < 1$. Let $Y = vZU + \xi$, where $v > 0$, $\mathbb{E}U^2 = 1$, $\mathbb{E}\xi = 0$, $\mathbb{E}\xi^2 < \infty$, and $U, \xi, (Z, S)$ are mutually independent. Thus $g = \rho v > 0$ and $\mathbb{E}[Y \mid S, U] = gSU$. The experts predict $f_\pm = w_\pm U$, and the router selects the $+$ expert with probability

$$p_+(S) = \mathrm{sigmoid}((\beta + \alpha S)/T), \qquad x = p_+(1), \quad y = p_+(-1), \qquad T > 0. \tag{1}$$

Here $\mathrm{sigmoid}(z) = (1 + e^{-z})^{-1}$, while x and y are the conditional probabilities of selecting the $+$ expert for the two cue values. Finite logits give $x, y \in (0, 1)$.

The global phase diagram requires only the following moment conditions:

$$\mathbb{P}(S = \pm 1) = \tfrac{1}{2}, \quad \mathbb{E}[U^2 \mid S] = 1, \quad \mathbb{E}[UY \mid S] = gS, \quad \mathbb{E}Y^2 < \infty. \tag{2}$$

The teacher model satisfies these conditions and is used again for recovery in Section 4; its independence assumptions are not needed for the global classification. Write $\bar{f} = p_+ f_+ + (1 - p_+) f_-$ and define the selected-expert and averaged-output risks by

$$\mathcal{R}_{\mathrm{sel}} = \tfrac{1}{2}\mathbb{E}[p_+(f_+ - Y)^2 + (1 - p_+)(f_- - Y)^2], \qquad \mathcal{R}_{\mathrm{avg}} = \tfrac{1}{2}\mathbb{E}[(\bar{f} - Y)^2].$$

Their difference is half the conditional expert-output variance, a classical distinction (Jacobs et al., 1991) related to ensemble ambiguity (Krogh & Vedelsby, 1994). We vary its weight at fixed routing and balancing to isolate the prediction objective. For $0 \le q \le 1$, define

$$\mathcal{R}_q = \mathcal{R}_{\mathrm{avg}} + \frac{q}{2}\mathbb{E}[\mathrm{Var}_{J \sim p(\cdot|S)}(f_J \mid U, S)] = q\mathcal{R}_{\mathrm{sel}} + (1 - q)\mathcal{R}_{\mathrm{avg}}. \tag{3}$$

Here the variance is over expert selection conditional on the example, distinct from stochastic-gradient variance. Thus $q = 1$ gives selected-expert loss, whereas $q = 0$ gives deterministic averaged-output loss. If L independent selected outputs are averaged for one input before squared loss is applied, expectation over selections gives $q = 1/L$.

For a batch of B cues, let $\bar{S} = B^{-1}\sum_i S_i$ and $\chi = \mathbb{E}\bar{S}^2 \in [0, 1]$. Each cue has the population marginal, but cues within a batch may be dependent. With balancing weight λ, the expected quadratic penalty on the batch-mean router load is

$$\lambda\mathbb{E}[(\bar{p} - \tfrac{1}{2})^2 + (1 - \bar{p} - \tfrac{1}{2})^2], \qquad \bar{p} = B^{-1}\sum_i p_+(S_i).$$

Independent cues give $\chi = 1/B$; exactly cue-balanced batches give $\chi = 0$, and a shared cue across the batch gives $\chi = 1$. We next introduce coordinates that separate expert and routing contrast and put the joint gradient flow in reduced form.

2.2 REDUCED FORMULATION AND JOINT GRADIENT FLOW

To separate shared prediction and load from specialization, set $m = (w_+ + w_-)/2$, $d = (w_+ - w_-)/2$, $a = x - y$, and $b = x + y - 1$. Here (m, b) describes the common expert component and marginal load imbalance, while (d, a) describes expert and routing contrasts. The global result will show $m(t), b(t) \to 0$. Moreover, $p_+(S) = (1 + b + aS)/2$, $|a| + |b| \le 1$, and the marginal load of the $+$ expert is $(1 + b)/2$. In these coordinates, the batch penalty reduces to

$$\lambda\mathbb{E}[(\bar{p} - \tfrac{1}{2})^2 + (1 - \bar{p} - \tfrac{1}{2})^2] = \tfrac{\lambda}{2}(b^2 + \chi a^2), \qquad \bar{p} = \tfrac{1}{2}(1 + b + a\bar{S}), \quad k := \lambda\chi. \tag{4}$$

Thus λ penalizes marginal imbalance, whereas $k = \lambda\chi$ sets the pressure against cue-dependent routing. Batch composition can change that pressure even at fixed population loads. After subtracting the constant $\mathbb{E}[Y^2]/2$, the objective becomes

$$F_q = \tfrac{1}{2}m^2 + bmd + \tfrac{1}{2}[q + (1 - q)(a^2 + b^2)]d^2 - gad + \tfrac{\lambda}{2}b^2 + \tfrac{k}{2}a^2. \tag{5}$$

At $q = 1$, this is the selected-expert objective, with no time or parameter rescaling.

Throughout the binary population analysis, we use Euclidean gradient flow in (m, d, α, β), so $\dot{w}_\pm = -2\partial_{w_\pm} F_q$. Unit Euclidean flow in (w_+, w_-) would rescale the expert dynamics. Write $F_a = \partial_a F_q$, with analogous notation for the other coordinates. With $c_T = 2/T^2$, $F_x = F_a + F_b$, and $F_y = F_b - F_a$, the flow is

$$\dot{m} = -F_m, \quad \dot{d} = -F_d, \quad \dot{x} = -c_T x^2(1-x)^2 F_x, \quad \dot{y} = -c_T y^2(1-y)^2 F_y. \tag{6}$$

The mobility factors $x^2(1-x)^2$ and $y^2(1-y)^2$ reappear in the recovery analysis of Section 4. Appendix A.1 gives the partial derivatives and the selected-expert form of this flow.

2.3 Global phase diagram and the selected-expert endpoint

In (m, d, a, b) coordinates, let $O = (0, 0, 0, 0)$ and $S_\pm = (0, \pm g, \pm 1, 0)$. The latter are fully specialized cue-conditional predictors with equal marginal loads. For $0 < q < 1$ and $qg^2 < k < g^2/q$, let $P_\pm = (0, \pm d_*, \pm a_*, 0)$, where

$$a_*^2 = \frac{g\sqrt{q/k} - q}{1 - q}, \qquad d_* = \sqrt{\frac{k}{q}}\, a_* > 0. \tag{7}$$

Partial specialization means $0 < |a_*| < 1$ and $d_* \neq 0$: routing remains probabilistic while the experts differ. It is not a scalar interpolation between full specialization and homogenization. For $q < 1/2$, increasing k within the partial regime decreases routing contrast but initially increases expert separation; d_*^2 reaches its maximum at $k/g^2 = 1/(4q)$.

The next theorem classifies the possible limits of every finite trajectory and identifies those attained generically. It also rules out expert divergence and asymptotic starvation, which a stationary-point calculation cannot exclude.

Theorem 1 (Global phase diagram for binary two-expert training)**.** *Assume* (2)*, $g, T, \lambda > 0$, $\chi \in [0, 1]$, and $q \in (0, 1]$. Let $k = \lambda\chi$ and consider* (6)*. Every finite initialization generates a global trajectory with bounded expert parameters and $x(t), y(t) \in (0, 1)$ at every finite time. The state (m, d, a, b) converges to a single limit satisfying $m_\infty = b_\infty = 0$. Both limiting marginal loads equal $1/2$; hence asymptotic starvation is excluded.*

For $0 < q < 1$, the complete list of possible limits is

Regime	*Possible limits*
$0 \le k \le qg^2$	O, S_+, S_-
$qg^2 < k < g^2/q$	O, P_+, P_-
$k \ge g^2/q$	O.

When $k < g^2/q$, O has a three-dimensional local stable manifold and a four-dimensional Lebesgue-null basin. Consequently, almost every finite initialization converges to full specialization in the first regime and to a partially specialized equilibrium in the second. The equilibria $P_\pm$ are locally exponentially asymptotically stable. For $k \le qg^2$, including equality, $S_\pm$ are locally asymptotically stable relative to the compactified feasible state space. At $k = g^2/q$, O is nonhyperbolic but locally asymptotically stable and attracts every finite initialization.

For $q = 1$ and $k < g^2$, the possible limits are O, S_+, S_-, and the basin of O is Lebesgue-null. When $k > g^2$, every finite initialization converges to O. At $k = g^2$, each trajectory converges to a single point satisfying

$$m_\infty = b_\infty = 0, \qquad d_\infty = g a_\infty, \qquad |a_\infty| < 1,$$

with $\beta(t) \to 0$ and $\alpha(t) \to 2T \operatorname{arctanh}(a_\infty)$.

The resulting global regimes are summarized in Figure 1(a); the $q = 0$ boundary shown there is analyzed separately in Section 3.2. At criticality, an integrable-drift relation selects one interior limit on the continuum for each trajectory (Appendix A.12). Uniform initial routing with distinct experts avoids the exceptional basin at $q = 1, k < g^2$ and selects nonzero contrast at criticality (Appendix A.14). The case $q = 1, \lambda = 0$ is treated separately from Theorem 1: every finite initialization then converges either to full specialization or to zero experts with nondegenerate cue-independent routing (Appendix A.15).

2.4 WHY THE CLASSIFICATION IS GLOBAL

A local curvature calculation locates a stability boundary of the homogeneous state but does not establish a global phase diagram. The global argument must control expert growth, exclude starvation at the softmax boundary, and show that every trajectory converges to a single admissible limit.

For every $q > 0$, a disk of radius $\max\{\sqrt{m(0)^2 + d(0)^2}, g/q\}$ is forward invariant for the expert subsystem. Bounded expert gradients and reciprocal-probability estimates keep finite-logit trajectories in the interior at every finite time. For $s = x + y$, the population barrier $\mathcal{H}_+ = L(x) + L(y) - (m^2 + d^2)/T^2$, where $L(p) = \log[p/(1-p)] + 1/(1-p)$, satisfies

$$\dot{\mathcal{H}}_+ \geq \frac{2\lambda}{T^2} s[1 - (1+\chi)s] \quad \text{when } s \leq 4q. \tag{8}$$

The quadratic form behind this bound is nonnegative near zero load for every $q > 0$, and $|a| \leq s$ makes the adverse contrast term quadratic. Approaching starvation would force $\mathcal{H}_+ \to -\infty$, contrary to its eventual nonnegative drift. Exchanging experts excludes $s \to 2$.

Dissipation reduces the remaining limits to stationary components. On the balanced-load slice, minimizing over the experts exposes the two thresholds:

$$\Phi_q(a, 0) = \frac{k}{2}a^2 - \frac{g^2 a^2}{2[q + (1-q)a^2]}, \qquad \partial_a \Phi_q(a, 0) = a\left[k - \frac{qg^2}{[q + (1-q)a^2]^2}\right]. \tag{9}$$

The nonzero roots give (7); they meet full specialization at $k = qg^2$ and the homogeneous state at $k = g^2/q$. Boundedness and the barrier confine trajectories; dissipation and boundary signs make the classification global. For $0 < q < 1$, admissible components are isolated; at $q = 1, k = g^2$, integrable drift selects one point on the critical line. Appendix A proves the result. Appendix G shows why routing geometry matters: projected routing can reach permanently starved vertices from an open set of interior initial states.

3 LOCAL STRUCTURE AND THE PURE-AVERAGING BOUNDARY

3.1 A LOCAL CURVATURE–COUPLING CRITERION

The Hessian at the homogeneous binary equilibrium identifies the upper homogeneous-state boundary in Figure 1(a). In a multi-expert model, we ask how expert-separation curvature, router–residual coupling, and balancing curvature jointly determine local stability. Consider $E \geq 2$ separately parameterized real-valued experts $f_{w_j}(U)$ and a softmax router with fixed, bounded, finite-dimensional features $\boldsymbol{\psi}(S)$:

$$p_j(S) = \frac{\exp(\theta_j^\top \boldsymbol{\psi}(S)/T)}{\sum_{\ell=1}^{E} \exp(\theta_\ell^\top \boldsymbol{\psi}(S)/T)}, \qquad \sum_j \theta_j = 0, \quad 0 < T < \infty.$$

Router and expert inputs may be dependent, but the model excludes a shared trainable encoder. Use the prediction-dispersion family in (3) and the balancing penalty $(\lambda E/2)\mathbb{E}_{\text{batch}} \sum_j (\bar{p}_j - 1/E)^2$, with $\lambda \geq 0$. Batch observations have the population marginals but may be dependent within a batch. Assume that f_w is locally C^3, with square-integrable local envelopes for the function and its first three parameter derivatives, and that $\mathbb{E}Y^2 < \infty$.

Let w_0 be a stationary point of the single-expert population risk, and set $\varepsilon = Y - f_{w_0}(U)$ and $J_w = \nabla_w f_{w_0}(U)$. Define

$$\begin{aligned} &\Sigma_J = \mathbb{E}[J_w J_w^\top], \quad \mathsf{N} = \mathbb{E}[\varepsilon \nabla_w^2 f_{w_0}(U)], \quad \mathsf{H} = \Sigma_J - \mathsf{N}, \\ &\mathsf{H}_q = q\Sigma_J - \mathsf{N}, \quad \mathsf{G} = \mathbb{E}[\boldsymbol{\psi}(S)\varepsilon J_w^\top], \quad \mathsf{C}_\psi = \mathbb{E}_{\text{batch}}[\bar{\boldsymbol{\psi}}\bar{\boldsymbol{\psi}}^\top]. \end{aligned} \tag{10}$$

Here $\bar{\boldsymbol{\psi}}$ is the batch-mean feature vector. Thus C_ψ is an uncentered second moment and need not be a covariance matrix.

Theorem 2 (Local curvature–coupling criterion)**.** *Under these assumptions, suppose* $\mathsf{H} \succ 0$*. Then* $w_j = w_0, \theta_j = 0$ *is stationary. After rescaling the router coordinates by temperature, the Hessian is congruent to one common-expert block* H *and* $E - 1$ *contrast blocks* $E^{-1}\mathcal{K}_q$*, where*

$$\mathcal{K}_q = \begin{pmatrix} \lambda \mathsf{C}_\psi & -\mathsf{G} \\ -\mathsf{G}^\top & \mathsf{H}_q \end{pmatrix}. \tag{11}$$

For gradient flow with a smooth positive-definite metric on the constrained parameter space, the equilibrium is locally exponentially asymptotically stable if and only if $\mathcal{K}_q \succ 0$. If $\mathcal{K}_q$ has a negative eigenvalue, the equilibrium is unstable. Moreover,

$$\mathcal{K}_q \succeq 0 \iff \mathsf{H}_q \succeq 0, \quad \mathsf{G}\ker \mathsf{H}_q = \{0\}, \quad \lambda \mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^{\dagger}\mathsf{G}^\top \succeq 0. \tag{12}$$

In particular, if some $v \in \ker \mathsf{H}_q$ has $\mathsf{G}v \neq 0$, a negative-curvature direction remains at every finite balancing strength. A singular positive-semidefinite $\mathcal{K}_q$ alone does not determine nonlinear stability.

When $\mathsf{H}_q \succ 0$, the strict criterion has the ordinary Schur form

$$\mathcal{K}_q \succ 0 \quad \iff \quad \mathsf{H}_q \succ 0 \text{ and } \lambda \mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^{-1}\mathsf{G}^\top \succ 0. \tag{13}$$

A negative expert-contrast curvature cannot be stabilized by balancing. Coupling through G likewise makes a flat expert direction unstable at every finite balancing strength, as the kernel condition in (12) shows. When $\mathsf{H}_q \succ 0$, balancing curvature must dominate the effective coupling $\mathsf{G}\mathsf{H}_q^{-1}\mathsf{G}^\top$.

The matrix G correlates router features with residual-weighted expert sensitivity. This coupling arises from the router-only cue in the binary model and can also reflect structured residuals under expert misspecification. If $\boldsymbol{\psi}(S)$ is measurable from U and $f_{w_0}(U) = \mathbb{E}[Y \mid U]$, then $\mathbb{E}[\varepsilon \mid U] = 0$ and $\mathsf{G} = \mathsf{N} = 0$: the residual coupling vanishes locally in this matched-information case.

For the binary linear model, $w_0 = 0$, $\boldsymbol{\psi}(S) = (1, S)^\top$, $\mathsf{C}_\psi = \mathrm{diag}(1, \chi)$, $\mathsf{G} = (0, g)^\top$, and $\mathsf{H}_q = q$. The nontrivial contrast block is

$$\begin{pmatrix} k & -g \\ -g & q \end{pmatrix}, \qquad qk = g^2 \text{ at the homogeneous curvature boundary.} \tag{14}$$

This yields $k = g^2$ at $q = 1$ and the upper boundary $k = g^2/q$ for $q < 1$. The lower boundary $k = qg^2$ comes from the specialized states rather than the homogeneous Hessian. Theorem 2 describes local behavior at the homogeneous equilibrium; the binary analysis provides the global classification and exclusion of boundary starvation. Appendix B gives the expansion, Schur conditions, exact correspondence with the binary flow metric, and examples with singular curvature.

3.2 Pure averaging: non-attainment and invariant-set escape

At $q = 0$, the linear contrast curvature in (14) vanishes while router coupling remains nonzero, so the positive-q expert-boundedness argument no longer applies. This raises two questions: can finite parameters attain the infimum, and can the flow escape toward it? The proposition answers both, with its dynamical claim restricted to a specified invariant set.

Proposition 3 (Pure averaging: non-attainment and invariant-set escape)**.** *In the binary linear model, let $q = 0$, $g, T, \lambda > 0$, and $k = \lambda\chi > 0$, and let F_0 denote the objective in* (5) *at $q = 0$. This objective has infimum $-g^2/2$, which no finite expert parameters attain. The homogeneous equilibrium is a strict saddle for every finite balancing strength.*

The plane $m = b = 0$ is invariant. Every finite initial point on this plane satisfying $0 < a_0 < 1$, $d_0 > 0$, and $0 < a_0 d_0 \leq g$ generates a trajectory defined for all $t \geq 0$ that satisfies

$$d(t) \to +\infty, \qquad a(t) \to 0, \qquad a(t)d(t) \to g. \tag{15}$$

An exact square decomposition gives non-attainment:

$$F_0 + \frac{g^2}{2} = \frac{1}{2}(m + bd)^2 + \frac{1}{2}(ad - g)^2 + \frac{\lambda}{2}b^2 + \frac{k}{2}a^2. \tag{16}$$

For $k > 0$, equality would require both $a = 0$ and $ad = g$. The infimum is instead approached as $a \downarrow 0$, $d = g/a$. Appendix C proves dynamical escape on the stated invariant plane.

Along these trajectories, the marginal expert loads remain $1/2$ and both conditional routing probabilities approach $1/2$, while the expert parameters diverge with opposite signs. Nevertheless, $\bar{f} = (m + bd + adS)U$ converges in mean square to gSU. Finite-parameter attainment and expert boundedness fail, but cue-dependent prediction persists. Nonlinear experts can have positive contrast curvature through $-\mathsf{N}$ even at $q = 0$; thus pure averaging need not be locally unstable in a nonlinear model. Appendix C proves the proposition without using the positive-q bounds.

4 LOAD RECOVERY UNDER SELECTED-EXPERT TRAINING

4.1 POPULATION RECOVERY FROM SATURATED ROUTING

Under joint selected-expert training ($q = 1$), initialize $\alpha(0) = 0, \beta(0) = -M$; then the $+$ expert has $x(0) = y(0) = \delta_0 = \mathrm{sigmoid}(-M/T)$. For fixed $\delta_0 < r < 1/4$, define $\tau_r = \inf\{t \geq 0 : (x(t) + y(t))/2 = r\}$. Near zero load, softmax mobility is quadratic in the rare routing probability. Since δ_0 is of order $e^{-M/T}$, this suggests a reciprocal, exponentially long recovery scale. The next theorem establishes that scale for the joint flow.

Theorem 4 (Population recovery time)**.** *Under the teacher model, fix $\lambda, T > 0$ and $0 < r < 1/4$. With r, T, λ, g, χ and the initial expert weights fixed, there are constants $0 < c_1 \leq c_2 < \infty$, independent of M, such that*

$$c_1 e^{M/T} \leq \tau_r \leq c_2 e^{M/T} \tag{17}$$

for all sufficiently large M. Thus $\tau_r = \Theta(e^{M/T})$ as $M \to \infty$.

The lower bound controls reciprocal rare probability; the upper bound uses the starvation barrier (Appendix D). Near $q = 1, k = g^2$, a vanishing unstable eigenvalue can delay visible specialization (Appendix A.15). The scale varies M at fixed parameters; changing T also changes mobility through T^{-2}.

Figure 2 connects batch composition to the routing-contrast penalty through χ and near-zero softmax mobility to the $e^{M/T}$ recovery scale.

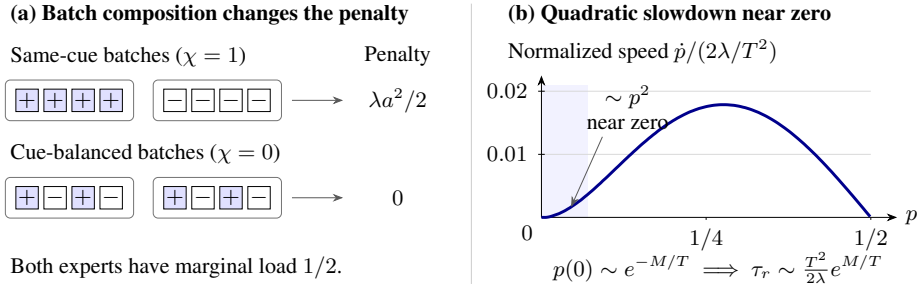


Figure 2: **Two mechanisms in the selected-expert model. (a)** Batch composition changes χ and the routing-contrast penalty at balanced marginal loads: same-cue batches give $\chi = 1$, whereas exactly cue-balanced batches give $\chi = 0$. **(b)** Softmax probability mobility is quadratic near zero load. On the zero-expert invariant line, $p(0) \asymp e^{-M/T}$ gives $\tau_r \sim (T^2/2\lambda)e^{M/T}$; Theorem 4 proves the same exponential order while expert weights evolve.

4.2 SIMULTANEOUS CATEGORICAL UPDATES AND RECOVERY BOUNDS

Categorical updates require more than a population recovery bound. A rare assignment can produce an order-one no-baseline score increment while its mean scales with the rare load; it also updates an expert. We therefore bound each path through the one-step logit change and control the upper recovery scale through a corrected potential.

For independent batch cues, $\chi = 1/B$. We use the same teacher and simultaneous categorical router–expert updates, with batch size and step size fixed as M grows. Assume

$$|U| \leq K, \qquad |Y| \leq D|U| \quad \text{almost surely}, \qquad K \geq 1. \tag{18}$$

Each iteration draws B independent teacher examples and conditionally independent $I_i \sim \mathrm{Bernoulli}(p_i)$, where $p_i = p_+(S_i)$; these draws are independent of the past. Define the dispatch load, mean routing probability, and selected-expert loss by

$$\widehat{p}_B := B^{-1} \sum_i I_i, \qquad \bar{p}_B := B^{-1} \sum_i p_i, \qquad \ell_i = \tfrac{1}{2}(w_{I_i} U_i - Y_i)^2,$$

where $w_1 = w_+$ and $w_0 = w_-$. With $\boldsymbol{\psi}_i := \boldsymbol{\psi}(S_i) = (1, S_i)^\top$, $\theta = (\beta, \alpha)^\top$, and step size $\eta > 0$, the simultaneous updates are

$$\theta' = \theta - \eta \left\{ \frac{1}{BT} \sum_i \ell_i (I_i - p_i) \boldsymbol{\psi}_i + \frac{2\lambda(2\widehat{p}_B - 1)}{BT} \sum_i p_i (1 - p_i) \boldsymbol{\psi}_i \right\}, \tag{19}$$

$$w'_+ = w_+ - \frac{2\eta}{B} \sum_i I_i (w_+ U_i - Y_i) U_i, \qquad w'_- = w_- - \frac{2\eta}{B} \sum_i (1 - I_i)(w_- U_i - Y_i) U_i. \tag{20}$$

The same assignments drive both updates. The first router term is the no-baseline score estimator, $\nabla_\theta \log \mathbb{P}(I_i \mid S_i) = (I_i - p_i)\boldsymbol{\psi}_i / T$. Balancing holds $\widehat{p}_B$ fixed and differentiates through $\bar{p}_B$. Since $\mathbb{E}[\widehat{p}_B \mid \text{batch}] = \bar{p}_B$, the mean update follows (5) at $q = 1, \chi = 1/B$. Hard top-one dispatch has a different assignment mean and prediction gradient.

For the bounds, set $R \geq \max\{D, |w_+(0)|, |w_-(0)|\}$ and $C_{\mathrm{upd}} = R^2 K^2 + \lambda$; then $g \leq D \leq R$. The step-size condition below bounds expert iterates by R, and C_{upd} sets the one-step scale. Initialize $\alpha_0 = 0, \beta_0 = -M$, set $\delta_0 = \mathrm{sigmoid}(-M/T)$, and define $\tau_r = \inf\{n \geq 0 : (x_n + y_n)/2 \geq r\}$. Appendix E gives the score variance.

Proposition 5 (Pathwise recovery lower bound)**.** *Assume* (18) *and the updates* (19)–(20)*. If* $\eta \leq 1/(2K^2)$ *and* $\delta_0 < r$*, every realization satisfies*

$$\tau_r \geq \frac{T^2}{4\eta C_{\mathrm{upd}}} \left(\delta_0^{-1} - \frac{1}{r} \right). \tag{21}$$

In particular, $\delta_0 \leq r/2$ *gives* $\tau_r \geq T^2 e^{M/T} / (8\eta C_{\mathrm{upd}})$.

The step-size condition preserves $|w_\pm(n)| \leq R$ (Appendix E.2). A baseline can preserve the mean gradient while changing pathwise behavior, so the bound depends on the estimator.

The pathwise bound excludes substantially faster recovery. The next theorem controls long recovery times in the second moment despite active-expert fluctuations.

Theorem 6 (Second moment of the recovery time)**.** *Assume* (18)*, the updates* (19)–(20)*, and* $0 < \lambda < 4gR$*. Fix any* $B \geq 1$*. There exists* $r_0 > 0$ *such that, for every fixed* $r \in (0, r_0]$*, there exists* $\eta_0(r) > 0$*, both independent of* M *and the confidence level, with the following property. Whenever* $0 < \eta \leq \eta_0(r)$ *and* $\mathrm{sigmoid}(-M/T) < r$*,*

$$\mathbb{E}\left[\left(\frac{\eta\lambda}{T^2} e^{-M/T} \tau_r \right)^2 \right] \leq C \exp\left(\frac{w_+(0)^2}{2T^2} \right), \tag{22}$$

where C *is a universal constant. Consequently, for every* $0 < \epsilon < 1$*,*

$$\mathbb{P}\left(\tau_r \leq \left\lceil C \epsilon^{-1/2} e^{w_+(0)^2/(4T^2)} \frac{T^2}{\eta\lambda} e^{M/T} \right\rceil \right) \geq 1 - \epsilon. \tag{23}$$

Explicit admissible choices are given in Appendix E.

Together with (21), the theorem determines the dependence on M for a fixed admissible step size and target load, including parameter values for which the population flow specializes. Appendix F treats $\lambda \geq 4gR$ under separate conditions. The results concern a sufficiently small fixed positive load; experiments with larger thresholds provide qualitative checks beyond the range of the theorem.

4.3 A CORRECTED POTENTIAL FOR SHARED ROUTER–EXPERT NOISE

The population barrier fails sample by sample: the squared active-expert weight adds $O(\eta^2)$ corner drift from persistent sampling noise (Appendix E.6). A rare-expert-only barrier avoids this drift but leaves a first-order aw_{a} interaction. Let $w_{\mathrm{r}} = w_+$ be the rare expert, $w_{\mathrm{a}} = w_-$ the active expert, and $L(p) = \log[p/(1-p)] + 1/(1-p)$. The cross term below cancels that interaction in conditional expectation through active-expert and product increments:

$$\mathcal{J} = \mathcal{H}_0 - \frac{g a w_{\mathrm{a}}}{2T^2}, \qquad \mathcal{H}_0 = L(x) + L(y) - \frac{w_{\mathrm{r}}^2}{2T^2}. \tag{24}$$

The correction leaves only $O(s^2)$ interaction and noise terms. At a small recovery threshold and admissible step size, linear balancing dominates them, giving conditional drift of $\mathcal{J}$ at least μs, where $\mu = \eta\lambda/T^2$ (Appendix E.3–E.4). An exponential transform gives a stopped-process estimate; restarting it yields (22) and the recovery-time tail.

5 EMPIRICAL EVIDENCE WITH LEARNED REPRESENTATIONS

We test whether balancing-group composition changes routing association in a four-layer Transformer-MoE with two four-expert hard top-one layers. Its router uses learned hidden states and receives no domain label. Within each of five seeds, the compared conditions share initialization and supervised sequence sampling. At $\lambda_{\text{aux}} = 0.05$, domain-balanced and single-domain auxiliary groups each contain eight sequences; only their domain composition changes.

Mean expert-load coefficient of variation (CV) is 0.0515 ± 0.0168 versus 0.0500 ± 0.0112, while domain–expert mutual information (MI) is 0.0326 ± 0.0212 versus 0.00262 ± 0.00069 nats (Figure 3). The MI difference favors domain-balanced groups in all five pairs, while load-CV differences have mixed signs. These results provide qualitative evidence that routing association can vary despite comparable aggregate loads, while the reported mean squared difference between expert outputs remains similar across conditions. Appendix I.4 gives the full sweep, validation perplexity, and routing matrices; Appendix I.3 retains the MNIST comparison.

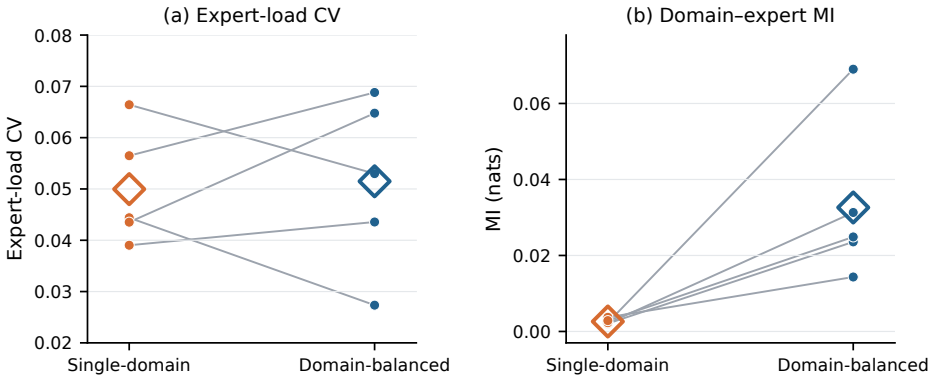


Figure 3: Paired effect of auxiliary-group composition in the Transformer-MoE. Five paired runs at $\lambda_{\text{aux}} = 0.05$ share initialization and supervised sequence sampling; both auxiliary groupings contain eight sequences and differ only in domain composition. (a) Expert-load CV has mixed paired changes and similar means. (b) Domain–expert MI is higher under domain-balanced grouping in all five pairs. Diamonds denote condition means.

6 CONCLUSION

Expert specialization in a jointly trained MoE is multidimensional. Balanced utilization, input-dependent routing, differences between expert predictors, and predictive signal need not move together. Stronger balancing can reduce routing contrast while initially increasing expert separation; nearly uniform routing can coexist with cue-dependent prediction. Claims of specialization should therefore state what is measured.

The prediction objective and balancing scope shape limiting states, while softmax routing governs how quickly an underused expert recovers. Asymptotic non-starvation does not ensure recovery within a finite training horizon, and balanced loads do not imply identical predictors. Batch composition can also change routing pressure without changing population utilization.

Our global classification concerns a binary linear model; the multi-expert criterion is local with fixed routing features, and neural experiments offer qualitative evidence. The broader challenge is to identify which mechanisms govern routing structure, expert differentiation, and recovery time under shared representation learning. A theory that answers this question would better predict the behavior of jointly trained MoEs.

AI USE STATEMENT

Generative AI assisted with mathematical derivations and proof checks, literature review, symbolic and finite-state checks, experiment implementation and analysis, manuscript preparation, and LaTeX development.

REPRODUCIBILITY STATEMENT

Section 2 defines the prediction-dispersion family, moment assumptions, and gradient normalization. Section 4.2 specifies the selected-expert categorical algorithm and its assumptions. Appendices A–G contain the proofs of the population, local-structure, pure-averaging, recovery, and routing-geometry results. The experimental appendices describe data generation, training settings, and evaluation metrics. Code for reproducing the reported experiments and figures is included in the supplementary material.

REFERENCES

Sebastian Allmeier and Nicolas Gast. Computing the bias of constant-step stochastic approximation with Markovian noise. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137873–137902, 2024. doi: 10.52202/079017-4379. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f949c1f490beb42124a267b7476cd353-Abstract-Conference.html.

Lizhang Chen, Jonathan Li, Qi Wang, Runlong Liao, Shuozhe Li, Chen Liang, Ni Lao, and Qiang Liu. ϕ-balancing for mixture-of-experts training. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2605.15403.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23049–23062, 2022. doi: 10.52202/068431-1675. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Shivam Garg, Samuele Tosatto, Yangchen Pan, Martha White, and A. Rupam Mahmood. An alternate policy gradient estimator for softmax policies. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6630–6689, 2022. URL https://proceedings.mlr.press/v151/garg22b.html.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better MoE. In *Advances in Neural Information Processing Systems*, volume 38, pp. 54383–54425, 2025. doi: 10.52202/085713-1629. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4598de7d243d528e38eb0c5d8155fb52-Abstract-Conference.html.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79. URL https://direct.mit.edu/neco/article/3/1/79/5560/Adaptive-Mixtures-of-Local-Experts.

Kukjin Kang and Jong-Hoon Oh. Statistical mechanics of the mixture of experts. In *Advances in Neural Information Processing Systems*, volume 9, pp. 183–189, 1996. URL https://proceedings.neurips.cc/paper/1996/file/a7d8ae4569120b5bec12e7b6e9648b86-Paper.pdf.

Ryotaro Kawata, Kohsei Matsutani, Yuri Kinoshita, Naoki Nishikawa, and Taiji Suzuki. Mixture of experts provably detect and learn the latent cluster structure in gradient-based learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29390–29448, 2025. URL https://proceedings.mlr.press/v267/kawata25a.html.

O. M. Kiselev. A minimal bifurcation model of load imbalance in a softmax mixture-of-experts router. *arXiv preprint arXiv:2605.29121*, 2026. URL https://arxiv.org/abs/2605.29121.

Sara Klein, Simon Weissmann, and Leif Döring. Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1VeQ6VBbev.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, 1994. URL https://papers.nips.cc/paper_files/paper/1994/hash/b8c37e33defde51cf91e1e03e51657da-Abstract.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://bottou.org/papers/lecun-98h.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, 201:707–802, 2023. doi: 10.1007/s10107-022-01920-6. URL https://doi.org/10.1007/s10107-022-01920-6.

Hongbo Li, Sen Lin, Lingjie Duan, Yingbin Liang, and Ness B. Shroff. Theory on mixture-of-experts in continual learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/17a234c91f746d9625a75cf8a8731ee2-Paper-Conference.pdf.

Jiaang Li, Haibin Chen, Langming Liu, Yujin Yuan, Yadao Wang, Yizhen Zhang, Chengting Yu, Xin Tong, Weidong Zhang, Shilei Liu, Wenbo Su, and Bo Zheng. Expert divergence learning for MoE-based language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0c8f0ede04b01a976bedd97cfc2df37d-Abstract-Conference.html.

Fangshuo Liao and Anastasios Kyrillidis. Guided by the experts: Provable feature learning dynamic of soft-routed mixture-of-experts. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3898–3906, 2026. URL https://proceedings.mlr.press/v300/liao26a.html.

Jincheng Mei and Ian Osband. Delightful gradients accelerate corner escape. *arXiv preprint arXiv:2605.11908*, 2026. URL https://arxiv.org/abs/2605.11908.

Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5005–5018, 2025. doi: 10.18653/v1/2025.acl-long.249. URL https://aclanthology.org/2025.acl-long.249/.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Xi Wang, Soufiane Hayou, and Eric Nalisnick. The myth of expert specialization in MoEs: Why routing reflects geometry, not necessarily domain expertise. *arXiv preprint arXiv:2604.09780*, 2026. URL https://arxiv.org/abs/2604.09780.

A GLOBAL BINARY CLASSIFICATION AND SELECTED-EXPERT REFINEMENTS

We prove Theorem 1 under (2). The subsequent selected-expert refinements use the teacher model and the same gradient normalization. Set $c_T = 2/T^2$. At $q = 1, k = g^2$, define

$$\begin{aligned} \Xi(p) &= -p^{-1} + (1-p)^{-1} + 2\log\frac{p}{1-p}, \\ \Xi_-(x,y) &:= \frac{\Xi(x) - \Xi(y)}{2c_T}, \qquad \Xi_+(x,y) := \frac{\Xi(x) + \Xi(y)}{2c_T}. \end{aligned} \tag{A.14}$$

The single-point critical selection relation is

$$2T^2\left[\frac{a_\infty}{1-a_\infty^2} + \operatorname{arctanh}(a_\infty)\right] + g^2 a_\infty = \Xi_-(x(0),y(0)) + gd(0) - g\int_0^\infty b(t)m(t)\,dt. \tag{A.15}$$

The integral is absolutely convergent, as shown below. Along a trajectory, $\Xi_\pm$ abbreviates $\Xi_\pm(x(t), y(t))$.

Proof of Theorem 1. We first derive the objective and flow identities, then classify the compactified stationary components and exclude inadmissible limits. The critical continuum at $q = 1$ requires a separate selection argument.

A.1 OBJECTIVE DECOMPOSITION

Completing squares in (5) gives

$$F_q + \frac{g^2}{2} = \frac{1}{2}(m+bd)^2 + \frac{1}{2}(ad-g)^2 + \frac{q}{2}(1-a^2-b^2)d^2 + \frac{\lambda}{2}b^2 + \frac{k}{2}a^2. \tag{A.16}$$

On the feasible diamond, $a^2+b^2 \le (|a|+|b|)^2 \le 1$. Hence $F_q \ge -g^2/2$ for $q \ge 0$ and nonnegative balancing coefficients. We use both the expanded form (5) and the decomposition (A.16) below.

A.2 GRADIENT FLOW AND NORMALIZATION

Let F_m, F_d, F_a, F_b denote the partial derivatives of F_q. Differentiation gives

$$\begin{aligned} F_m &= m + bd, \\ F_d &= bm + \big[q + (1-q)(a^2+b^2)\big]d - ga, \\ F_a &= \big[k + (1-q)d^2\big]a - gd, \\ F_b &= md + (1-q)bd^2 + \lambda b. \end{aligned} \tag{A.17}$$

Since $F_x = F_a + F_b$ and $F_y = F_b - F_a$, writing $r_x = x(1-x)$ and $r_y = y(1-y)$ gives

$$F_\alpha = \frac{r_x F_x - r_y F_y}{T}, \qquad F_\beta = \frac{r_x F_x + r_y F_y}{T}. \tag{A.18}$$

Consequently, Euclidean flow in the stated coordinates induces

$$\begin{aligned} \dot m &= -F_m, & \dot d &= -F_d, \\ \dot x &= -c_T x^2(1-x)^2 F_x, & \dot y &= -c_T y^2(1-y)^2 F_y, \qquad c_T = \frac{2}{T^2}. \end{aligned} \tag{A.19}$$

The factor 2 and the two powers of T^{-1} follow from (A.18) and the conversion from logit to probability velocity.

The expert-coordinate relation is

$$F_{w_+} = \frac{F_m + F_d}{2}, \qquad F_{w_-} = \frac{F_m - F_d}{2}, \qquad \dot w_\pm = -2F_{w_\pm}. \tag{A.20}$$

At $q = 1$, let $s = x+y = 1+b$, $h_x = F_x = md-gd+\lambda b+ka$, and $h_y = F_y = md+gd+\lambda b-ka$. The selected-expert flow is

$$\dot{w}_+ = -sw_+ + ga, \quad \dot{w}_- = -(2-s)w_- - ga, \quad \dot{x} = -c_T x^2(1-x)^2 h_x, \quad \dot{y} = -c_T y^2(1-y)^2 h_y. \tag{A.21}$$

Equations (A.19)–(A.21) recover the selected-expert flow exactly. Euclidean flow in (w_+, w_-) with unit coefficients would use a different normalization.

A.3 Expert boundedness on the maximal existence interval

For $z = (m, d)^\top$, the expert subsystem is

$$\begin{aligned} \dot{z} &= -A_q(a,b)z + \begin{pmatrix} 0 \\ ga \end{pmatrix}, \\ A_q(a,b) &= \begin{pmatrix} 1 & b \\ b & q + (1-q)(a^2+b^2) \end{pmatrix} \\ &= q \begin{pmatrix} 1 & b \\ b & 1 \end{pmatrix} + (1-q) \begin{pmatrix} 1 & b \\ b & a^2+b^2 \end{pmatrix}. \end{aligned} \tag{A.22}$$

The second matrix is positive semidefinite because its quadratic form is $(m+bd)^2 + a^2d^2$. The first has minimum eigenvalue $1 - |b|$. For $0 < q \le 1$,

$$z^\top A_q z \ge q(1-|b|)\|z\|^2.$$

The estimate degenerates at $|b| = 1$ and therefore does not give uniform positive definiteness on the closed router domain. Feasibility also gives $|a| \le 1 - |b|$, so

$$\frac{1}{2}\frac{d}{dt}\|z\|^2 = -z^\top A_q z + gad \le -(1-|b|)\big(q\|z\|^2 - g\|z\|\big). \tag{A.23}$$

The radial derivative is nonpositive whenever $\|z\| \ge g/q$. Hence

$$m(t)^2 + d(t)^2 \le \max\left\{ m(0)^2 + d(0)^2, \frac{g^2}{q^2} \right\}. \tag{A.24}$$

The argument remains valid at the degenerate boundary because it does not divide by $1 - |b|$. The bound holds throughout the maximal finite-logit solution. We next use it to prove global existence. The bound need not remain uniform as $q \downarrow 0$.

A.4 Finite-time probability-boundary exclusion and global existence

The vector field in (m, d, x, y) is polynomial, and the probability-boundary faces are invariant under its extension. On the product of the expert disk (A.24) and $[0,1]^2$, both F_x and F_y are bounded; let C_F be a common absolute bound.

For $p = x$ or y, with its corresponding derivative F_p, (A.19) yields

$$\frac{d}{dt}\frac{1}{p} = c_T(1-p)^2 F_p, \qquad \frac{d}{dt}\frac{1}{1-p} = -c_T p^2 F_p. \tag{A.25}$$

Therefore, as long as the finite-logit solution exists,

$$p(t) \ge \frac{1}{p(0)^{-1} + c_T C_F t}, \qquad 1 - p(t) \ge \frac{1}{(1-p(0))^{-1} + c_T C_F t}. \tag{A.26}$$

Neither endpoint can be reached in finite time. If the maximal existence time were finite, the expert bound and (A.26) would confine the trajectory to a compact subset of the finite-parameter domain, where the solution could be continued. Equivalently, one can continue the bounded polynomial flow and invert

$$\alpha = \frac{T}{2}\big(\operatorname{logit} x - \operatorname{logit} y\big), \qquad \beta = \frac{T}{2}\big(\operatorname{logit} x + \operatorname{logit} y\big).$$

Thus the trajectory exists globally, with finite logits at every finite time. These estimates alone do not exclude asymptotic starvation or divergent logits as time tends to infinity.

A.5 Starvation barriers

Define

$$L(p) = \log \frac{p}{1-p} + \frac{1}{1-p}, \qquad \mathcal{H}_+ = L(x) + L(y) - \frac{m^2+d^2}{T^2}. \tag{A.27}$$

Since $L'(p) = 1/[p(1-p)^2]$, direct differentiation gives

$$\frac{\dot{\mathcal{H}}_+}{c_T} = mF_m + dF_d - aF_a - sF_b = Q_q(m,d,s) + \lambda s(1-s) - ka^2, \tag{A.28}$$

where

$$Q_q(m,d,s) = m^2 + (s-2)md + [1-(1-q)s]d^2. \tag{A.29}$$

The two terms proportional to gad cancel. The coefficient of md is $2b - s = s - 2$, and the coefficient of d^2 simplifies to $q + (1-q)b^2 - (1-q)sb = 1 - (1-q)s$.

Write $w_r = m + d$ and $w_a = m - d$. Then

$$4Q_q(m,d,s) = qs\, w_r^2 + 2(1-q)s\, w_r w_a + [4-(2-q)s]w_a^2 = \begin{pmatrix} w_r & w_a \end{pmatrix} \mathbf{Q}_q(s) \begin{pmatrix} w_r \\ w_a \end{pmatrix},$$

$$\mathbf{Q}_q(s) = \begin{pmatrix} qs & (1-q)s \\ (1-q)s & 4-(2-q)s \end{pmatrix}. \tag{A.30}$$

Its determinant is

$$\det \mathbf{Q}_q(s) = s(4q - s). \tag{A.31}$$

For $s > 0$, the leading diagonal entry is positive, so $\mathbf{Q}_q(s) \succeq 0$ whenever $s \le 4q$. At $s = 0$, the matrix equals $\operatorname{diag}(0, 4)$. We use this positivity only near zero load; for $q < 1/2$, it need not hold throughout the router domain.

Because $|a| \le s$ and $k = \lambda\chi$, for $0 \le s \le 4q$,

$$\dot{\mathcal{H}}_+ \ge c_T \lambda s[1 - (1+\chi)s]. \tag{A.32}$$

Choose $0 < \delta < \min\{4q, 1/(1+\chi)\}$. If $s(t) \to 0$, then $s(t) \le \delta$ for all t beyond some finite t_0. Hence $\mathcal{H}_+(t) \ge \mathcal{H}_+(t_0) > -\infty$ for $t \ge t_0$. Yet $x, y \to 0$ would give $L(x) + L(y) \to -\infty$ and, because the experts are bounded, $\mathcal{H}_+ \to -\infty$, a contradiction.

The expert-swapped barrier is

$$\mathcal{H}_- = L(1-x) + L(1-y) - \frac{m^2+d^2}{T^2}. \tag{A.33}$$

Exchanging the experts sends (m, d, a, b) to $(m, -d, -a, -b)$ and s to $2 - s$. The same barrier argument therefore excludes $s(t) \to 2$.

This rules out $s(t) \to 0$ and $s(t) \to 2$; Appendix A.8 completes the exclusion of starvation points from the omega-limit set.

At $q = 1$, (A.30) reduces to $4Q_1 = sw_+^2 + (4 - s)w_-^2$, and (A.28) becomes the selected-expert barrier drift with its $1/(2T^2)$ expert-square coefficients.

A.6 Interior stationary points and expert elimination

For $|b| < 1$, set

$$D_q(a, b) = q(1 - b^2) + (1 - q)a^2 > 0.$$

The stationary expert equations give

$$m = -bd, \qquad d = \frac{ga}{D_q(a, b)}. \tag{A.34}$$

The corresponding reduced objective is

$$\Phi_q(a, b) = \frac{\lambda}{2}b^2 + \frac{k}{2}a^2 - \frac{g^2a^2}{2D_q(a, b)}. \tag{A.35}$$

The following exact decomposition also holds:

$$F_q = \frac{1}{2}(m + bd)^2 + \frac{D_q(a, b)}{2}\left(d - \frac{ga}{D_q(a, b)}\right)^2 + \Phi_q(a, b). \tag{A.36}$$

At an interior stationary point, both probability mobilities are positive. The remaining equations are

$$b(\lambda - qd^2) = 0, \qquad a[k + (1 - q)d^2] - gd = 0. \tag{A.37}$$

Suppose $b \neq 0$. Then $qd^2 = \lambda > 0$. Multiplying the second equation in (A.37) by a and the equation $D_q d = ga$ by d gives

$$qd^2(1 - b^2) = ka^2, \qquad \text{hence} \qquad 1 - b^2 = \chi a^2. \tag{A.38}$$

This contradicts

$$\chi a^2 \leq a^2 < (1 - |b|)^2 < 1 - b^2.$$

This argument does not divide by χ, a, or d and therefore includes $\chi = 0$. Every interior stationary point satisfies $b = m = 0$.

On the balanced-load slice, $D_q(a, 0) = q + (1 - q)a^2$. Hence

$$\Phi_q(a, 0) = \frac{k}{2}a^2 - \frac{g^2a^2}{2D_q(a, 0)}, \qquad \partial_a\Phi_q(a, 0) = a\left[k - \frac{qg^2}{D_q(a, 0)^2}\right]. \tag{A.39}$$

For $0 < q < 1$, a nonzero interior root requires $q < D_q(a, 0) < 1$. Thus $D_q(a, 0) = g\sqrt{q/k}$ has a solution precisely when $qg^2 < k < g^2/q$, giving (7). There are exactly two roots, with matching signs of a and d.

At $k = qg^2$, the nonzero root reaches $a^2 = 1$ and leaves the interior. At $k = g^2/q$, it merges with $a = 0$. At $k = 0$, no nonzero interior root exists. The classification includes $\chi = 0$ and $\chi = 1$.

For $q = 1$, (A.39) becomes $a(k - g^2) = 0$. Thus the only interior stationary point is O unless $k = g^2$, in which case the interior stationary line is $m = b = 0$, $d = ga$, $|a| < 1$.

A.7 ALL BOUNDARY STATIONARY POINTS

Boundary stationarity is defined for the polynomial extension of (A.19). When a boundary mobility vanishes, the corresponding probability derivative of F_q need not vanish.

Representative open edge and its root count Use $u \in (0, 1)$ as an edge parameter in this subsection, and take $x = 1, y = u$. Then $a = 1 - u, b = u$. Solving the expert equations gives

$$d(u) = \frac{g}{1 + (2q - 1)u}, \qquad m = -ud(u). \tag{A.40}$$

The denominator is positive for $0 < q \leq 1$ and $0 \leq u \leq 1$. With $\phi_q(u) = \Phi_q(1 - u, u)$,

$$\phi_q'(u) = \lambda[(1 + \chi)u - \chi] + \frac{qg^2}{[1 + (2q - 1)u]^2}. \tag{A.41}$$

Because y is the unsaturated coordinate, the edge is stationary exactly at a zero of (A.41).

If $\chi = 0$, the derivative is positive. If $\chi > 0$ and $k \leq qg^2$, it is also positive when $(1+\chi)u-\chi \geq 0$. In the remaining case, $\lambda = k/\chi \leq qg^2/\chi$, $\chi \leq 1$, and $2q - 1 \leq 1$ give

$$\phi_q'(u) \geq qg^2 \left[2u - 1 + \frac{1}{(1 + u)^2}\right] = qg^2 \frac{u^2(3 + 2u)}{(1 + u)^2} > 0.$$

Thus no open-edge stationary point exists when $k \leq qg^2$.

If $k > qg^2$, then

$$\phi_q'(0) = qg^2 - k < 0, \qquad \phi_q'(1) = \lambda + \frac{g^2}{4q} > 0.$$

Moreover,

$$\phi_q''(u) = \lambda(1 + \chi) - \frac{2(2q - 1)qg^2}{[1 + (2q - 1)u]^3} > 0.$$

For $2q - 1 \leq 0$, both terms are nonnegative and the first is positive. For $2q - 1 > 0$, use $\lambda(1 + \chi) = k(1 + \chi)/\chi > 2qg^2$ and $[1 + (2q - 1)u]^3 \geq 1$. Hence each open edge has exactly one root, denoted u_*.

At the representative root, $F_y = 0$, so $F_a = F_b$. Equation (A.41) gives

$$qd(u_*)^2 = \lambda[\chi - (1 + \chi)u_*].$$

Therefore

$$F_x = 2F_b = 2\lambda u_*[1 - \chi + (1 + \chi)u_*] =: \sigma_{\text{edge}} > 0. \tag{A.42}$$

The expression remains strictly positive when $\chi = 1$.

The four open edges The objective and flow respect the expert and sign symmetries

$$(m, d, a, b) \mapsto (m, -d, -a, -b), \qquad (m, d, a, b) \mapsto (-m, d, a, -b).$$

By symmetry, the representative calculation gives the complete list below. These points exist only when $k > qg^2$.

(x, y)	(a, b)	(m, d)	(F_x, F_y)
$(1, u_*)$	$(1 - u_*, u_*)$	$(-u_* d(u_*), d(u_*))$	$(\sigma_{\text{edge}}, 0)$
$(u_*, 1)$	$(-(1 - u_*), u_*)$	$(u_* d(u_*), -d(u_*))$	$(0, \sigma_{\text{edge}})$
$(0, 1 - u_*)$	$(-(1 - u_*), -u_*)$	$(-u_* d(u_*), -d(u_*))$	$(-\sigma_{\text{edge}}, 0)$
$(1 - u_*, 0)$	$(1 - u_*, -u_*)$	$(u_* d(u_*), d(u_*))$	$(0, -\sigma_{\text{edge}})$

These points are stationary for the compactified field even though their normal derivatives are nonzero. We exclude them as limits only after establishing singleton convergence.

The four corners and two starvation lines The corner expert equations give exactly:

(x, y)	(a, b)	Stationary expert coordinates
$(1, 0)$	$(1, 0)$	$(m, d) = (0, g)$
$(0, 1)$	$(-1, 0)$	$(m, d) = (0, -g)$
$(0, 0)$	$(0, -1)$	$(m, d) = (r, r), r \in \mathbb{R}$
$(1, 1)$	$(0, 1)$	$(m, d) = (r, -r), r \in \mathbb{R}$

The last two rows describe the starvation stationary lines. At $(x, y) = (0, 0)$, the unused expert $w_+ = 2r$ is arbitrary and $w_- = 0$; at $(1, 1)$, $w_+ = 0$ and $w_- = 2r$ is arbitrary.

At S_+,

$$F_x = k - qg^2, \qquad F_y = qg^2 - k; \tag{A.43}$$

at S_- the two derivatives are reversed.

Together with the interior calculation, these tables exhaust the compactified stationary set. For $0 < q < 1$, all non-starvation stationary points are isolated at fixed parameters. For $q = 1, k = g^2$, the specialized corners join the interior line to form the closed component $m = b = 0, d = ga, |a| \leq 1$. The starvation lines remain separate connected components.

A.8 Dissipation, omega-limit components, and singleton convergence

The exact dissipation identity is

$$\dot{F}_q = -\mathcal{D}, \qquad \mathcal{D} = F_m^2 + F_d^2 + c_T \left[x^2 (1 - x)^2 F_x^2 + y^2 (1 - y)^2 F_y^2 \right]. \tag{A.44}$$

The trajectory lies in the compact set $\mathcal{K}$ obtained by combining the expert disk in (A.24) with $[0, 1]^2$. Because F_q is bounded below, $\int_0^\infty \mathcal{D}(t)\, dt < \infty$. Both $\mathcal{D}$ and the vector field are polynomials, so $\dot{\mathcal{D}}$ is bounded on $\mathcal{K}$ and $\mathcal{D}(t)$ is uniformly continuous.

Since $\mathcal{D}$ is nonnegative, integrable, and uniformly continuous, $\mathcal{D}(t) \to 0$.

At every omega-limit point, $\mathcal{D} = 0$, so the compactified velocity vanishes. On a probability boundary, this need not imply that F_x or F_y vanishes.

The omega-limit set is nonempty and compact. It is connected as the intersection of nested, compact, connected closures of trajectory tails. Hence it lies in one connected component of the stationary set.

For $0 < q < 1$, suppose this component were the $(x, y) = (0, 0)$ starvation line. Then $s(t) \to 0$; otherwise a bounded subsequence with $s(t_n) \geq \varepsilon > 0$ would have an omega-limit point outside the line. This contradicts the starvation barrier. The other starvation line similarly forces $s(t) \to 2$ and is excluded.

Every remaining component is a singleton. Therefore the omega-limit set contains one point, and the trajectory converges to it. The normal signs from the boundary calculation then rule out inadmissible boundary limits.

For example, convergence to $(x, y) = (1, u_*)$ would give $F_x \to \sigma_{\text{edge}} > 0$ and, by (A.25),

$$\frac{d}{dt}\frac{1}{1-x} = -c_T x^2 F_x \le -c$$

eventually for some $c > 0$. A positive reciprocal cannot decrease at this rate indefinitely. The signs in the table give analogous contradictions on the other three edges using $1/(1-y)$, $1/x$, and $1/y$, respectively. When $k > qg^2$, (A.43) excludes the specialized singleton limits by the same argument.

This proves the list of limits for $0 < q < 1$. The same stationary-component argument proves the classification for $q = 1, k \neq g^2$.

At $q = 1, k = g^2$, the starvation components are still excluded, but the remaining component is the closed critical line. The argument so far gives

$$m(t) \to 0, \qquad b(t) \to 0, \qquad d(t) - ga(t) \to 0.$$

The selection argument below shows that each trajectory converges to a single interior point on this line.

A.9 Stability and the measure-zero exceptional basin

At O, put $\gamma = 1/(4T^2)$. In the order (m, b, a, d), the Jacobian is

$$J_O = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -\gamma\lambda & 0 & 0 \\ 0 & 0 & -\gamma k & \gamma g \\ 0 & 0 & g & -q \end{pmatrix}. \tag{A.45}$$

The contrast eigenvalues are

$$\mu_\pm = -\frac{q+\gamma k}{2} \pm \frac{1}{2}\sqrt{(q-\gamma k)^2 + 4\gamma g^2}, \qquad \mu_+\mu_- = \gamma(qk - g^2). \tag{A.46}$$

If $qk < g^2$, the contrast block has one positive and one negative eigenvalue. Together with -1 and $-\gamma\lambda$, these make the equilibrium hyperbolic with a three-dimensional local stable manifold in finite-logit coordinates.

Let Φ_t denote the flow map and $W^s_{\text{loc}}(O)$ a local stable manifold. Every trajectory converging to O eventually remains in its defining neighborhood, so

$$W^s(O) = \bigcup_{n=0}^{\infty} \Phi_n^{-1}\big(W^s_{\text{loc}}(O)\big). \tag{A.47}$$

Each finite-time flow map is a smooth local diffeomorphism onto its image. Consequently, each preimage in (A.47) is locally three-dimensional and has four-dimensional Lebesgue measure zero; the countable union is also null. The argument requires no global backward flow. The smooth diffeomorphism between finite logits and interior probabilities preserves the null-set conclusion.

Partial equilibria At P_+, and analogously at P_- after sign changes, the Hessian splits into (m, b) and (a, d) blocks:

$$H_{mb} = \begin{pmatrix} 1 & d_* \\ d_* & \lambda + (1-q)d_*^2 \end{pmatrix}, \qquad H_{ad} = \begin{pmatrix} k + (1-q)d_*^2 & 2(1-q)a_*d_* - g \\ 2(1-q)a_*d_* - g & q + (1-q)a_*^2 \end{pmatrix}. \tag{A.48}$$

The stationary relations give $qd_*^2 = ka_*^2$. Thus

$$\det H_{mb} = \lambda - qd_*^2 = \lambda(1 - \chi a_*^2) > 0, \qquad \det H_{ad} = \frac{4g^2 q(1-q)a_*^2}{[q + (1-q)a_*^2]^2} > 0. \tag{A.49}$$

The diagonal entries are also positive. Hence the full Hessian is positive definite, and the interior logit transformation is nonsingular. Euclidean gradient flow in logits is locally exponentially asymptotically stable at each partial equilibrium.

Fully specialized equilibria below the lower threshold Near S_+ write $z = 1 - a$, so feasibility gives $z \geq |b|$. If $k < qg^2$, Taylor expansion of the smooth function Φ_q near $(1, 0)$ gives

$$\Phi_q(1 - z, b) - \Phi_q(1, 0) = (qg^2 - k)z + O(z^2), \qquad |b| \leq z. \tag{A.50}$$

The remainder is uniform on this cone. Thus the reduced objective increases under sufficiently small nonzero feasible router perturbations. The positive expert-square terms in (A.36) make S_+ a strict constrained local minimum of the full objective. The same argument applies to S_-.

To obtain asymptotic stability from strict constrained minimality, we also use isolation of these stationary points. Choose a sufficiently small relative closed ball around one such point P containing no other compactified stationary point. The continuous function $F_q - F_q(P)$ has a positive minimum on the relative boundary of the ball. Because F_q is nonincreasing, a lower sublevel set containing P cannot cross this boundary. Trajectories starting sufficiently near P remain in a compact neighborhood, and dissipation confines their omega-limit sets to P.

Sublevel sets in arbitrarily small neighborhoods give Lyapunov stability. The absence of other stationary points in one such neighborhood gives local attraction. The same argument applies at both critical thresholds after strict minimality and isolation are established.

A.10 Upper critical threshold: $k = g^2/q$, $0 < q < 1$

At this parameter value,

$$\Phi_q(a, 0) = \frac{g^2(1-q)a^4}{2q[q + (1-q)a^2]}. \tag{A.51}$$

This identity has leading term $g^2(1-q)a^4/(2q^2)$ and is positive for every sufficiently small nonzero a. The transverse difference is

$$\Phi_q(a, b) - \Phi_q(a, 0) = \frac{b^2}{2}\left[\lambda - \frac{g^2 q a^2}{D_q(a, b)D_q(a, 0)}\right]. \tag{A.52}$$

Near O, the bracket is at least $\lambda/2 > 0$. Together, (A.36), (A.51), and (A.52) establish strict local minimality of the full objective.

The stationary classification makes O isolated at this parameter value, so the sublevel-set argument gives local asymptotic stability. By (A.45), one eigenvalue is zero and the other three are negative; the equilibrium is nonhyperbolic.

The quartic expansion establishes local stability, while the stationary-component classification gives global attraction for every finite-logit initialization.

A.11 Lower critical threshold: $k = qg^2$, $0 < q < 1$

Near S_+, set $z = 1 - a \geq |b|$. The constrained expansion is

$$\Phi_q(a, b) - \Phi_q(1, 0) = 2q(1-q)g^2 z^2 + \frac{\lambda - qg^2}{2} b^2 + O(z^3). \tag{A.53}$$

Because D_q is positive near $(1,0)$, Φ_q has bounded third derivatives there. Its Taylor remainder is bounded by a constant times $(z+|b|)^3$, hence by a constant times z^3 on the feasible cone. The first-order term vanishes at $k = qg^2$; the mixed second-order term vanishes because Φ_q is even in b.

Because $\chi \le 1$, we have $\lambda \ge k = qg^2$. The first quadratic coefficient in (A.53) is strictly positive and controls the cubic remainder for sufficiently small z, even when $\chi = 1$ makes the displayed b^2 coefficient zero. If $z = 0$, feasibility forces $b = 0$. The expert squares then establish strict constrained local minimality of S_+.

No additional stationary continuum appears: the nonzero stationary equation forces $a^2 = 1$, and the open-edge equation has no root. The specialized points are isolated, so the sublevel-set argument proves their asymptotic stability at equality.

A.12 The $q = 1$ critical continuum and single-point selection

Fix $q = 1, k = g^2$. The stationary-component argument shows that the trajectory approaches the closed critical line, so $m, b, d - ga \to 0$. It remains to prove convergence of a.

For the functions in (A.14),

$$\Xi'(p) = \frac{1}{p^2(1-p)^2}.$$

Equations (A.19) and (A.21) give

$$\frac{d}{dt}\Xi_-(x,y) = gd - ka, \qquad \frac{d}{dt}\Xi_+(x,y) = -md - \lambda b, \tag{A.54}$$

and, at criticality,

$$\frac{d}{dt}\big(\Xi_-(x,y) + gd\big) = -gbm. \tag{A.55}$$

The quantity $\Xi_- + gd$ can drift along the full four-dimensional flow, although it is conserved on the invariant plane $m = b = 0$. We show that the drift is integrable.

From (A.21) and $|a| \le 1 \pm b$, the excesses $(|w_\pm| - g)_+$ satisfy upper Dini-derivative bounds

$$D^+(|w_+| - g)_+ \le -(1+b)(|w_+| - g)_+, \qquad D^+(|w_-| - g)_+ \le -(1-b)(|w_-| - g)_+.$$

Since $b \to 0$, both rates are eventually at least $1/2$. Thus

$$\begin{aligned} e(t) &:= \big(\max\{|w_+(t)|, |w_-(t)|\} - g\big)_+ \in L^1([0,\infty)), \\ |m| + |d| &= \max\{|w_+|, |w_-|\} \le g + e(t). \end{aligned} \tag{A.56}$$

The function Ξ is increasing and satisfies $\Xi(1-p) = -\Xi(p)$. Hence Ξ_+ and b have the same sign. For $V = |\Xi_+| + g|m|$, equations (A.54) and $\dot m = -m - bd$ imply

$$\begin{aligned} D^+V &\le (|d| - g)|m| + (g|d| - \lambda)|b| \\ &\le -m^2 - g|bm| - (\lambda - g^2)|b| + e(t)(|m| + g|b|). \end{aligned} \tag{A.57}$$

Criticality implies $\chi > 0$ and $\lambda = g^2/\chi \ge g^2$. By (A.56) and boundedness, the last term in (A.57) is integrable. Integrating the inequality bounds V and gives

$$\int_0^\infty |b(t)m(t)|\, dt < \infty.$$

Equation (A.55) now shows that $\Xi_- + gd$ has a finite limit. Boundedness of d then bounds Ξ_-, while V bounds Ξ_+. Therefore

$$\Xi(x) = c_T(\Xi_- + \Xi_+), \qquad \Xi(y) = c_T(\Xi_+ - \Xi_-)$$

are bounded. Because Ξ maps $(0, 1)$ increasingly onto $\mathbb{R}$, both probabilities stay uniformly away from the endpoints. Thus finite-logit trajectories cannot converge to the endpoints of the critical line.

On $b = 0$, substitution into (A.14) gives

$$\Xi_-(x, y) = 2T^2 \left[\frac{a}{1 - a^2} + \operatorname{arctanh} a \right]. \tag{A.58}$$

Consequently, every subsequential limit satisfies the selection relation (A.15). The left-hand side has derivative

$$\frac{4T^2}{(1 - a^2)^2} + g^2 > 0, \qquad -1 < a < 1, \tag{A.59}$$

and ranges from $-\infty$ to $+\infty$. Thus exactly one limiting value of a is possible, proving convergence of the full trajectory to a single point. Inverting the router gives the stated limits of α and β.

Every point on the interior critical line is a finite-logit stationary initialization, so the selected limit genuinely depends on initialization. For $0 < q < 1$, (A.39) instead has isolated roots, and the quartic term at the upper threshold is positive away from zero. The continuum occurs only at $q = 1, k = g^2$.

This completes the proof of Theorem 1. □

A.13 Selected-expert identities for later results

For $q = 1$, put $s = x + y = 1 + b$, $R_0 = \max\{g, |w_+(0)|, |w_-(0)|\}$, and $C_{\mathrm{grad}} = R_0^2/4 + gR_0 + \lambda$. The selected-expert flow (A.21) preserves $|w_\pm| \le R_0$ and gives

$$\dot{m} = -m - bd, \qquad \dot{d} = -d - bm + ga. \tag{25}$$

For $p = x$ or y with $F_p = h_p$,

$$\frac{d}{dt}\frac{1}{p} = c_T(1 - p)^2 h_p, \qquad \frac{d}{dt}\frac{1}{1 - p} = -c_T p^2 h_p. \tag{26}$$

Both derivatives have absolute value at most $c_T C_{\mathrm{grad}}$. The selected-expert probability barrier is

$$L(p) = \log\frac{p}{1 - p} + \frac{1}{1 - p}, \qquad \mathcal{H}_+ = L(x) + L(y) - \frac{w_+^2 + w_-^2}{2T^2}. \tag{27}$$

Its drift and a near-zero-load bound are

$$\dot{\mathcal{H}}_+ = \frac{sw_+^2 + (4 - s)w_-^2}{2T^2} + \frac{2\lambda}{T^2}s(1 - s) - \frac{2k}{T^2}a^2, \tag{28}$$

$$\dot{\mathcal{H}}_+ \ge c_T\lambda s\{1 - (1 + \chi)s\}. \tag{29}$$

The (a, d) linearization at O is

$$\begin{pmatrix} \dot{a} \\ \dot{d} \end{pmatrix} = \begin{pmatrix} -k/(4T^2) & g/(4T^2) \\ g & -1 \end{pmatrix} \begin{pmatrix} a \\ d \end{pmatrix}. \tag{30}$$

On the zero-weight invariant line with $\alpha = 0$, $x = y = p$ and

$$\dot{p} = \frac{2\lambda}{T^2}p^2(1 - p)^2(1 - 2p), \qquad \tau_r \sim \frac{T^2}{2\lambda}e^{M/T}. \tag{31}$$

A.14 UNIFORM-ROUTER INITIALIZATION AT $q = 1$

Write $\gamma = (4T^2)^{-1}$. To analyze uniform-router initialization, use the symmetries

$$(m, d, a, b) \mapsto (m, -d, -a, -b), \qquad (m, d, a, b) \mapsto (-m, d, a, -b)$$

to reduce to $m(0) \geq 0$, $d(0) > 0$. The cone

$$m, d, a \geq 0, \qquad b \leq 0 \tag{32}$$

is invariant. Invariance of the m, d faces follows from (25). At $a = 0$, the coefficients in the probability equations coincide, and $\dot{a} = 2c_T x^2(1-x)^2 gd \geq 0$. At $b = 0$, they also coincide, and $\dot{b} = -2c_T x^2(1-x)^2 md \leq 0$. Finally, $\dot{d} \geq -d$, so $d(t) > 0$ at every finite time.

To exclude convergence to the identical-expert equilibrium within the cone, define coordinates that remove the vanishing coefficients from the routing equations:

$$\Xi(p) = -p^{-1} + (1-p)^{-1} + 2\log\frac{p}{1-p}, \quad \Xi_-(x, y) = \frac{\Xi(x) - \Xi(y)}{2c_T}, \quad \Xi_+(x, y) = \frac{\Xi(x) + \Xi(y)}{2c_T}. \tag{33}$$

These coordinates obey

$$\frac{d}{dt}\Xi_-(x, y) = gd - ka, \qquad \frac{d}{dt}\Xi_+(x, y) = -md - \lambda b. \tag{34}$$

Because Ξ is increasing and odd around $1/2$, Ξ_- and Ξ_+ have the signs of a and b, respectively. Near the identical-expert equilibrium, $a/\Xi_- \to \gamma$. If a cone trajectory converged to this equilibrium with $\lambda > 0$ and $k < g^2$, then for sufficiently small $\varepsilon > 0$ it would eventually satisfy

$$\frac{d}{dt}\Xi_-(x, y) \geq -k(\gamma + \varepsilon)\Xi_- + gd, \qquad \dot{d} \geq g(\gamma - \varepsilon)\Xi_- - d.$$

Equivalently, the coordinatewise comparison is

$$\begin{pmatrix} \frac{d}{dt}\Xi_-(x, y) \\ \dot{d} \end{pmatrix} \geq \begin{pmatrix} -k(\gamma + \varepsilon) & g \\ g(\gamma - \varepsilon) & -1 \end{pmatrix} \begin{pmatrix} \Xi_- \\ d \end{pmatrix}, \tag{35}$$

Choose ε so that $k(\gamma + \varepsilon) < g^2(\gamma - \varepsilon)$. The comparison matrix has positive off-diagonal entries and negative determinant, so it has a positive eigenvalue and a strictly positive left eigenvector. The corresponding positive linear functional of (Ξ_-, d) grows exponentially, contradicting convergence to zero. The opposite specialized point lies outside the cone. This proves specialization. The argument also covers any initial state in (32) with $d(0) > 0$, including $w_+(0) > 0, w_-(0) = 0, \alpha(0) = 0, \beta(0) = -M$.

At $\lambda = 0$, $\frac{d}{dt}\Xi_-(x, y) = gd > 0$ on this cone. After any positive time, $\Xi_- > 0$ and continues to increase, ruling out convergence to a finite identical-expert equilibrium with $a = \Xi_- = 0$. The zero-balancing convergence result below therefore gives specialization.

A.15 ZERO BALANCING, PREDICTION RISK, AND SPECIALIZATION RATES AT $q = 1$

When $q = 1, \lambda = 0$, the interior stationary equations (A.37) give $m = d = a = 0$ with arbitrary $b \in (-1, 1)$. Formula (A.41) is strictly positive on each open edge, leaving the specialized corners and the two starvation lines as the other stationary components. The possible limit components are therefore isolated specialized points and the identical-expert continuum, including its endpoints. Both barriers obtained by exchanging experts are globally nondecreasing by (28). They exclude starvation even as a subsequential limit: approaching either starvation line would force the corresponding barrier to $-\infty$. If the limit set lies on the identical-expert continuum, it is a compact subset of its interior. The probabilities are then eventually bounded away from zero and one, and $w_+, w_-, a \to 0$.

For all sufficiently large times, $\mathcal{H}_+$ is bounded above. With $\lambda = 0$, the identity (28) implies $\int_0^\infty (w_+^2 + w_-^2)\, dt < \infty$. Since $|md| \leq (w_+^2 + w_-^2)/4$, equation (34) shows that Ξ_+ converges. As $a \to 0$, this coordinate approaches $\Xi((1+b)/2)/c_T$, which is strictly increasing in b. Hence b converges to one value in $(-1, 1)$, completing the zero-balancing classification.

Evaluate the prediction part of (5) to obtain the limiting risks. It equals $-g^2/2$ at a specialized state, zero at an identical-expert state, and $-g^2a^2/2$ on the critical line. Adding $\mathbb{E}[Y^2]/2$ gives the absolute risks corresponding to the improvements stated in Section 2. The identity $\mathbb{E}[Y \mid S, U] = gSU$ shows that the specialized prediction is the conditional mean.

Finally, suppose the trajectory converges to $(g, -g, 1, 0)$, and set $\kappa = g^2 - k > 0$. Then $h_x \to -\kappa$, $h_y \to \kappa$, and (26) gives

$$\frac{d}{dt}\frac{1}{1-x} \to c_T\kappa, \qquad \frac{d}{dt}\frac{1}{y} \to c_T\kappa.$$

Integrating and dividing by t gives $1 - x \sim y \sim 1/(c_T\kappa t)$. The logits recovered from the probabilities satisfy $\log(x/(1-x)) = \log(c_T\kappa t) + o(1)$ and $\log(y/(1-y)) = -\log(c_T\kappa t) + o(1)$. Therefore $\alpha = T\log(c_T\kappa t) + o(1)$, $\beta = o(1)$, and $a = 1 - T^2/(\kappa t) + o(t^{-1})$.

A.16 Batch-cue second moments

For independent cues, all off-diagonal terms in $\mathbb{E}\bar{S}^2 = B^{-2}\sum_{i,j}\mathbb{E}[S_iS_j]$ vanish, giving $\chi = 1/B$. If all examples share one latent task and the cues are conditionally independent given it, then $\mathbb{E}[S_iS_j] = \rho^2$ for $i \neq j$, so

$$\chi = \rho^2 + \frac{1-\rho^2}{B}.$$

If the batch is partitioned into n_{grp} equally sized independent task groups, only within-group off-diagonal terms contribute ρ^2, giving $\chi = \rho^2/n_{\mathrm{grp}} + (1-\rho^2)/B$. These examples illustrate the batch statistic in (4); in general, χ differs from $1/B$.

B Local curvature–coupling theory

We use the assumptions of Section 3.

Let $E \geq 2$ experts belong to the same real-valued function family $f_w(U)$, with separate parameters $w_j \in \mathbb{R}^{d_w}$. Let $\boldsymbol{\psi}(S) \in \mathbb{R}^{d_\psi}$ be a fixed, bounded feature vector. The expert input U and router input S may be dependent. For fixed $0 < T < \infty$, use

$$p_j(S) = \frac{\exp(\theta_j^\top \boldsymbol{\psi}(S)/T)}{\sum_{\ell=1}^{E}\exp(\theta_\ell^\top \boldsymbol{\psi}(S)/T)}, \qquad \sum_{j=1}^{E}\theta_j = 0. \tag{B.1}$$

The zero-sum constraint removes common-logit translations but permits redundant routing features.

For the prediction-dispersion family $0 \leq q \leq 1$ and $\lambda \geq 0$, define

$$\begin{aligned}\mathcal{F}_q = \frac{1}{2}\mathbb{E}[(\bar{f} - Y)^2] + \frac{q}{2}\mathbb{E}\left[\sum_{j=1}^{E} p_j(f_{w_j} - \bar{f})^2\right] \\ + \frac{\lambda E}{2}\mathbb{E}_{\mathrm{batch}}\sum_{j=1}^{E}(\bar{p}_j - 1/E)^2, \qquad \bar{f} = \sum_j p_j f_{w_j}.\end{aligned} \tag{B.2}$$

Here $\bar{p}_j = B^{-1}\sum_i p_j(S_i)$ for a batch of size B. Batch cues have the population marginals but may be dependent within a batch. At $E = 2$, the prefactor $\lambda E/2$ equals λ, matching the binary balancing normalization.

Assume that $f_w(U)$ is C^3 near w_0, that the function and its first three parameter derivatives have square-integrable local envelopes, and that $\mathbb{E}[Y^2] < \infty$. These conditions justify differentiation and the second-order expansions under expectation. The parameterization does not include a shared trainable encoder.

Define

$$\begin{gathered}\varepsilon = Y - f_{w_0}(U), \qquad J_w = \nabla_w f_{w_0}(U), \qquad J_{ww} := \nabla_w^2 f_{w_0}(U), \\ \Sigma_J = \mathbb{E}[J_w J_w^\top], \qquad \mathsf{N} = \mathbb{E}[\varepsilon J_{ww}], \qquad \mathsf{H} = \Sigma_J - \mathsf{N}, \qquad \mathsf{H}_q = q\Sigma_J - \mathsf{N}, \\ \mathsf{G} = \mathbb{E}[\boldsymbol{\psi}(S)\varepsilon J_w^\top], \qquad \mathsf{C}_\psi = \mathbb{E}_{\mathrm{batch}}[\bar{\boldsymbol{\psi}}\bar{\boldsymbol{\psi}}^\top], \qquad \bar{\boldsymbol{\psi}} = B^{-1}\sum_i \boldsymbol{\psi}(S_i).\end{gathered} \tag{B.3}$$

The matrix C_ψ is the uncentered second moment of the batch-mean routing features and need not be a covariance matrix.

Quadratic form and block criterion For perturbations $w_j = w_0 + u + \Delta w_j$ and $\theta_j = T\widetilde{\theta}_j$, impose $\sum_j \Delta w_j = \sum_j \widetilde{\theta}_j = 0$. Let $\Delta_2\mathcal{F}_q$ denote the homogeneous quadratic term, including the factor $1/2$ in the Hessian convention.

Suppose that w_0 is stationary for the single-expert population risk and that $\mathsf{H} \succ 0$. Then $w_j = w_0$, $\theta_j = 0$ is stationary for $\mathcal{F}_q$. Its second-order term is

$$\Delta_2\mathcal{F}_q = \frac{1}{2}u^\top \mathsf{H} u + \frac{1}{2E}\sum_{j=1}^{E}\left[\Delta w_j^\top \mathsf{H}_q \Delta w_j - 2\widetilde{\theta}_j^\top \mathsf{G}\Delta w_j + \lambda\widetilde{\theta}_j^\top \mathsf{C}_\psi \widetilde{\theta}_j\right]. \tag{B.4}$$

After an expert-index contrast decomposition, the Hessian is congruent to one common-expert block H and $E-1$ blocks $E^{-1}\mathcal{K}_q$, where

$$\mathcal{K}_q = \begin{pmatrix} \lambda\mathsf{C}_\psi & -\mathsf{G} \\ -\mathsf{G}^\top & \mathsf{H}_q \end{pmatrix}. \tag{B.5}$$

For gradient flow with a smooth positive-definite metric on the constrained parameter space, the homogeneous equilibrium is locally exponentially asymptotically stable if and only if $\mathcal{K}_q \succ 0$. A negative eigenvalue of $\mathcal{K}_q$ implies both linear and nonlinear instability. Positive semidefiniteness has the exact characterization

$$\mathcal{K}_q \succeq 0 \iff \begin{cases} \mathsf{H}_q \succeq 0, \\ \mathsf{G}\ker(\mathsf{H}_q) = \{0\}, \\ \lambda\mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^\dagger\mathsf{G}^\top \succeq 0. \end{cases} \tag{B.6}$$

A negative-curvature expert-contrast direction remains unstable under balancing. The same holds at every finite balancing strength for a zero-curvature expert-contrast direction coupled nontrivially through G. If $\mathcal{K}_q$ is singular positive semidefinite, second-order information alone does not determine nonlinear stability.

Proof of Theorem 2. We establish stationarity and the quadratic expansion, then identify the contrast blocks and analyze their stability implications.

Homogeneous-point stationarity Stationarity of the single-expert risk means

$$\nabla_w \frac{1}{2}\mathbb{E}[(f_w(U) - Y)^2]\Big|_{w=w_0} = -\mathbb{E}[\varepsilon J_w] = 0.$$

At $w_j = w_0$, $\theta_j = 0$, the routing probabilities equal $1/E$. The gradient with respect to each expert is $-\mathbb{E}[\varepsilon J_w]/E = 0$. The output-variance term and its first derivatives vanish because the expert outputs coincide.

When the experts coincide, the prediction objective is independent of the router. Uniform routing also makes every batch-balancing residual $\bar{p}_j - 1/E$ zero. Hence the router gradient vanishes.

This argument uses unconditional single-expert stationarity; it does not require $\mathbb{E}[\varepsilon J_w \mid S] = 0$. At nonuniform routing, identical experts need not have zero individual gradients.

Softmax Taylor expansion Take a perturbation path

$$w_j(t) = w_0 + t(u + \Delta w_j), \qquad \theta_j(t) = t\Theta_j, \qquad \sum_j \Delta w_j = \sum_j \Theta_j = 0, \qquad \widetilde{\theta}_j := \Theta_j/T. \tag{B.7}$$

Set $z_j(S) = \widetilde{\theta}_j^\top \boldsymbol{\psi}(S)$, so $\sum_j z_j = 0$. Expanding numerator and denominator of (B.1) gives

$$p_j(t) = \frac{1}{E} + \frac{t}{E} z_j + \frac{t^2}{2E}\left(z_j^2 - \frac{1}{E}\sum_\ell z_\ell^2\right) + O(t^3). \tag{B.8}$$

For each fixed perturbation direction, bounded routing features make the remainder uniform over samples. The second-order probability coefficients sum to zero.

Expert Taylor expansion The expert expansion is

$$f_{w_j(t)}(U) = f_{w_0}(U) + tJ_w^\top(u + \Delta w_j) + \frac{t^2}{2}(u + \Delta w_j)^\top J_{ww}(u + \Delta w_j) + O(t^3). \tag{B.9}$$

The remainder is bounded by a square-integrable envelope times a direction-dependent constant. Its product with the residual is integrable by Cauchy–Schwarz, as are products involving lower derivatives. Thus the assumptions justify the risk expansion through second order.

Average-risk quadratic form Multiplying (B.8) and (B.9) and summing over experts gives

$$\begin{aligned}\bar{f}(t) = f_{w_0} + tJ_w^\top u + t^2\Bigg[&\frac{1}{2}u^\top J_{ww}u + \frac{1}{2E}\sum_j \Delta w_j^\top J_{ww}\Delta w_j \\ &+ \frac{1}{E}\sum_j z_j J_w^\top \Delta w_j\Bigg] + O(t^3).\end{aligned} \tag{B.10}$$

Terms coupling the common expert perturbation to expert contrasts vanish because $\sum_j \Delta w_j = 0$; terms coupling it to the router vanish because $\sum_j z_j = 0$. The pure second-order router term multiplying f_{w_0} vanishes because the second-order probability coefficients sum to zero. These cancellations hold for each sample, regardless of dependence between S and U.

Since the baseline prediction residual is $-\varepsilon$, the quadratic term in the average risk is

$$\Delta_2\mathcal{R}_{\text{avg}} = \frac{1}{2}u^\top(\Sigma_J - \mathsf{N})u - \frac{1}{2E}\sum_j \Delta w_j^\top \mathsf{N}\Delta w_j - \frac{1}{E}\sum_j \widetilde{\theta}_j^\top \mathsf{G}\Delta w_j. \tag{B.11}$$

The linear term vanishes because $\mathbb{E}[\varepsilon J_w] = 0$. Equation (B.11) gives the signs $-\mathsf{N}$ and $-\mathsf{G}$ directly from the objective. In particular, the coupling is not multiplied by q.

Conditional output-variance quadratic form At first order,

$$f_{w_j(t)} - \bar{f}(t) = tJ_w^\top \Delta w_j + O(t^2).$$

Therefore

$$\sum_j p_j(t)(f_{w_j(t)} - \bar{f}(t))^2 = \frac{t^2}{E}\sum_j (J_w^\top \Delta w_j)^2 + O(t^3),$$

and hence

$$\Delta_2 \mathcal{R}_{\text{var}} = \frac{q}{2E} \sum_j \Delta w_j^\top \Sigma_J \Delta w_j. \tag{B.12}$$

A first-order router perturbation multiplied by the squared output difference contributes only at third order. Hence the expert-contrast curvature is $q\Sigma_J - \mathsf{N}$, not $q(\Sigma_J - \mathsf{N})$.

Balancing quadratic form Averaging (B.8) over a batch gives

$$\bar{p}_j(t) - 1/E = \frac{t}{E} \widetilde{\theta}_j^\top \bar{\boldsymbol{\psi}} + O(t^2).$$

The balancing term contributes

$$\frac{\lambda E}{2E^2} \sum_j \mathbb{E}_{\text{batch}}[(\widetilde{\theta}_j^\top \bar{\boldsymbol{\psi}})^2] = \frac{\lambda}{2E} \sum_j \widetilde{\theta}_j^\top \mathsf{C}_\psi \widetilde{\theta}_j. \tag{B.13}$$

For independent identically distributed batch cues with feature mean μ, $\mathsf{C}_\psi = B^{-1} \operatorname{Cov}(\boldsymbol{\psi}) + \mu\mu^\top$. This differs from $B^{-1}\mathbb{E}[\boldsymbol{\psi}\boldsymbol{\psi}^\top]$ when $B > 1$ and $\mu \neq 0$.

Adding (B.11)–(B.13) gives (B.4), with an $O(t^3)$ expected remainder along each fixed direction. In the original router perturbations Θ_j, the quadratic term is

$$\Delta_2 \mathcal{F}_q = \frac{1}{2} u^\top \mathsf{H} u + \frac{1}{2E} \sum_j \left[\Delta w_j^\top \mathsf{H}_q \Delta w_j - \frac{2}{T} \Theta_j^\top \mathsf{G} \Delta w_j + \frac{\lambda}{T^2} \Theta_j^\top \mathsf{C}_\psi \Theta_j \right]. \tag{B.14}$$

Expert-index contrast decomposition and temperature congruence Choose $Q \in \mathbb{R}^{E \times (E-1)}$ with

$$Q^\top Q = I, \qquad Q^\top \mathbf{1} = 0, \qquad \Delta w_j = \sum_{h=1}^{E-1} Q_{jh} \Delta w^{(h)}, \qquad \widetilde{\theta}_j = \sum_{h=1}^{E-1} Q_{jh} \Delta \theta^{(h)}. \tag{B.15}$$

Orthogonality converts each expert-index sum in (B.4) into a sum over h without cross terms. Thus

$$\Delta_2 \mathcal{F}_q = \frac{1}{2} u^\top \mathsf{H} u + \frac{1}{2E} \sum_{h=1}^{E-1} \begin{pmatrix} \Delta \theta^{(h)} \\ \Delta w^{(h)} \end{pmatrix}^\top \mathcal{K}_q \begin{pmatrix} \Delta \theta^{(h)} \\ \Delta w^{(h)} \end{pmatrix}. \tag{B.16}$$

In these coordinates, the Hessian is $\operatorname{diag}(\mathsf{H}, E^{-1}\mathcal{K}_q, \ldots, E^{-1}\mathcal{K}_q)$. If the common expert coordinate is normalized as $\sqrt{E}\, u$, its block becomes $E^{-1}\mathsf{H}$. The total dimension is $d_w + (E-1)(d_\psi + d_w) = Ed_w + (E-1)d_\psi$, consistent with the zero-sum constraint.

For original router coordinates rather than $\widetilde{\theta}_j = \Theta_j / T$, each contrast block contains

$$\mathcal{K}_{q,T} = \begin{pmatrix} \lambda \mathsf{C}_\psi / T^2 & -\mathsf{G}/T \\ -\mathsf{G}^\top / T & \mathsf{H}_q \end{pmatrix} = D_T \mathcal{K}_q D_T, \qquad D_T = \operatorname{diag}(T^{-1} I_{d_\psi}, I_{d_w}). \tag{B.17}$$

These matrices are congruent and have the same inertia, although their eigenvalues generally differ. Each Hessian contrast block also carries a factor of $1/E$. Thus neither expert count nor fixed positive temperature changes the sign criterion, but both can affect rates.

The blocks decouple only to second order. Higher-order terms may couple common and contrast modes, and a general positive-definite metric may mix the blocks in the linearized dynamics. The Hessian inertia is unchanged.

Stability implications In local coordinates on the zero-sum parameter space, write the gradient flow as $\dot{z} = -\mathcal{M}(z)\nabla\mathcal{F}_q(z)$, where $\mathcal{M}$ is a smooth symmetric positive-definite inverse metric. Let $\mathcal{H}_{\text{full}}$ denote the full Hessian at the homogeneous stationary point. Since the gradient vanishes there,

$$D\dot{z}(0) = -\mathcal{M}(0)\mathcal{H}_{\text{full}}, \qquad -\mathcal{M}(0)\mathcal{H}_{\text{full}} \sim -\mathcal{M}(0)^{1/2}\mathcal{H}_{\text{full}}\mathcal{M}(0)^{1/2}. \tag{B.18}$$

The right-hand matrix is symmetric and has the inertia of $-\mathcal{H}_{\text{full}}$. Because the common-expert block $\mathsf{H} \succ 0$ by assumption, $\mathcal{K}_q \succ 0$ makes $\mathcal{H}_{\text{full}}$ positive definite and the linearization Hurwitz. A negative direction of $\mathcal{K}_q$ gives a positive eigenvalue of the linearization and hence nonlinear instability.

Conversely, local exponential asymptotic stability of a smooth equilibrium implies exponential decay of the derivative flow there, so the linearization must be Hurwitz. By (B.18), this requires $\mathcal{H}_{\text{full}} \succ 0$, equivalently $\mathcal{K}_q \succ 0$ under $\mathsf{H} \succ 0$.

If $\mathcal{K}_q$ is singular positive semidefinite, the linearization has zero eigenvalues but no positive eigenvalue. The full parameter point cannot be locally exponentially asymptotically stable, although nonlinear Lyapunov stability and slower attraction remain possible. The examples below show different outcomes.

Ordinary Schur complement Suppose $\mathsf{H}_q \succ 0$. Completing squares gives

$$\begin{aligned} &v^\top \mathsf{H}_q v - 2a^\top \mathsf{G} v + \lambda a^\top \mathsf{C}_\psi a \\ &\quad = (v - \mathsf{H}_q^{-1}\mathsf{G}^\top a)^\top \mathsf{H}_q (v - \mathsf{H}_q^{-1}\mathsf{G}^\top a) + a^\top(\lambda \mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^{-1}\mathsf{G}^\top)a. \end{aligned} \tag{B.19}$$

Therefore

$$\begin{aligned} \mathcal{K}_q \succeq 0 &\iff \lambda \mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^{-1}\mathsf{G}^\top \succeq 0, \\ \mathcal{K}_q \succ 0 &\iff \lambda \mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^{-1}\mathsf{G}^\top \succ 0. \end{aligned} \qquad (\mathsf{H}_q \succ 0). \tag{B.20}$$

These conditions concern matrix definiteness; equality alone does not determine nonlinear stability.

Generalized Schur complement If $\mathcal{K}_q \succeq 0$, the pure expert direction forces $\mathsf{H}_q \succeq 0$. For $v_0 \in \ker \mathsf{H}_q$, any a, and any real t,

$$\begin{pmatrix} a \\ tv_0 \end{pmatrix}^\top \mathcal{K}_q \begin{pmatrix} a \\ tv_0 \end{pmatrix} = \lambda a^\top \mathsf{C}_\psi a - 2t\, a^\top \mathsf{G} v_0.$$

Nonnegativity for every t forces $\mathsf{G}v_0 = 0$. Thus $\mathsf{G}\ker \mathsf{H}_q = \{0\}$ is necessary.

Under this condition, $\operatorname{range}(\mathsf{G}^\top) \subseteq \operatorname{range}(\mathsf{H}_q)$, so $\mathsf{H}_q\mathsf{H}_q^\dagger\mathsf{G}^\top = \mathsf{G}^\top$. The exact identity is then

$$\begin{aligned} &v^\top \mathsf{H}_q v - 2a^\top \mathsf{G} v + \lambda a^\top \mathsf{C}_\psi a \\ &\quad = (v - \mathsf{H}_q^\dagger \mathsf{G}^\top a)^\top \mathsf{H}_q (v - \mathsf{H}_q^\dagger \mathsf{G}^\top a) + a^\top(\lambda \mathsf{C}_\psi - \mathsf{G}\mathsf{H}_q^\dagger \mathsf{G}^\top)a. \end{aligned} \tag{B.21}$$

This proves both directions of (B.6). The kernel condition is required when the inverse is replaced by a pseudoinverse. If H_q is singular, $(0, v_0)$ for nonzero $v_0 \in \ker \mathsf{H}_q$ prevents strict positive definiteness even when all semidefinite conditions hold.

Finite-balancing impossibility and routing-kernel obstructions If $v^\top \mathsf{H}_q v < 0$, the direction $(a, v) = (0, v)$ already has negative quadratic form; balancing cannot alter it.

Now suppose $v \in \ker \mathsf{H}_q$ and $h = \mathsf{G}v \neq 0$. Set $a = \tau h$ with $\tau > 0$. Then

$$\begin{pmatrix} \tau h \\ v \end{pmatrix}^\top \mathcal{K}_q \begin{pmatrix} \tau h \\ v \end{pmatrix} = \lambda \tau^2 h^\top \mathsf{C}_\psi h - 2\tau \|h\|^2 < 0 \tag{B.22}$$

for sufficiently small τ at every fixed finite λ. This explicitly constructs a negative-curvature direction.

The direction corresponds to a feasible multi-expert perturbation: set $\Delta w_j = Q_{j1} v$ and $\widetilde{\theta}_j = Q_{j1} a$ using the orthonormal contrast basis (B.15). Both zero-sum constraints hold, and the full quadratic term equals the negative expression in (B.22) divided by $2E$. Scaling the perturbation places it in any prescribed neighborhood. The router perturbations in the original coordinates are $\Theta_j = T\widetilde{\theta}_j$, valid for every fixed positive finite temperature.

Routing curvature gives an analogous obstruction. If $a_0 \in \ker \mathsf{C}_\psi$ and $h = \mathsf{G}^\top a_0 \neq 0$, take $v = \tau h$. Then

$$\begin{pmatrix} a_0 \\ \tau h \end{pmatrix}^\top \mathcal{K}_q \begin{pmatrix} a_0 \\ \tau h \end{pmatrix} = -2\tau \|h\|^2 + \tau^2 h^\top \mathsf{H}_q h < 0 \tag{B.23}$$

for sufficiently small positive τ. Thus increasing λ cannot stabilize a signal-coupled routing direction that balancing does not penalize. Condition (B.6) includes this obstruction.

If G annihilates an expert zero-curvature direction, this argument does not apply. Higher-order terms then determine its behavior.

Linear-expert specialization For $f_w(U) = w^\top U$, one has $J_{ww} = 0$, $\mathsf{N} = 0$, and $\Sigma_J = \Sigma := \mathbb{E}[UU^\top]$. Hence $\mathsf{H} = \Sigma$ and $\mathsf{H}_q = q\Sigma$.

For $q > 0$ and $\Sigma \succ 0$, (B.20) becomes

$$\mathcal{K}_q \succeq 0 \iff \lambda q\, \mathsf{C}_\psi - \mathsf{G}\Sigma^{-1}\mathsf{G}^\top \succeq 0, \tag{B.24}$$

The same equivalence holds for strict positive definiteness. The coupling retains the residual:

$$\mathsf{G} = \mathbb{E}[\boldsymbol{\psi}(S)(Y - w_0^\top U)U^\top].$$

When $w_0 \neq 0$ and the inputs are correlated, the coupling generally differs from $\mathbb{E}[\boldsymbol{\psi} Y U^\top]$.

If also $\mathsf{C}_\psi \succ 0$, define

$$\nu_{\mathrm{spec}} := \left\| \mathsf{C}_\psi^{-1/2} \mathsf{G} \Sigma^{-1/2} \right\|_{\mathrm{op}}^2. \tag{B.25}$$

Then $\lambda q > \nu_{\mathrm{spec}}$ gives local exponential asymptotic stability, $\lambda q < \nu_{\mathrm{spec}}$ gives instability, and $\lambda q = \nu_{\mathrm{spec}}$ gives a singular positive-semidefinite Hessian. The spectral threshold alone does not determine nonlinear stability at equality.

At $q = 0$, linear experts have $\mathsf{H}_0 = 0$. Thus a nonzero G implies instability at every finite balancing strength. If $\mathsf{G} = 0$, the contrast Hessian is only semidefinite and second-order information is insufficient.

If Σ is singular for linear experts, $\mathsf{H} \succ 0$ fails in the full parameter space. The theorem applies to an explicitly reduced space only if its expert curvature is positive definite.

Reduction to the binary scalar model Take $E = 2$, $f_w(U) = wU$, and $\boldsymbol{\psi}(S) = (1, S)^\top$. The binary moment assumptions give $w_0 = 0$, $\Sigma = 1$, and

$$\mathsf{G} = \begin{pmatrix} 0 \\ g \end{pmatrix}, \qquad \mathsf{C}_\psi = \mathrm{diag}(1, \chi), \qquad \mathcal{K}_q = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda\chi & -g \\ 0 & -g & q \end{pmatrix}. \tag{B.26}$$

For $\lambda > 0$, the nontrivial contrast determinant is $q\lambda\chi - g^2$. At $q = 1$, its zero recovers the selected-expert threshold $\lambda\chi = g^2$.

The binary parameters correspond to

$$\theta_+ = \frac{1}{2}\begin{pmatrix}\beta\\ \alpha\end{pmatrix}, \qquad \theta_- = -\frac{1}{2}\begin{pmatrix}\beta\\ \alpha\end{pmatrix}, \qquad u = m, \qquad v_+ = d, \quad v_- = -d. \tag{B.27}$$

The difference between the two softmax scores is $(\beta+\alpha S)/T$, as required; the factor $1/2$ is therefore necessary.

Substituting (B.27) into (B.14) gives

$$\Delta_2 \mathcal{F}_q = \frac{1}{2}m^2 + \frac{q}{2}d^2 - \frac{g}{2T}\alpha d + \frac{\lambda}{8T^2}\beta^2 + \frac{k}{8T^2}\alpha^2. \tag{B.28}$$

Since $a = \alpha/(2T) + O(\|(\alpha,\beta)\|^3)$ and $b = \beta/(2T) + O(\|(\alpha,\beta)\|^3)$, this is the quadratic part of (5). The term bmd is cubic and does not enter the homogeneous Hessian.

The original Euclidean metric in (m, d, α, β) pulls back to

$$\frac{1}{2}\sum_{j=\pm}(\mathrm{d}w_j)^2 + 2\sum_{j=\pm}\|\mathrm{d}\theta_j\|^2 = \mathrm{d}m^2 + \mathrm{d}d^2 + \mathrm{d}\alpha^2 + \mathrm{d}\beta^2. \tag{B.29}$$

This smooth positive-definite metric is covered by Theorem 2. It recovers (A.45), including every factor of T and the selected-expert time scale. A unit Euclidean metric in all w_j, θ_j would preserve the inertia criterion but change those time scales.

This completes the proof of Theorem 2. □

B.1 Degenerate cases and parameterization scope

Theorem 2 gives a local second-order characterization at the homogeneous stationary point. Global multi-expert dynamics and singular semidefinite cases require additional analysis. The parameterization uses separately parameterized experts and fixed routing features without a shared trainable encoder.

The constraint $\sum_j \theta_j = 0$ removes the common-logit gauge, but redundant routing features can create additional null directions. For example, if $h^\top \boldsymbol{\psi}(S) = 0$ almost surely, router perturbations in direction h leave the predictions unchanged. Such degeneracies remain unless the parameter space is explicitly quotiented further.

Identical singular Hessians with different nonlinear stability The following pair of models shows why a singular semidefinite Hessian does not determine stability. Take $E = 4$, $q = 0$, $\lambda = 1$, $T = 2.7$, X uniform on $\{-1, 1\}$, $Y = 1$, and $\boldsymbol{\psi} = 1$. Consider

$$f_w^{(+)}(X) = wX + w^4, \qquad f_w^{(-)}(X) = wX - w^4, \qquad w_0 = 0.$$

Both models have

$$\Sigma_J = 1, \quad \mathsf{N} = 0, \quad \mathsf{H} = 1, \quad \mathsf{H}_0 = 0, \quad \mathsf{G} = 0, \quad \mathsf{C}_\psi = 1, \qquad \mathcal{K}_0 = \mathrm{diag}(1, 0). \tag{B.30}$$

For the plus model, uniform routing with $(w_1, w_2, w_3, w_4) = (r, r, -r, -r)$ is invariant under Euclidean flow in the expert parameters and centered router parameters. The routing gradient vanishes, and

$$\mathcal{F}_0 - \mathcal{F}_0(0) = -r^4 + \frac{1}{2}r^8, \qquad \dot{r} = (1 - r^4)r^3. \tag{B.31}$$

Every sufficiently small positive r grows away from zero. Hence the homogeneous point is unstable despite its positive-semidefinite Hessian.

For the minus model, put $M_1 = \sum_j p_j w_j$ and $M_4 = \sum_j p_j w_j^4 \geq 0$. Then

$$\mathcal{F}_0 - \frac{1}{2} = \frac{1}{2}M_1^2 + M_4 + \frac{1}{2}M_4^2 + \frac{\lambda E}{2}\sum_j (p_j - 1/E)^2. \tag{B.32}$$

For finite logits, equality requires all $w_j = 0$ and uniform routing. The expert stationary equations are $M_1 + 4(1 + M_4)w_j^3 = 0$ for all j. They force equal expert weights, and substitution then makes the common weight zero. The remaining balancing stationary equations force uniform probabilities and, under the zero-sum constraint, zero router parameters. The homogeneous point is an isolated strict local minimum. A local invariant-sublevel-set argument gives local asymptotic stability, but the singular Hessian precludes exponential stability.

Thus identical quadratic data can correspond to instability or nonexponential asymptotic stability; second-order information alone does not determine nonlinear stability.

Pure averaging need not be unstable for nonlinear experts A second example uses independent Rademacher variables U_1, U_2, with $S = U_1U_2$, $Y = -U_2$, and $f_w(U) = wU_1 + w^2U_2$. At $w_0 = 0$, with feature $\psi(S) = S$ and single-cue batches,

$$\Sigma_J = 1, \qquad \mathsf{N} = -2, \qquad \mathsf{H} = 3, \qquad \mathsf{H}_0 = 2, \qquad \mathsf{G} = -1, \qquad \mathsf{C}_\psi = 1.$$

Consequently,

$$\mathcal{K}_0 = \begin{pmatrix} \lambda & 1 \\ 1 & 2 \end{pmatrix} \succ 0 \quad \text{when} \quad \lambda > \frac{1}{2}. \tag{B.33}$$

Theorem 2 gives local exponential stability under pure averaging. The router and expert inputs are correlated, and $-\mathsf{N}$ supplies positive contrast curvature. Thus the linear failure boundary proved next does not extend to every nonlinear averaged-output model.

C PURE-AVERAGING BOUNDARY

We prove Proposition 3 using the $q = 0$ objective.

Proof of Proposition 3. We derive the averaged-prediction objective, establish non-attainment, and analyze the stated invariant region without using the positive-q expert bound.

Pure averaged-output objective The average prediction is

$$\bar{f} = (m + bd + adS)U.$$

Applying (2) to the half-squared loss gives

$$\frac{1}{2}\mathbb{E}[(\bar{f} - Y)^2] - \mathbb{E}[Y^2]/2 = \frac{1}{2}(m + bd)^2 + \frac{1}{2}a^2d^2 - gad.$$

Adding the batch penalty from (4) yields

$$F_0 = \frac{1}{2}m^2 + bmd + \frac{1}{2}(a^2 + b^2)d^2 - gad + \frac{\lambda}{2}b^2 + \frac{k}{2}a^2. \tag{C.2}$$

Completing squares gives

$$F_0 + \frac{g^2}{2} = \frac{1}{2}(m + bd)^2 + \frac{1}{2}(ad - g)^2 + \frac{\lambda}{2}b^2 + \frac{k}{2}a^2. \tag{C.3}$$

Infimum, non-attainment, and optimization edge cases All terms on the right of (C.3) are nonnegative. For $0 < \delta < 1$, the finite-parameter state

$$m = b = 0, \qquad a = \delta, \qquad d = \frac{g}{\delta}$$

is feasible, with $\beta = 0$ and $\alpha = 2T \operatorname{arctanh} \delta$. Its objective is

$$F_0 = -\frac{g^2}{2} + \frac{k}{2}\delta^2 \longrightarrow -\frac{g^2}{2}. \tag{C.4}$$

Hence the lower bound is the infimum. If finite expert parameters attained it when $k > 0$, the vanishing squares would require both $a = 0$ and $ad = g$, contradicting $g > 0$. Boundary routing probabilities do not remove this contradiction when expert parameters remain finite.

The assumptions matter in the following equality cases for $\lambda > 0$:

Parameters	Infimum	Finite-parameter attainment
$g > 0, k > 0$	$-g^2/2$	Impossible
$g > 0, k = 0$	$-g^2/2$	$m = b = 0$, $ad = g$
$g = 0, k > 0$	0	$m = b = a = 0$, any finite d
$g = 0, k = 0$	0	$m = b = 0$, $ad = 0$

The $g = 0$ rows describe a limiting model outside the positive-signal assumptions of the proposition. If $\lambda = 0$, then $k = 0$ under the stated parameter relation; the minimization conditions become $m + bd = 0$ and $ad = g$, with no requirement that $b = 0$.

In the original generative model, the unshifted infimum is

$$\mathbb{E}[Y^2]/2 - \frac{g^2}{2} = \frac{v^2(1-\rho^2) + \mathbb{E}[\xi^2]}{2}.$$

The unshifted prediction risk remains nonnegative.

Strict saddle at the homogeneous equilibrium In (m, b, a, d) coordinates, differentiation of (C.2) gives

$$\nabla^2 F_0(O) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ 0 & 0 & k & -g \\ 0 & 0 & -g & 0 \end{pmatrix}. \tag{C.5}$$

The contrast block has determinant $-g^2 < 0$ and eigenvalues $(k \pm \sqrt{k^2 + 4g^2})/2$. Hence the Hessian has a negative direction for every finite k when $g > 0$.

In the original (α, d) coordinates, the block is

$$\begin{pmatrix} k/(4T^2) & -g/(2T) \\ -g/(2T) & 0 \end{pmatrix},$$

with determinant $-g^2/(4T^2) < 0$. The original-metric linearization in (a, d) is $\left(\begin{smallmatrix} -\gamma k & \gamma g \\ g & 0 \end{smallmatrix}\right)$, where $\gamma = 1/(4T^2)$, and has the positive eigenvalue

$$\frac{-\gamma k + \sqrt{\gamma^2 k^2 + 4\gamma g^2}}{2}.$$

The saddle is dynamically unstable under the specified flow. For linear experts, the second parameter derivative vanishes; nonlinear models can instead obtain contrast curvature through $-\mathsf{N}$, as illustrated in Appendix B.1.

Exact invariant-plane dynamics Let $h = m + bd$ and $e = g - ad$. The four derivatives are

$$F_m = h, \qquad F_d = bh - ae, \qquad F_a = ka - de, \qquad F_b = dh + \lambda b. \tag{C.6}$$

The full flow has $\dot{m} = -F_m$, $\dot{d} = -F_d$, and the probability mobility (A.19). On $m = b = 0$, both F_m and F_b vanish. Moreover, $x = (1 + a)/2$ and $y = (1 - a)/2$, so the two probability mobilities agree and their sum gives $\dot{b} = 0$. Local uniqueness establishes invariance of the plane.

On that plane,

$$\dot{d} = a(g - ad), \qquad \dot{a} = c(a)[d(g - ad) - ka], \qquad c(a) = \frac{(1 - a^2)^2}{4T^2}. \tag{C.7}$$

Here $\beta = 0$ and $a = \tanh(\alpha/(2T))$. Equivalently,

$$\dot{\alpha} = \frac{1 - a^2}{2T}[d(g - ad) - ka].$$

Thus (C.7) is not Euclidean gradient flow in (a, d); its router mobility is induced by the logit coordinates.

Forward invariance with finite-time strict inequalities Consider

$$\mathcal{R} = \{(a, d) : 0 < a < 1,\ d > 0,\ 0 < ad \le g\}.$$

At $a = 0, d > 0$, $\dot{a} = gd/(4T^2) > 0$. At $a = 1$, $\dot{a} = 0$. At $d = 0, a > 0$, $\dot{d} = ga > 0$. At $ad = g$ with $0 < a < 1$,

$$\frac{d}{dt}(ad) = d\dot{a} + a\dot{d} = -c(a)kg < 0.$$

Because $\dot{a} = 0$ on the boundary $a = 1$, we must also rule out finite-time arrival at that boundary.

On any interval before a possible first exit from $\mathcal{R}$, equations (C.7) imply

$$0 \le \dot{d} \le g, \qquad d_0 \le d(t) \le d_0 + gt, \qquad \dot{a} \ge -\frac{k}{4T^2}a, \qquad a(t) \ge a_0 e^{-kt/(4T^2)} > 0. \tag{C.8}$$

Also,

$$\begin{aligned} &\frac{d}{dt}\frac{1}{1 - a} = \frac{(1 + a)^2}{4T^2}[d(g - ad) - ka] \le \frac{g}{T^2}(d_0 + gt), \\ &\frac{1}{1 - a(t)} \le \frac{1}{1 - a_0} + \frac{gd_0}{T^2}t + \frac{g^2}{2T^2}t^2. \end{aligned} \tag{C.9}$$

Thus a cannot reach 1 in finite time. The lower bounds exclude finite-time arrival at $a = 0$ or $d = 0$. A first crossing from $ad < g$ to $ad = g$ would require a nonnegative product derivative, contradicting the strictly negative derivative on that boundary. If $a_0 d_0 = g$, the trajectory first enters $ad < g$ and cannot return to the boundary.

The first-exit argument therefore gives

$$0 < a(t) < 1, \qquad d(t) \ge d_0 > 0, \qquad 0 < a(t)d(t) < g \quad (t > 0). \tag{C.10}$$

In particular, d is nondecreasing and strictly increasing for $t > 0$.

Global finite-time existence The vector field (C.7) is polynomial in (a, d) and hence locally Lipschitz. If the maximal existence time were finite, say τ, (C.8) would confine the trajectory to the compact set $[0, 1]\times[d_0, d_0+g\tau]$, where the solution could be continued. Therefore expert parameters cannot diverge in finite time.

Equation (C.9) also keeps a away from 1 on every finite time interval, so $\alpha = 2T \operatorname{arctanh} a$ remains finite. The solution is global in the finite-logit parameterization, without using the positive-q disk bound.

Expert divergence Assume for contradiction that d is bounded. Its monotonicity gives a finite limit $d_{\max} \geq d_0 > 0$. The trajectory then lies in the compact set

$$\mathcal{K} = \{(a, d) : 0 \leq a \leq 1,\ d_0 \leq d \leq d_{\max},\ ad \leq g\}.$$

This compactness statement concerns probability coordinates and does not assume bounded logits.

Use the planar Lyapunov function

$$V(a, d) = \frac{1}{2}(ad - g)^2 + \frac{k}{2}a^2. \tag{C.11}$$

Its exact dissipation is

$$\dot{V} = -a^2(g - ad)^2 - c(a)[d(g - ad) - ka]^2. \tag{C.12}$$

Because $V \geq 0$, the nonnegative dissipation is integrable. Its time derivative is bounded on $\mathcal{K}$, so it is uniformly continuous and tends to zero by the argument used for (A.44). Every omega-limit point is stationary for the compactified planar flow, including at the zero-mobility boundary $a = 1$.

The stationary points in $0 \leq a \leq 1, d \geq 0$ are exactly

$$(a, d) = (0, 0) \quad \text{and} \quad (a, d) = (1, g). \tag{C.13}$$

Indeed, at $a = 0$, $\dot{a} = gd/(4T^2)$ forces $d = 0$. In $0 < a < 1$, $\dot{d} = 0$ forces $ad = g$, but then $\dot{a} = -c(a)ka < 0$. At $a = 1$, $\dot{a} = 0$ and $\dot{d} = g - d$, forcing $d = g$.

On the full compactified plane, the only additional stationary point is the sign-reflected point $(-1, -g)$, outside the positive invariant region. The bound $d \geq d_0 > 0$ excludes $(0, 0)$. If $(1, g)$ were not in $\mathcal{K}$, no omega-limit point would be possible. Otherwise the nonempty omega-limit set is $\{(1, g)\}$, and compactness gives $a(t) \to 1$, $d(t) \to g$.

At this putative limit, $F_a = ka - d(g - ad) \to k > 0$. For all sufficiently large times, (C.7) gives $\dot{a} < 0$ because $a < 1$ at every finite time. Hence $a(t) \leq a(t_0) < 1$ for a fixed late t_0, contradicting $a(t) \to 1$.

Thus d is unbounded; monotonicity gives $d(t) \to +\infty$.

Router homogenization From (C.10),

$$0 < a(t) \leq \frac{g}{d(t)}. \tag{C.14}$$

Expert divergence therefore gives $a(t) \to 0$, without requiring monotonic decrease of $a(t)$.

Product convergence and exclusion of repeated excursions Set $z = ad$. Directly from (C.7),

$$\dot{z} = [a^2 + c(a)d^2](g - z) - c(a)kz. \tag{C.15}$$

Fix $0 < \varepsilon < g$. Since $a(t) \to 0$, eventually $a(t) \leq 1/2$ and hence

$$c(a(t)) \geq c_* := \frac{9}{64T^2} > 0. \tag{C.16}$$

Since $d(t) \to \infty$, choose t_ε so that (C.16) holds and $d(t)^2 \geq 2kg/\varepsilon$ for every $t \geq t_\varepsilon$. Whenever $t \geq t_\varepsilon$ and $z(t) \leq g - \varepsilon$,

$$\dot{z} \geq c(a)[d^2\varepsilon - kg] \geq c_* kg =: \delta_\varepsilon > 0. \tag{C.17}$$

If $z(t_\varepsilon) < g - \varepsilon$, it reaches $g - \varepsilon$ within $(g - \varepsilon - z(t_\varepsilon))/\delta_\varepsilon$ additional time. The derivative remains positive on the lower boundary $z = g - \varepsilon$ thereafter, so the trajectory cannot cross downward. The upper bound $z \leq g$ is already invariant. Therefore

$$g - \varepsilon \leq z(t) \leq g \quad \text{for all sufficiently large } t. \tag{C.18}$$

As ε was arbitrary, $z(t) \to g$. In particular, the trajectory eventually enters and remains in every interval $[g - \varepsilon, g]$.

Interpretation and scope This establishes the three limits in (15). Along these trajectories, $m = b = 0$ at all times, so both marginal expert loads equal $1/2$. Moreover,

$$x = \frac{1+a}{2} \to \frac{1}{2}, \qquad y = \frac{1-a}{2} \to \frac{1}{2}, \qquad \beta = 0, \qquad \alpha = 2T \operatorname{arctanh} a \to 0,$$

whereas $w_+ = d \to +\infty$ and $w_- = -d \to -\infty$. Uniform limiting routing therefore coexists with diverging expert parameters.

The averaged prediction nevertheless satisfies

$$\mathbb{E}[(\bar{f}(t) - gSU)^2] = (a(t)d(t) - g)^2 \to 0, \qquad F_0(t) + \frac{g^2}{2} = \frac{1}{2}(a(t)d(t) - g)^2 + \frac{k}{2}a(t)^2 \to 0. \tag{C.19}$$

Under the teacher model, gSU is the conditional mean. Under the weaker moment assumptions, it is the cue-conditional linear prediction but need not be the full conditional mean.

For fixed $0 < \eta < 1$, the transformation on this plane

$$a \mapsto \eta a, \qquad d \mapsto d/\eta \tag{C.20}$$

preserves ad and thus the averaged-prediction signal, while multiplying the contrast-balancing cost by η^2. The pure averaged-output objective has no expert-separation penalty to prevent this compensation in the linear model.

The specified invariant set is two-dimensional and has four-dimensional Lebesgue measure zero, so Proposition 3 establishes escape only on this set rather than on a full-dimensional basin. For example, $a = b = d = 0$ with finite m_0 has $m(t) = m_0 e^{-t}$ and converges to the homogeneous point.

The finite-logit assumption also matters. At the compactified boundary initialization $a_0 = 1$, the solution is $a(t) = 1$ and $d(t) = g + (d_0 - g)e^{-t}$, which is bounded. When $k = 0$, finite minimizers exist, so the non-attainment and escape conclusions require positive contrast penalization.

This completes the proof of Proposition 3. □

D Population recovery for selected-expert training

Appendix A.13 defines the sharper selected-expert constants R_0 and C_{grad}.

The assumption $r < 1/4$ gives $1 - 2(1+\chi)r > 0$. Set $\varphi(p) := -\log(1-p) + (1-p)^{-1}$. The explicit finite-$M$ bounds are

$$\frac{T^2}{2C_{\mathrm{grad}}}\left(\delta_0^{-1} - \frac{1}{r}\right) \le \tau_r \le \frac{T^2}{2\lambda[1-2(1+\chi)r]\delta_0} \exp\left\{\varphi(2r) - \varphi(\delta_0) + \frac{w_+(0)^2 + w_-(0)^2}{4T^2}\right\}. \tag{36}$$

At time τ_r, at least one of x, y is at least r. Apply (26) from their common initial value δ_0 to obtain the lower bound in (36).

For the upper bound, write $L(p) = \log p + \varphi(p)$ and $W = e^{-\mathcal{H}_+/2}$. Before escape, $\varphi(x), \varphi(y) \le \varphi(2r)$, so

$$sW = \frac{x+y}{\sqrt{xy}} \exp\left\{-\frac{\varphi(x)+\varphi(y)}{2} + \frac{w_+^2 + w_-^2}{4T^2}\right\} \ge 2e^{-\varphi(2r)}.$$

By (29), $\dot{W} \le -(c_T\lambda[1-2(1+\chi)r]/2)sW \le -c_T\lambda[1-2(1+\chi)r]e^{-\varphi(2r)}$. If escape never occurred, this inequality would eventually force $W < 0$. Hence escape occurs in finite time, and integration up to escape gives

$$\tau_r \le \frac{e^{\varphi(2r)}W(0)}{c_T\lambda[1-2(1+\chi)r]}, \qquad W(0) = \delta_0^{-1}e^{-\varphi(\delta_0)+(w_+(0)^2+w_-(0)^2)/(4T^2)}.$$

Substitution of the initial barrier value gives the stated upper bound. Since $\delta_0^{-1} = 1 + e^{M/T}$ and $\varphi(\delta_0) \to 1$, the lower and upper bounds have the same asymptotic order.

On the invariant line with zero expert weights, p solves (31). For a fixed $r < 1/4$, direct separation of variables gives

$$\tau_r = \frac{T^2}{2\lambda}\int_{\delta_0}^{r} \frac{du}{u^2(1-u)^2(1-2u)} \sim \frac{T^2}{2\lambda}e^{M/T}.$$

The quadratic near-zero slowdown behind this asymptotic is illustrated in Figure 2(b). For every finite M, $p(t) \to 1/2$ as $t \to \infty$. For fixed t, the derivative bound (26) gives $1/p(t) \ge \delta_0^{-1} - c_T C_{\mathrm{grad}} t$, so $p(t) \to 0$ as $M \to \infty$. Thus the load depends on whether $t \to \infty$ or $M \to \infty$ is taken first.

E Discrete estimates for recovery when $0 < \lambda < 4gR$

We prove the one-step estimates used in Theorem 6, including the dependence between router and expert updates. The simultaneous algorithm is given in Section 4.2; primes denote values after one iteration, and Δ denotes an increment. Let $\mathcal{F}_n$ be the history before iteration n, and write $\mathbb{E}_n$ for conditional expectation given $\mathcal{F}_n$. Admissible parameters have the scales

$$r \lesssim \min\left\{1, \frac{\lambda T^2}{gR(T^2 + C_{\mathrm{upd}})}\right\},$$
$$\eta \lesssim \min\left\{K^{-2}, \frac{T^2}{C_{\mathrm{upd}}}, \frac{\lambda}{R^2K^4(B^{-1}+2r)}, \frac{\lambda T^2}{C_{\mathrm{upd}}^2(B^{-1}+2r)}\right\},$$

where the comparison constants are universal. These are sufficient conditions for the recovery bound, not phase-transition thresholds.

One explicit choice satisfying Theorem 6 is:

$$0 < r \le \min\left\{\frac{1}{16}, \frac{\lambda T^2}{8gR(T^2 + 32C_{\mathrm{upd}})}\right\}, \tag{37}$$

$$0 < \eta \le \min\left\{\frac{1}{2K^2}, \frac{T^2}{64C_{\mathrm{upd}}}, \frac{\lambda}{32R^2K^4(B^{-1}+2r)}, \frac{\lambda T^2}{4096C_{\mathrm{upd}}^2(B^{-1}+2r)}\right\}. \tag{38}$$

The first step-size bound makes every expert update a convex combination. Under these conditions, we obtain the explicit estimates

$$\mathbb{E}\left[\left(\frac{\eta\lambda}{T^2}e^{-M/T}\tau_r\right)^2\right] \le 256\exp\left\{\frac{w_+(0)^2}{2T^2}\right\}, \tag{39}$$

$$\mathbb{P}\left(\tau_r \le \left\lceil \frac{16}{\sqrt{\epsilon}}e^{w_+(0)^2/(4T^2)}\frac{T^2}{\eta\lambda}e^{M/T}\right\rceil\right) \ge 1-\epsilon. \tag{40}$$

The pathwise estimate for the no-baseline update is stated in Proposition 5 and proved below.

E.1 Bounded experts, unbiasedness, and categorical moments

If $U_i \neq 0$, condition (18) gives $|Y_i/U_i| \le D$. Each selected expert update is a convex combination of the previous weight and target slopes in $[-D, D]$, with nonnegative coefficients $2\eta I_i U_i^2/B$ (or $2\eta(1-I_i)U_i^2/B$). Their sum is at most $2\eta K^2 \le 1$. When $U_i = 0$, the corresponding terms vanish because (18) forces $Y_i = 0$. Hence both experts remain in $[-R, R]$, and $0 \le \ell_i \le 2R^2K^2$.

For a Bernoulli variable I with success probability p, $\ell_I(I-p)$ equals $\ell_+(1-p)$ with probability p and $-\ell_- p$ with probability $1-p$. Its mean and variance are

$$\mathbb{E}[\ell_I(I-p)] = p(1-p)(\ell_+ - \ell_-), \qquad \mathrm{Var}(\ell_I(I-p)) = p(1-p)[(1-p)\ell_+ + p\ell_-]^2. \tag{41}$$

The mean identity proves conditional unbiasedness of the prediction gradient. Conditional on a batch, $\mathbb{E}\widehat{p}_B = \bar{p}_B$ and $\nabla_\theta \bar{p}_B = (BT)^{-1}\sum_i p_i(1-p_i)\boldsymbol{\psi}_i$. The conditional mean of the balancing term is

$$2\lambda(2\bar{p}_B - 1)\nabla_\theta \bar{p}_B = \lambda\nabla_\theta\{(\bar{p}_B - \tfrac{1}{2})^2 + (1-\bar{p}_B - \tfrac{1}{2})^2\}.$$

Averaging over batches gives the router gradient of (5). The expert updates give $\mathbb{E}_n\Delta w_{\mathrm{r}} = \eta(-sw_{\mathrm{r}} + ga)$, where $w_{\mathrm{r}} = w_+$ and $w_{\mathrm{a}} = w_-$. For the active expert,

$$\mathbb{E}[\Delta w_{\mathrm{a}} \mid \mathcal{F}_n] = -\eta(2-s)w_{\mathrm{a}} - \eta g a. \tag{42}$$

For $c \in \{-1, 1\}$, write $p_c = p_+(c)$ and $z_c = (\beta + c\alpha)/T$. The exact router increment is

$$\Delta z_c = -\frac{2\eta}{BT^2}\sum_{i:S_i=c}\{\ell_i(I_i - p_c) + 2\lambda(2\widehat{p}_B - 1)p_c(1-p_c)\}. \tag{43}$$

Using the $q = 1$ derivatives of (5), conditional unbiasedness implies

$$\mathbb{E}_n\Delta z_+ = -\frac{2\eta}{T^2}x(1-x)h_x, \qquad \mathbb{E}_n\Delta z_- = -\frac{2\eta}{T^2}y(1-y)h_y. \tag{44}$$

Here h_x, h_y are evaluated at the current weights and probabilities. For independent examples, $\chi = 1/B$.

Conditional independence of the assignments gives

$$\mathbb{E}_n\widehat{p}_B = s/2, \qquad \mathbb{E}_n\widehat{p}_B^2 \le \mathbb{E}_n\bar{p}_B^2 + \mathbb{E}_n\bar{p}_B/B \le s^2 + s/(2B).$$

The pointwise bound $\bar{p}_B \le s$ makes these inequalities valid even when data within a fresh batch are correlated but have the same marginals. Consequently,

$$\mathbb{E}_n(\widehat{p}_B + s)^2 \le s/B + 3s^2. \tag{45}$$

With $\zeta = \eta C_{\mathrm{upd}}/T^2$, the bound on the loss and (43) give

$$|\Delta z_+| + |\Delta z_-| \le 4\zeta(\widehat{p}_B + s), \tag{46}$$

$$|\Delta w_{\mathrm{r}}| \le 4\eta R K^2\widehat{p}_B, \qquad |\Delta w_{\mathrm{a}}| \le 4\eta R K^2(1-\widehat{p}_B), \tag{47}$$

$$\mathbb{E}_n(\Delta w_{\mathrm{r}})^2 \le 16\eta^2R^2K^4 s(B^{-1} + s). \tag{48}$$

For the first inequality, the absolute prediction contribution is at most $4\eta R^2K^2(\widehat{p}_B + \bar{p}_B)/T^2$, while the balancing contribution is at most $4\eta\lambda\bar{p}_B/T^2$. Their sum is at most $4\zeta(\widehat{p}_B + s)$.

E.2 Proof of the pathwise lower bound

In (43), an example with $I_i = 1$ contributes a nonpositive prediction term to Δz_c. An example with $I_i = 0$ contributes at most $4\eta R^2 K^2 p_c/(BT^2)$, and each balancing term contributes at most $4\eta\lambda p_c/(BT^2)$. Thus

$$\Delta z_c \le \frac{4\eta C_{\mathrm{upd}}}{T^2} p_c.$$

If $\Delta z_c \le 0$, inverse odds cannot decrease. Otherwise,

$$e^{-z_c} - e^{-z_c'} = e^{-z_c}(1 - e^{-\Delta z_c}) \le e^{-z_c}\Delta z_c \le \frac{4\eta C_{\mathrm{upd}}}{T^2} e^{-z_c} p_c \le \frac{4\eta C_{\mathrm{upd}}}{T^2}.$$

Since $p^{-1} = 1 + e^{-z}$, a reciprocal probability decreases by at most $4\eta C_{\mathrm{upd}}/T^2$ in one step. At escape, at least one reciprocal is at most $1/r$; summing its decreases proves (21). If $\delta_0 \le r/2$, then $\delta_0^{-1} - 1/r \ge \delta_0^{-1}/2 \ge e^{M/T}/2$, giving the final bound in the proposition. Bounded increments also keep every logit finite after any finite number of updates.

E.3 The barrier and routing-contrast estimates

For $p = \operatorname{sigmoid} z$, the function $L(\operatorname{sigmoid} z) = z + 1 + e^z$ has derivative $1/(1-p)$ and is convex. Apply its tangent-line lower bound to the probability terms of $\mathcal{H}_0 = L(x) + L(y) - w_{\mathrm{r}}^2/(2T^2)$. Together with (44) and the exact increment of w_{r}^2, this gives

$$\mathbb{E}_n \Delta\mathcal{H}_0 \ge -\frac{2\eta}{T^2}(xh_x + yh_y) - \frac{w_{\mathrm{r}}}{T^2}\mathbb{E}_n \Delta w_{\mathrm{r}} - \frac{1}{2T^2}\mathbb{E}_n(\Delta w_{\mathrm{r}})^2.$$

Substituting $md = (w_{\mathrm{r}}^2 - w_{\mathrm{a}}^2)/4$, $d = (w_{\mathrm{r}} - w_{\mathrm{a}})/2$, and $\mathbb{E}_n \Delta w_{\mathrm{r}} = \eta(-sw_{\mathrm{r}} + ga)$ yields the conditional drift identity

$$\mathbb{E}[\Delta\mathcal{H}_0 \mid \mathcal{F}_n] \ge \frac{2\eta}{T^2}\left\{\frac{s(w_{\mathrm{r}}^2 + w_{\mathrm{a}}^2)}{4} - \frac{gaw_{\mathrm{a}}}{2} + \lambda s(1-s) - \lambda\chi a^2\right\} - \frac{\mathbb{E}[(\Delta w_{\mathrm{r}})^2 \mid \mathcal{F}_n]}{2T^2}. \tag{49}$$

Assume $s \le 1/8$, $\zeta \le 1/64$, and $\eta K^2 \le 1/2$. By (46), every logit increment has magnitude below $1/8$. Along the segment between the old and new logit values, $\operatorname{sigmoid}(z+t) \le e^{|t|}\operatorname{sigmoid}(z) \le 2p$. The first two sigmoid derivatives therefore have absolute value at most $2p$ on that segment, and the derivative of $L(\operatorname{sigmoid} z)$ is at most two. Both old and new expert weights have magnitude at most R, so

$$|\Delta\mathcal{H}_0| \le 12\zeta(\widehat{p}_B + s), \qquad \mathbb{E}_n(\Delta\mathcal{H}_0)^2 \le 512\zeta^2 s(B^{-1} + s). \tag{50}$$

The probability terms contribute at most $8\zeta(\widehat{p}_B + s)$, and the squared-weight term contributes at most $R|\Delta w_{\mathrm{r}}|/T^2 \le 4\zeta\widehat{p}_B$. Squaring and applying (45) gives the second inequality.

Next, we bound the contrast increment and its interaction with the active-expert update:

$$|\Delta a| \le 8\zeta s(\widehat{p}_B + s), \tag{51}$$

$$|\mathbb{E}_n \Delta a| \le 5\zeta s^2, \qquad |\mathbb{E}_n(\Delta a \Delta w_{\mathrm{a}})| \le 48\eta R K^2 \zeta s^2. \tag{52}$$

The mean value theorem gives $|\Delta a| \le 2x|\Delta z_+| + 2y|\Delta z_-|$, proving (51). Using $|h_x|, |h_y| \le R^2/4 + gR + \lambda \le 2C_{\mathrm{upd}}$, a Taylor expansion at the old logits gives

$$\begin{aligned}
|\mathbb{E}_n \Delta a| &\le x(1-x)|\mathbb{E}_n \Delta z_+| + y(1-y)|\mathbb{E}_n \Delta z_-| + x\mathbb{E}_n(\Delta z_+)^2 + y\mathbb{E}_n(\Delta z_-)^2 \\
&\le 4\zeta(x^2 + y^2) + 16\zeta^2 s\mathbb{E}_n(\widehat{p}_B + s)^2 \\
&\le 4\zeta s^2 + 16\zeta^2 s^2(B^{-1} + 3s) \le 5\zeta s^2.
\end{aligned}$$

For the product of increments, apply the pointwise bounds before taking expectations:

$$|\mathbb{E}_n(\Delta a \Delta w_{\mathrm{a}})| \le 32\eta R K^2 \zeta s\, \mathbb{E}_n[(\widehat{p}_B + s)(1 - \widehat{p}_B)] \le 32\eta R K^2 \zeta s(\mathbb{E}_n \widehat{p}_B + s) = 48\eta R K^2 \zeta s^2.$$

Thus, without assuming independence of the updates,

$$|w_{\mathrm{a}}\mathbb{E}_n \Delta a + \mathbb{E}_n(\Delta a \Delta w_{\mathrm{a}})| \le 32R\zeta s^2, \tag{53}$$

because $5 + 48\eta K^2 \le 29 < 32$. The conditions concern the state before each update, so the estimates also cover the step that crosses the escape boundary.

E.4 Corrected drift and two Lyapunov inequalities

The correction follows the auxiliary-potential approach used in stochastic approximation analysis (Allmeier & Gast, 2024). Its coefficients account for the coupled router–expert update. For $\mathcal{J} = \mathcal{H}_0 - gaw_{\mathrm{a}}/(2T^2)$ and $W = e^{-\mathcal{J}/2}$, the product expansion gives the identity

$$\mathbb{E}_n \Delta\mathcal{J} = \mathbb{E}_n \Delta\mathcal{H}_0 - \frac{g}{2T^2}\{a\mathbb{E}_n \Delta w_{\mathrm{a}} + w_{\mathrm{a}}\mathbb{E}_n \Delta a + \mathbb{E}_n(\Delta a \Delta w_{\mathrm{a}})\}.$$

Combining (49), (42), and (53) cancels the order-s term involving the active expert and gives

$$\begin{aligned}\mathbb{E}_n \Delta\mathcal{J} \geq \frac{\eta}{T^2}\left\{\frac{s(w_{\mathrm{r}}^2 + w_{\mathrm{a}}^2)}{2} - \frac{gsaw_{\mathrm{a}}}{2} + \frac{g^2a^2}{2} + 2\lambda s(1-s) - 2\lambda\chi a^2\right\} \\ - \frac{\mathbb{E}_n(\Delta w_{\mathrm{r}})^2}{2T^2} - \frac{16gR\zeta}{T^2}s^2.\end{aligned} \tag{54}$$

Apply $|a| \leq s$, $\chi \leq 1$, and (48), then drop the nonnegative square terms to obtain

$$\mathbb{E}_n \Delta\mathcal{J} \geq \frac{\eta s}{T^2}\left\{2\lambda - 4\lambda s - \frac{gR}{2}\left(1 + \frac{32C_{\mathrm{upd}}}{T^2}\right)s - 8\eta R^2K^4(B^{-1} + s)\right\}. \tag{55}$$

Under (37)–(38), the three subtracted terms in braces are at most $\lambda/2$, $\lambda/8$, and $\lambda/4$, respectively. Hence $\mathbb{E}_n \Delta\mathcal{J} \geq \mu s$, where $\mu = \eta\lambda/T^2$.

The recovery-threshold condition also gives

$$\frac{2gRr}{T^2} \leq \frac{\lambda}{4(T^2 + 32C_{\mathrm{upd}})} \leq \frac{1}{128}. \tag{56}$$

The identity $a'w'_{\mathrm{a}} - aw_{\mathrm{a}} = a\Delta w_{\mathrm{a}} + w'_{\mathrm{a}}\Delta a$, the bound $|w'_{\mathrm{a}}| \leq R$, and the increment estimates give

$$\begin{aligned}|\Delta(\mathcal{J} - \mathcal{H}_0)| &\leq \frac{g}{2T^2}\{4\eta RK^2 s + 8R\zeta s(\widehat{p}_B + s)\} \\ &\leq 2\zeta s + 4(2gRr/T^2)\zeta(\widehat{p}_B + s) \leq 6\zeta(\widehat{p}_B + s).\end{aligned}$$

The second line uses $gRK^2 \leq C_{\mathrm{upd}}$. Together with (50), it gives $|\Delta\mathcal{J}| \leq 20\zeta(\widehat{p}_B + s) < 1/2$ and

$$\mathbb{E}_n(\Delta\mathcal{J})^2 \leq 400\zeta^2(s/B + 3s^2) \leq 2048\zeta^2 s(B^{-1} + s).$$

Combining these bounds with the corrected drift gives

$$\mathbb{E}_n \Delta\mathcal{J} \geq \mu s, \qquad |\Delta\mathcal{J}| \leq 20\zeta(\widehat{p}_B + s) < \tfrac{1}{2}, \qquad \mathbb{E}_n(\Delta\mathcal{J})^2 \leq 2048\zeta^2 s(B^{-1} + s). \tag{57}$$

For the exponential transforms, (38) ensures $\eta C_{\mathrm{upd}}^2(B^{-1} + 2r)/(\lambda T^2) \leq 1/4096$. For $|h| \leq 1/2$, the inequalities

$$e^{-h/2} \leq 1 - h/2 + h^2/4, \qquad e^{-h} \leq 1 - h + h^2$$

combined with (57) give

$$\begin{aligned}\mathbb{E}_n \Delta W &\leq W(-\mu s/2 + 512\zeta^2 s(B^{-1} + s)) \leq -\mu s W/4, \\ \mathbb{E}_n \Delta(W^2) &\leq W^2(-\mu s + 2048\zeta^2 s(B^{-1} + s)) \leq -\mu s W^2/2.\end{aligned}$$

Thus

$$\mathbb{E}_n \Delta W \leq -\frac{\mu}{4}sW, \qquad \mathbb{E}_n \Delta(W^2) \leq -\frac{\mu}{2}sW^2. \tag{58}$$

To obtain a uniform additive drift, note that

$$\begin{aligned}se^{-\mathcal{H}_0/2} &= \frac{x+y}{\sqrt{xy}}\sqrt{(1-x)(1-y)}\exp\left\{-\frac{1}{2(1-x)} - \frac{1}{2(1-y)} + \frac{w_{\mathrm{r}}^2}{4T^2}\right\} \\ &\geq 2(1-2r)e^{-1/(1-2r)} \geq \frac{1}{2}.\end{aligned}$$

Moreover, $W = e^{-\mathcal{H}_0/2}e^{gaw_{\mathrm{a}}/(4T^2)}$, and (56) bounds the second factor below by $e^{-1/4} > 1/2$. Hence $sW \geq 1/4$, giving

$$\mathbb{E}_n \Delta W \leq -\mu/16, \qquad \mathbb{E}_n \Delta(W^2) \leq -\mu W/8. \tag{59}$$

E.5 Stopping, restarting, and the second moment

For each finite N, sum the conditional drift inequalities over states $n < \tau_r \wedge N$. This includes the increment on the final step. Since W and W^2 are nonnegative,

$$\frac{\mu}{16}\mathbb{E}(\tau_r \wedge N) \le W_0, \qquad \frac{\mu}{8}\mathbb{E}\sum_{n<\tau_r\wedge N} W_n \le W_0^2.$$

Monotone convergence as $N \to \infty$ gives

$$\mathbb{E}\tau_r \le \frac{16W_0}{\mu}, \qquad \mathbb{E}\sum_{n<\tau_r} W_n \le \frac{8W_0^2}{\mu}. \tag{60}$$

In particular, $\tau_r < \infty$ almost surely.

The same argument applies conditionally at any deterministic n and gives

$$\mathbf{1}_{\{n<\tau_r\}}\mathbb{E}[\tau_r - n \mid \mathcal{F}_n] \le \mathbf{1}_{\{n<\tau_r\}}\frac{16W_n}{\mu}.$$

The conditions depend only on the state before escape and remain valid after restarting the argument. For integer τ_r, the identity $\tau_r^2 = \sum_{n<\tau_r}(2(\tau_r - n) - 1)$ and interchange of the nonnegative sum and expectation give

$$\mathbb{E}\tau_r^2 \le \frac{32}{\mu}\mathbb{E}\sum_{n<\tau_r} W_n \le \frac{256W_0^2}{\mu^2}. \tag{61}$$

Initially $x = y = \delta_0$ and $a = 0$, so $L(\delta_0) = -M/T + 1 + e^{-M/T}$. Thus $W_0 \le \exp\{M/T + w_{\mathrm{r}}(0)^2/(4T^2)\}$. Substitution proves (22); the Markov inequality applied to τ_r^2 gives (23). Appendix H.2 gives a numerical setting satisfying these sufficient conditions and tests the resulting recovery scaling.

E.6 Active-expert excursions before recovery

The stochastic barrier cannot reuse the squared-weight correction for the active expert from the population argument. We construct a bounded teacher under which the weight of the active expert exceeds any fixed threshold in $(0, v)$ before escape with probability tending to one as $M \to \infty$. This occurs for every fixed step size $0 < \eta < 1/2$. The squared stochastic increment also creates an adverse drift that persists at the starved corner, motivating (24).

Let U be uniform on $\{-1, 1\}$, take $\xi = 0$ and $D = R = v$, and initialize the expert weights in $[-v, v]$. Fix a recovery threshold $0 < r < 1/4$ and a weight threshold $u_c \in (0, v)$. Use independent batches and the algorithm of Section 4.2. Consider a block of L iterations in which every example has $Z = 1$ and every assignment goes to the active expert w_-. On this event,

$$w_-(n + L) = v + (w_-(n) - v)(1 - 2\eta)^L \ge v - 2v(1 - 2\eta)^L.$$

Choose L so that the lower bound exceeds u_c. Before escape, each example is assigned to the active expert with probability at least $1 - 2r$, regardless of its cue, and its latent task equals 1 with probability $1/2$. Thus a prescribed block has conditional probability at least

$$p_* = [(1 - 2r)/2]^{BL} > 0,$$

provided that every update in the block starts before escape.

Proposition 5 guarantees at least

$$N_M = \max\left\{0, \left\lfloor \frac{T^2}{4\eta C_{\mathrm{upd}}}(\delta_0^{-1} - 1/r)\right\rfloor - 1\right\}$$

iterations before escape on every path. Divide this interval into disjoint blocks of length L. Conditional on the preceding history, each block has probability at least p_* of producing the stated event. Iterated conditioning gives

$$\mathbb{P}\big(w_-(n) \le u_c \text{ for every } n < \tau_r\big) \le (1 - p_*)^{\lfloor N_M/L\rfloor} \longrightarrow 0 \quad \text{as } M \to \infty.$$

Subtracting $w_-^2/(2T^2)$, as in the population barrier, would introduce a noise term that persists at the corner. At $s = 0$ with $w_- = 0$, the active-expert update is $\Delta w_- = 2\eta v B^{-1} \sum_i Z_i$, so

$$\mathbb{E}(\Delta w_-)^2 = 4\eta^2 v^2/B.$$

Subtracting $w_-^2/(2T^2)$ from the barrier therefore adds a negative conditional drift $-2\eta^2 v^2/(BT^2)$ at the corner, whereas the balancing drift vanishes with s. The correction $-gaw_-/(2T^2)$ vanishes at the corner and leaves a remainder of order s^2, as shown in Appendix E.4.

F A RECOVERY BOUND WHEN $\lambda \geq 4gR$

When $\lambda \geq 4gR$, the active-expert bound controls its contribution to (49). An exponential supermartingale then bounds the escape time. The proof permits target marginal loads up to $1/16$.

Theorem 7. *Use the model, bounded-data condition, algorithm, and saturated initialization of Section 4.2. Suppose $\lambda \geq 4gR$. Fix $B \geq 1$, $0 < r \leq 1/16$, and $0 < \epsilon < 1/2$. There is a step-size threshold $\eta_0 > 0$, independent of M, such that, whenever $0 < \eta \leq \eta_0$ and $\delta_0 < r$,*

$$\mathbb{P}\left(\tau_r \leq \left\lceil C\frac{T^2}{\eta\lambda}\exp\left\{\frac{M}{T} + \frac{w_+(0)^2}{4T^2}\right\}\right\rceil\right) \geq 1 - \epsilon, \tag{62}$$

where C is universal. The conclusion also holds when batches are independent across time but examples within a batch are dependent, provided that they have the same teacher marginals and the categorical assignments are conditionally independent. The proof gives explicit sufficient conditions on η.

Proof. For explicit sufficient conditions, set

$$\zeta = \frac{\eta C_{\mathrm{upd}}}{T^2}, \qquad \nu_{\mathrm{var}} = \frac{\eta C_{\mathrm{upd}}^2}{\lambda T^2}(B^{-1} + 2r), \qquad \Gamma = 8192\nu_{\mathrm{var}} + 28\zeta.$$

It suffices to take

$$\eta \leq \frac{1}{2K^2}, \qquad \zeta \leq \frac{1}{64}, \qquad \eta R^2 K^4(B^{-1} + 2r) \leq \frac{\lambda}{32}, \qquad \Gamma \log(4/\epsilon) \leq \frac{1}{2}. \tag{63}$$

All four conditions hold for sufficiently small η, independently of M. In (62), one may take $C = 32$. Use $\mathcal{H}_0 = L(x) + L(y) - w_{\mathrm{r}}^2/(2T^2)$ and $W = e^{-\mathcal{H}_0/2}$, without the contrast correction. The estimates from Appendix E.1 through (50) remain valid with within-batch dependence; the population coefficient is $\chi = \mathbb{E}\bar{S}^2 \leq 1$. Before escape, $|w_{\mathrm{a}}| \leq R \leq \lambda/(4g)$ and $|a| \leq s$, so

$$\lambda\{1 - 2(1 + \chi)r\} - g|w_{\mathrm{a}}|/2 \geq 5\lambda/8.$$

By (49), (48), and (63),

$$\mathbb{E}_n \Delta\mathcal{H}_0 \geq \mu s, \qquad \mu = \eta\lambda/T^2.$$

The last condition in (63) implies $\nu_{\mathrm{var}} \leq 1/512$. Apply $e^{-h/2} \leq 1 - h/2 + h^2/4$ to (50) to obtain

$$\mathbb{E}_n \Delta W \leq -D_n, \qquad D_n = \frac{\mu}{4} W_n s_n, \tag{64}$$

$$\mathbb{E}_n(\Delta W)^2 \leq 512\zeta^2 W_n^2 s_n(B^{-1} + s_n), \qquad |\Delta W| \leq 12\zeta W_n(\widehat{p}_B + s_n). \tag{65}$$

The last two estimates use $|e^{-h/2} - 1| \leq |h|$ for $|h| \leq 1/2$. Applying the comparison in Appendix E.4 without the correction gives $sW \geq 1/2$ and hence $D_n \geq \mu/8$.

Let $\overline{W} = \exp\{M/T + w_{\mathrm{r}}(0)^2/(4T^2)\} \geq W_0$. We stop when the process escapes, when $W \geq 2\overline{W}$, or at a deterministic horizon N, and denote this stopping time by σ. At every state before stopping, (65) implies

$$|\Delta W| \leq 28\zeta\overline{W}, \qquad \mathbb{E}_n(\Delta W)^2 \leq 4096\nu_{\mathrm{var}}\overline{W} D_n.$$

With $\vartheta = (\Gamma\overline{W})^{-1}$, we have $\vartheta|\Delta W| \leq 1$ and $4096\vartheta\nu_{\mathrm{var}}\overline{W} \leq 1/2$. Using $e^z \leq 1 + z + z^2$ for $|z| \leq 1$ yields

$$\mathbb{E}_n e^{\vartheta\Delta W} \leq 1 - \vartheta D_n/2 \leq e^{-\vartheta D_n/2}.$$

It follows that

$$\exp\left\{\vartheta\left(W_{n\wedge\sigma} - W_0 + \tfrac{1}{2}\sum_{j<n\wedge\sigma} D_j\right)\right\}$$

is a nonnegative supermartingale with initial value one. Stopping at its first crossing of $2/\epsilon$ shows that, with probability at least $1-\epsilon/2$, simultaneously for all n,

$$W_{n\wedge\sigma} + \tfrac{1}{2}\sum_{j<n\wedge\sigma} D_j \le W_0 + \Gamma\overline{W}\log(2/\epsilon) \le \tfrac{3}{2}\overline{W}. \tag{66}$$

This event includes the last update and excludes stopping at $W \ge 2\overline{W}$. Choose $N = \lceil 32\overline{W}/\mu\rceil$. Without escape by N, the sum in (66) would be at least $N\mu/16 \ge 2\overline{W}$, a contradiction. The resulting probability is at least $1-\epsilon$. □

Corollary 1. *Suppose $\lambda \ge 4R^2K^2$, $|w_+(0)| \le T$, $r = 1/16$, and $0 < \epsilon < 1/2$. There are universal constants $c, C, M_0 > 0$ such that, if $M/T \ge M_0$ and*

$$0 < \eta \le c\min\left\{K^{-2}, \frac{T^2}{\lambda\log(4/\epsilon)}\right\},$$

then, on every sample path,

$$\tau_{1/16} \ge c\frac{T^2}{\eta\lambda}e^{M/T},$$

and, with probability at least $1-\epsilon$,

$$\tau_{1/16} \le 1 + C\frac{T^2}{\eta\lambda}e^{M/T}.$$

Proof. Here $C_{\mathrm{upd}} \le 5\lambda/4$, $B^{-1} + 2r \le 9/8$, and $\lambda \ge 4gR$. Thus a sufficiently small universal c in the stated step-size bound satisfies every condition in (63); in particular, $\Gamma \le C\eta\lambda/T^2$ for a universal constant C. Taking a sufficiently large universal M_0 gives $\delta_0 \le r/2$. Proposition 5 then supplies the lower bound, and Theorem 7 supplies the upper bound after absorbing $e^{w_+(0)^2/(4T^2)} \le e^{1/4}$ into a universal C. □

G Permanent starvation under projected routing

The no-starvation conclusion depends on routing geometry as well as the prediction objective. Replacing softmax routing with Euclidean projected routing can create permanently starved vertices. The comparison with Section 4.1 keeps the objective and expert updates fixed and projects the routing coordinates. Let $\mathcal{D} = \{(a,b) : |a| + |b| \le 1\}$. The projected router solves

$$(\dot{a}, \dot{b}) = \Pi_{T_{\mathcal{D}}(a,b)}(-F_a, -F_b),$$

where $T_{\mathcal{D}}(a,b)$ is the tangent cone and Π is Euclidean projection. The expert coordinates still obey (25). At the vertex $(a,b) = (0,-1)$, the $+$ expert has zero load.

Proposition 8. *The state $a = 0, b = -1, w_+ = u, w_- = 0$ is stationary under the projected flow if and only if*

$$\lambda \le u^2/4 - g|u|/2. \tag{67}$$

If $u > 2g$ and the inequality is strict, an open set of initial states with interior routing reaches the starved vertex in finite time. Thereafter, the unused expert remains fixed and the active expert tends to zero.

Proof. At the stated point, the expert derivatives vanish and

$$F_a = -gu/2, \qquad F_b = u^2/4 - \lambda.$$

The tangent cone consists of directions (v_a, v_b) satisfying $v_b \ge |v_a|$. The projection of $-\nabla F$ onto this cone vanishes exactly when $F_b \ge |F_a|$, which is (67).

For the open-set claim, strict inequality and continuity give a neighborhood on which

$$F_b - |F_a| \geq h > 0$$

for some h. Put $s = b + 1$. In the interior, $\dot{s} = -F_b \leq -h$. On the adjacent edge $a = s$, projection onto the boundary direction gives either the unprojected inward field or $\dot{s} = (-F_b - F_a)/2 \leq -h/2$. On $a = -s$, the analogous bound is $\dot{s} = (-F_b + F_a)/2 \leq -h/2$. Thus s reaches zero within $2s(0)/h$, provided the trajectory stays in the neighborhood. Since $|a| \leq s$, decreasing s also keeps the router near the vertex.

The expert vector field is bounded on a slightly larger compact neighborhood. Choose the initial router sufficiently close to the vertex and the initial experts in a smaller open neighborhood. The bound on hitting time is then shorter than the time needed for the expert coordinates to leave the larger neighborhood. The router reaches the vertex while $F_b > |F_a|$ still holds. At the vertex, $\dot{w}_+ = 0$ and $\dot{w}_- = -2w_-$. By uniform continuity near $(u, 0)$, the neighborhood can be chosen so that this inequality holds for every active weight between its value at the hitting time and zero. The router therefore remains at the vertex, the unused weight stays fixed, and the active weight decays to zero. These initial states form an open set with interior routing. □

The comparison uses the same selected-expert objective under two router parameterizations. The stationary inequality depends on the weight of the unused expert. Under softmax routing, Theorem 1 excludes asymptotic starvation from every finite initialization, regardless of the initial expert magnitudes.

H NUMERICAL VALIDATION OF SPECIALIZATION AND RECOVERY

The population transition experiment illustrates the three-regime prediction-dispersion phase diagram at $q = 0.5$. The population recovery experiments use the $q = 1$ selected-expert objective and illustrate the scale in Theorem 4. The sampled experiment examines recovery under the conditions of Theorem 6. Appendix I examines qualitative behavior in multi-expert models.

H.1 POPULATION TRANSITION AND RECOVERY

Transition. We integrate the joint flow of the $q = 0.5$ objective across both boundaries $k/g^2 = q$ and $k/g^2 = 1/q$, where $k = \lambda\chi$. Figure 4(a) reports the routing contrast $|a|$ and normalized expert separation $|d|/g$ at the common finite horizon $t = 10^4$, using 16 paired random initializations per setting. At this horizon, both quantities are near their full-specialization values below the first boundary, remain nonzero between the boundaries, and are near zero above the second. This finite-horizon pattern matches the three regimes of Theorem 1 but is not an asymptotic convergence test.

Recovery versus initial routing probability. At $T = 1$, we initialize $\alpha_0 = 0$, $\beta_0 = -M$ and compare initial expert weights $(1, 0)$ and $(0, 0)$ at target marginal load $r = 0.025$. Figure 4(b) plots first-passage time against inverse initial routing probability. On the zero-weight invariant line, the exact integral in Appendix D gives $\tau_r \sim (2\lambda\delta_0)^{-1}$ at $T = 1$. Joint training has the same $e^{M/T}$ dependence with a different prefactor.

Temperature scaling. We next compare jointly trained experts $(0.6, -0.3)$ at $T \in \{0.5, 1, 2\}$ and target load $r = 0.1$. Figure 4(c) shows an approximately linear relation between $\log \tau_r$ and M/T. Over the final four tested points, all three fitted slopes are within 0.02 of one. The invariant-line integral in Appendix D gives the asymptote $T^2 e^{M/T}/(2\lambda)$.

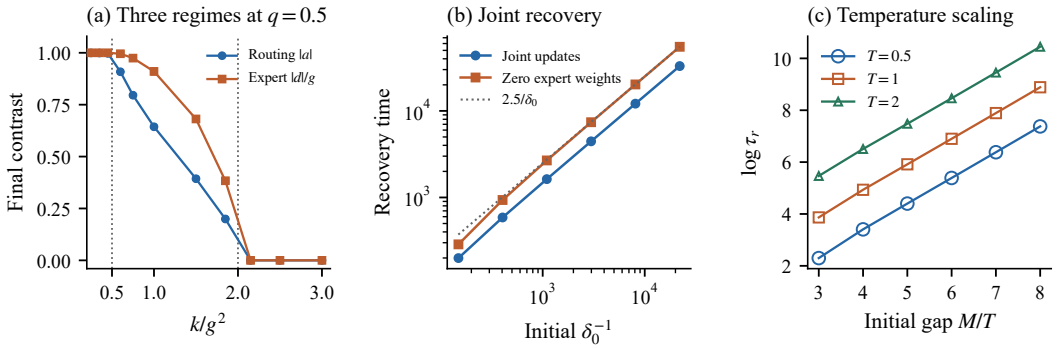


Figure 4: Population transition and recovery. (a) Routing contrast and normalized expert separation at $q = 0.5$ and the common finite horizon $t = 10^4$ across $k/g^2 = q$ and $k/g^2 = 1/q$. (b) Recovery time versus inverse initial routing probability. (c) Joint-training recovery at three temperatures.

H.2 STOCHASTIC RECOVERY WITH JOINTLY TRAINED EXPERTS

We test the exponential recovery scale using the simultaneous categorical updates (19)–(20). The bounded teacher has $U = 1$, $Y = Z$, $v = K = R = T = 1$, $\rho = 0.8$, and $g = 0.8$. We fix $\lambda = 0.2$ and initialize $(w_+, w_-) = (1, 0)$ and $(\alpha, \beta) = (0, -M)$. Prediction and balancing updates use the same sampled assignments.

We take $B = 512$, $\eta = 0.005$, $r = 0.00075$, and $M \in \{8, 9, 10\}$, with 64 seeds for each value; these settings satisfy the sufficient conditions of Theorem 6. Figure 5 and Table 1 report first-passage times. The exponentially normalized medians $(\eta\lambda/T^2)e^{-M/T}\,\mathrm{median}(\tau_r)$ are 0.2189, 0.2747, and 0.2909 across the three values of M.

For a robustness check outside the sufficient conditions of Theorem 6, we also simulate $B = 1$, $\eta = 0.05$, $r = 0.025$, and $M \in \{5, 6, 7, 8, 9\}$ with 128 seeds per value. The exponentially normal-

ized median remains between 0.2637 and 0.2967. Additional larger-step results and active-expert trajectories are provided in the supplementary material.

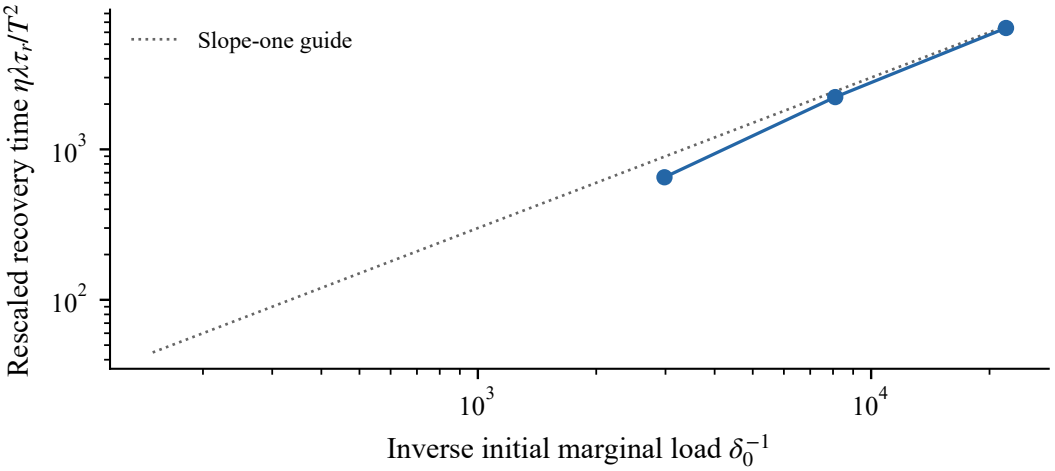


Figure 5: Recovery under categorical updates satisfying the theorem conditions. Rescaled first-passage times $\eta\lambda\tau_r/T^2$ versus inverse initial marginal load δ_0^{-1}; bars show interquartile ranges.

Table 1: Stochastic recovery in the theorem-covered regime. Medians and quartiles are measured in iterations. The exponentially normalized median is $(\eta\lambda/T^2)e^{-M/T}\,\mathrm{median}(\tau_r)$. All trajectories reached the target before the iteration cap.

M	n	Median τ_r	Interquartile range	Normalized median
8	64	652,506	[651,171.75, 653,636.75]	0.2189
9	64	2,225,650	[2,222,013.25, 2,228,515.50]	0.2747
10	64	6,407,879	[6,399,254.75, 6,414,847.25]	0.2909

I Evidence beyond two experts

The Transformer-MoE and controlled MNIST experiments change the group composition used by the balancing loss while keeping supervised batches fixed within paired comparisons. Four-task regression illustrates the distinction between marginal load balance and specialization with nonlinear experts. A separate batch-construction sweep includes MNIST models with four and eight experts.

I.1 Four-task regression beyond the binary model

The two-expert theory distinguishes balanced marginal loads from distinct predictors. We examine this distinction using four nonlinear experts and four scalar regression tasks in 16 input dimensions. Let $Z \in \{1,2,3,4\}$ denote the task label. For task z, a teacher vector t_z has norm 2, the input mean μ_z has norm 0.5, $X \mid Z = z \sim \mathcal{N}(\mu_z, I_{16})$, and $Y = t_z^\top X + \varepsilon$ for independent $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. A supplied cue $\widetilde{Z} \in \{1,2,3,4\}$ equals Z with probability 0.9 and otherwise indicates one of the other three tasks uniformly. Each batch has a uniformly chosen dominant task, assigned to each example with probability 0.8; the marginal task distribution remains uniform.

The $20 \to 32 \to 4$ router receives X and a one-hot encoding of $\widetilde{Z}$. Four $16 \to 32 \to 1$ experts receive only X. Both router and experts use a $\tanh$ hidden activation. Training differentiates the expected prediction loss over all four experts and a quadratic penalty on batch-mean routing probabilities:

$$\widehat{\mathcal{L}}_B = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{4} p_{ij}\frac{(h_j(X_i)-Y_i)^2}{2} + 4\lambda_{\mathrm{quad}}\sum_{j=1}^{4}\left(\bar{p}_j - \frac{1}{4}\right)^2, \qquad \bar{p}_j = \frac{1}{B}\sum_{i=1}^{B} p_{ij}. \tag{68}$$

The routing probability p_{ij} depends on X_i and its supplied cue, and h_j is the prediction of expert j. We vary $\lambda_{\text{quad}} \in \{0, 0.1, 0.3, 0.5, 0.7, 1, 3\}$ while fixing the data and optimization procedure. The objective averages over expert choice rather than using the sampled update in Section 4.2.

On a shared evaluation set of $N_{\text{eval}} = 20{,}000$ examples, we measure dataset-level and within-batch expert-load CV from routing probabilities. For mean loads L_j, $\text{CV} = 4\sqrt{\frac{1}{4}\sum_j (L_j - \frac{1}{4})^2}$. Task-routing purity uses n_{zj}, the number of evaluation examples from task z whose most likely expert is j. Expert-output diversity compares expert predictions on the same inputs:

$$\text{TaskPurity} = \frac{1}{N_{\text{eval}}} \sum_{j=1}^{4} \max_z n_{zj}, \quad \text{OutputDiversity} = \frac{1}{6N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} \sum_{j<k} \big(h_j(X_i) - h_k(X_i)\big)^2. \tag{69}$$

We also report the expected prediction loss $N_{\text{eval}}^{-1} \sum_i \sum_j p_{ij}(h_j(X_i) - Y_i)^2/2$. Task-routing purity measures task concentration in argmax groups without requiring a one-to-one task–expert matching. Expert-output diversity does not weight predictions by routing probabilities.

Moderate balancing ($\lambda_{\text{quad}} = 0.1, 0.3$) gives nearly uniform dataset-level expert loads while retaining strong task association and distinct expert predictions. As balancing increases, within-batch load variation, task-routing purity, and expert-output diversity decrease, while dataset-level loads remain nearly uniform. From $\lambda_{\text{quad}} = 0.1$ to 3, task-routing purity falls from 0.900 to 0.276 and expert-output diversity from 5.370 to 0.576. Figure 6 and Table 2 report the five-seed comparison. Here the intervention is balancing strength, in contrast to the balancing-group intervention in Section 5.

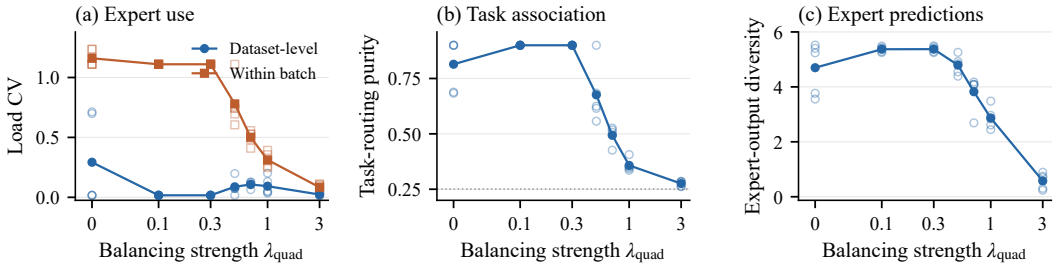


Figure 6: Four-task regression across λ_{quad}. (a) Dataset-level and within-batch expert-load CV, (b) task-routing purity, and (c) unweighted expert-output diversity. Curves show means over five seeds; open points show individual seeds.

Table 2: Four-task regression: means and sample standard deviations over five seeds. Dataset-level and within-batch expert loads use routing probabilities; both are reported as CV. Task-routing purity and expert-output diversity are defined in (69). Prediction loss averages one-half squared error over expert-selection probabilities.

λ_{quad}	Dataset-level CV	Within-batch CV	Task purity	Output diversity	Loss
0	0.293 ± 0.378	1.160 ± 0.069	0.814 ± 0.117	4.698 ± 0.954	0.501 ± 0.161
0.1	0.017 ± 0.000	1.110 ± 0.000	0.900 ± 0.000	5.370 ± 0.099	0.381 ± 0.005
0.3	0.017 ± 0.000	1.110 ± 0.000	0.900 ± 0.000	5.374 ± 0.106	0.381 ± 0.006
0.5	0.088 ± 0.067	0.780 ± 0.193	0.676 ± 0.133	4.792 ± 0.338	0.494 ± 0.074
0.7	0.107 ± 0.026	0.501 ± 0.057	0.494 ± 0.040	3.824 ± 0.635	0.660 ± 0.058
1	0.093 ± 0.072	0.312 ± 0.061	0.358 ± 0.028	2.868 ± 0.397	0.844 ± 0.036
3	0.023 ± 0.009	0.084 ± 0.030	0.276 ± 0.012	0.576 ± 0.288	1.360 ± 0.035

I.2 MNIST NEURAL MOES: SHARED SETUP AND METRICS

The neural studies use MNIST (LeCun et al., 1998), with 60,000 training and 10,000 test images. After scaling pixels to $[0, 1]$, we subtract 0.1307 and divide by 0.3081; no augmentation is used. Each

model has a learned $784 \to 64$ ReLU encoder, a linear $64 \to E$ softmax router at temperature one, and $E \in \{4, 8\}$ separate $64 \to 32 \to 10$ ReLU expert networks. The router receives the encoded image but not the digit label C. All components train jointly. Write $\mathbf{p}(X) = (p_1(X), \ldots, p_E(X))$ for the routing distribution and $h_j(X) \in \mathbb{R}^{10}$ for the expert logits. Let $J \in \{1, \ldots, E\}$ denote the selected expert index.

Table 3: The two neural dispatch and prediction-gradient rules. Here ℓ is cross entropy and $\mathrm{stopgrad}$ blocks differentiation through its argument.

	Categorical	Hard top-one
Assignment	$J \sim \mathrm{Cat}(\mathbf{p}(X))$	$J = \arg\max_j p_j(X)$
Prediction logits	$h_J(X)$	$p_J(X)h_J(X)$
Router prediction gradient	$\mathrm{stopgrad}(\ell(h_J(X), C))\nabla \log p_J(X)$	Differentiate through selected p_J with fixed index
Relation to Section 4.2	Sampled score-function form	Different gradient estimator

Under categorical routing, the selected expert and encoder also receive the ordinary prediction gradient, and no score-function baseline is used. Under hard top-one routing, the selected probability multiplies the logits, so the router receives a prediction gradient through $p_J(X)$ while the argmax index is held fixed; there is no score-function term. Both rules use the sampled auxiliary balancing loss

$$\lambda_{\mathrm{aux}} E \sum_{j=1}^{E} \mathrm{stopgrad}(\widehat{p}_j)\bar{p}_j, \qquad \widehat{p}_j := \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}\{J_i = j\}, \quad \bar{p}_j := \frac{1}{B}\sum_{i=1}^{B} p_j(X_i). \tag{70}$$

This is the objective of Fedus et al. (2022), with the two-expert normalization of (19). Treating $\widehat{p}_j$ as fixed makes the auxiliary gradient act through $\bar{p}_j$, matching the sampled-load construction of Section 4.2 when $E = 2$. The categorical and top-one assignment means nevertheless differ. Labels determine the supervised loss and constructed groups but do not enter the router.

Dispatch-load CV measures utilization balance, digit–expert mutual information (MI) measures routing association, and pairwise Jensen–Shannon divergence (JSD) measures differences between expert predictions independently of routing. We interpret JSD jointly with load, MI, and accuracy because large output differences can arise from underused experts without implying useful specialization.

Let $N_{\mathrm{eval}} = 10{,}000$, and let C_i be the label of test image i. For categorical routing, set $\pi_{ij} = p_j(X_i)$, integrating over assignment randomness; for top-one routing, set $\pi_{ij} = \mathbf{1}\{j = \arg\max_k p_k(X_i)\}$. Define $L_j = N_{\mathrm{eval}}^{-1} \sum_i \pi_{ij}$ and $Q_{cj} = N_{\mathrm{eval}}^{-1} \sum_{i:C_i=c} \pi_{ij}$. The reported dispatch metrics are

$$\mathrm{CV} = \frac{\sqrt{E^{-1}\sum_j (L_j - E^{-1})^2}}{E^{-1}}, \qquad Q_c = \sum_j Q_{cj},$$
$$\mathrm{MI} = \sum_{c,j} Q_{cj} \log \frac{Q_{cj}}{Q_c L_j}. \tag{71}$$

Terms with zero probability contribute zero. For $r_j(X) = \mathrm{softmax}(h_j(X))$, expert-output divergence is

$$\mathrm{JSD}_{\mathrm{out}} = \frac{1}{N_{\mathrm{eval}}\binom{E}{2}} \sum_{i=1}^{N_{\mathrm{eval}}} \sum_{j<k} \mathrm{JSD}\big(r_j(X_i), r_k(X_i)\big),$$
$$\mathrm{JSD}(v, w) = \frac{1}{2} D_{\mathrm{KL}}\left(v \middle\| \frac{v+w}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(w \middle\| \frac{v+w}{2}\right). \tag{72}$$

All logarithms are natural. Categorical accuracy integrates over $p(X)$ without additional assignment sampling; top-one accuracy uses the selected expert.

I.3 Controlled intervention on balancing-group composition

The first neural experiment fixes each supervised macro-batch and changes only the grouping of its examples for the auxiliary loss. Each 1,000-image macro-batch contains 100 draws with replacement from each digit, and the prediction loss uses all 1,000 images. The balancing loss averages (70)

over ten class-balanced groups (ten images per digit), ten single-class groups (100 images of one digit), ten random groups of size 100, or one global group.

For each routing rule and $\lambda_{\text{aux}} \in \{0.1, 1\}$, we run five paired seeds for 1,000 SGD updates at learning rate 0.05, without momentum or weight decay. Within each seed, conditions share initial parameters, supervised macro-batches, and categorical uniform variates; only the auxiliary-loss grouping changes. At $\lambda_{\text{aux}} = 1$ under categorical routing, the class-balanced, random, and global groupings give similar MI and JSD values. Single-class grouping reduces MI from 0.788 to 0.016 and JSD from 0.184 to 0.032 relative to class-balanced grouping. The same qualitative contrast appears under hard top-one routing. The resulting load CV also changes, so these conditions do not have matched final utilization. Figure 7 displays MI and JSD for all four groupings, and Table 4 includes load CV, accuracy, and the $\lambda_{\text{aux}} = 0.1$ conditions.

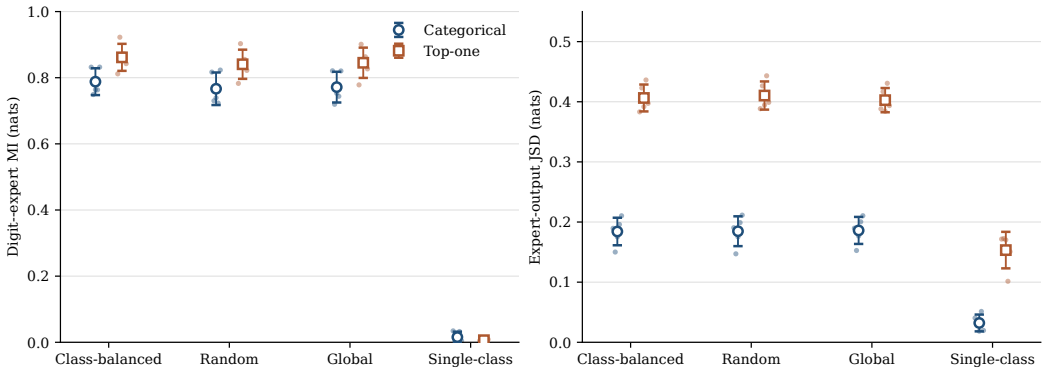


Figure 7: **Four-expert balancing-group intervention at $\lambda_{\text{aux}} = 1$.** Digit–expert mutual information (a) and unweighted expert-output JSD (b); standard deviations over five seeds.

Table 4: Controlled four-expert MNIST grouping experiment: mean ± sample standard deviation over five paired seeds. Accuracy is in percent, CV is dimensionless, and MI and JSD are in nats. Appendix I.2 defines the dispatch metrics and unweighted expert-output JSD.

Routing	λ_{aux}	Balance group	Accuracy (%)	Load CV	MI	JSD
Categorical	0.1	Class-balanced	95.10 ± 0.38	1.358 ± 0.445	0.262 ± 0.277	0.329 ± 0.036
Categorical	0.1	Random	95.12 ± 0.34	1.361 ± 0.439	0.260 ± 0.277	0.330 ± 0.035
Categorical	0.1	Global	95.12 ± 0.36	1.361 ± 0.441	0.262 ± 0.277	0.330 ± 0.037
Categorical	0.1	Single-class	95.10 ± 0.33	1.458 ± 0.373	0.121 ± 0.133	0.327 ± 0.032
Categorical	1	Class-balanced	94.26 ± 0.29	0.042 ± 0.005	0.788 ± 0.041	0.184 ± 0.023
Categorical	1	Random	94.22 ± 0.28	0.044 ± 0.013	0.767 ± 0.049	0.185 ± 0.025
Categorical	1	Global	94.27 ± 0.29	0.036 ± 0.014	0.772 ± 0.046	0.186 ± 0.022
Categorical	1	Single-class	93.56 ± 0.34	0.159 ± 0.063	0.016 ± 0.016	0.032 ± 0.014
Top-one	0.1	Class-balanced	93.55 ± 0.66	0.546 ± 0.153	0.725 ± 0.117	0.398 ± 0.086
Top-one	0.1	Random	93.63 ± 0.59	0.512 ± 0.115	0.752 ± 0.102	0.417 ± 0.059
Top-one	0.1	Global	93.87 ± 0.38	0.554 ± 0.160	0.705 ± 0.161	0.419 ± 0.052
Top-one	0.1	Single-class	93.01 ± 0.69	0.675 ± 0.245	0.252 ± 0.074	0.359 ± 0.045
Top-one	1	Class-balanced	93.90 ± 0.18	0.043 ± 0.027	0.862 ± 0.041	0.406 ± 0.022
Top-one	1	Random	93.87 ± 0.22	0.040 ± 0.021	0.841 ± 0.044	0.410 ± 0.023
Top-one	1	Global	93.96 ± 0.28	0.037 ± 0.023	0.845 ± 0.046	0.403 ± 0.020
Top-one	1	Single-class	93.53 ± 0.11	0.117 ± 0.028	0.006 ± 0.002	0.153 ± 0.030

For a broader robustness check, we vary the composition of the entire training batch across four- and eight-expert models and both routing rules. This intervention changes prediction and balancing updates simultaneously, so the sweep provides qualitative robustness evidence. Additional results from this sweep are provided in the supplementary numerical results.

I.4 TRANSFORMER-MOE DOMAIN GROUPING

We train a four-layer decoder-only Transformer with width 128, four attention heads, and a four-expert MoE feed-forward block in every second layer. The router makes a hard top-one token-level assignment, and the selected expert output is multiplied by its routing probability. The model is trained for 3,000 AdamW steps at learning rate $3{\cdot}10^{-4}$ with 200 warmup steps and weight decay 0.1. We use byte-level sequences from the arXiv, GitHub, Stack Exchange, and Wikipedia categories of the SlimPajama stream, with a batch of 32 sequences of length 128 and eight sequences per domain. Domain labels determine auxiliary groups and evaluation strata but are not model inputs. Within each seed, conditions reuse the model initialization and supervised sequence-sampling stream. The prediction update and auxiliary group size are fixed in the domain-balanced versus single-domain comparison; only auxiliary group composition changes.

We use the same four balancing strengths $\lambda_{\mathrm{aux}} \in \{0, 0.002, 0.01, 0.05\}$ for global, random, single-domain, and domain-balanced groups. Positive-strength non-global groups contain eight sequences. A single-domain group contains eight sequences from one domain; a domain-balanced group contains two from each domain. The global group contains all 32 sequences. At zero strength the grouping has no effect, so we train only the global condition. The auxiliary loss uses mean routing probabilities and hard dispatch counts for each group, with the latter held fixed during differentiation. The full configuration and five-seed records are included in the supplementary code archive.

For each MoE layer, expert-load CV uses the marginal top-one dispatch fractions, and domain–expert MI uses the joint counts of domain and selected expert. We average these two metrics across the two MoE layers. Expert-output MSE averages pairwise squared differences between all expert outputs on the same sampled hidden states, independently of dispatch. Validation perplexity is computed on 20 held-out batches of byte-level sequences per seed. Reported uncertainty is the sample standard deviation across seeds 0–4.

At $\lambda_{\mathrm{aux}} = 0.05$, the same-size comparison has similar mean load CV but a larger domain–expert MI with domain-balanced groups (Table 5). Each paired MI difference favors domain-balanced groups, while paired load-CV and expert-output MSE differences have mixed signs. Validation perplexity is slightly higher for domain-balanced groups in all five pairs. Figure 8 shows that the grouping difference in MI is more visible at stronger balancing; Figure 9 illustrates the routing matrices for one seed. These observations concern routing association and the reported output-distance measure, rather than task-specific predictive ability or the formal binary-model thresholds.

Table 5: Transformer-MoE results at $\lambda_{\mathrm{aux}} = 0.05$ (five-seed mean $\pm$ sample standard deviation). Both conditions use auxiliary groups of eight sequences. Expert-load CV, domain–expert MI, and expert-output MSE are averaged over two MoE layers.

Grouping	Load CV	Domain–expert MI	Validation PPL	Output MSE
Domain-balanced	0.0515 ± 0.0168	0.0326 ± 0.0212	9.222 ± 0.239	0.03630 ± 0.00647
Single-domain	0.0500 ± 0.0112	0.00262 ± 0.00069	9.147 ± 0.193	0.03609 ± 0.00729

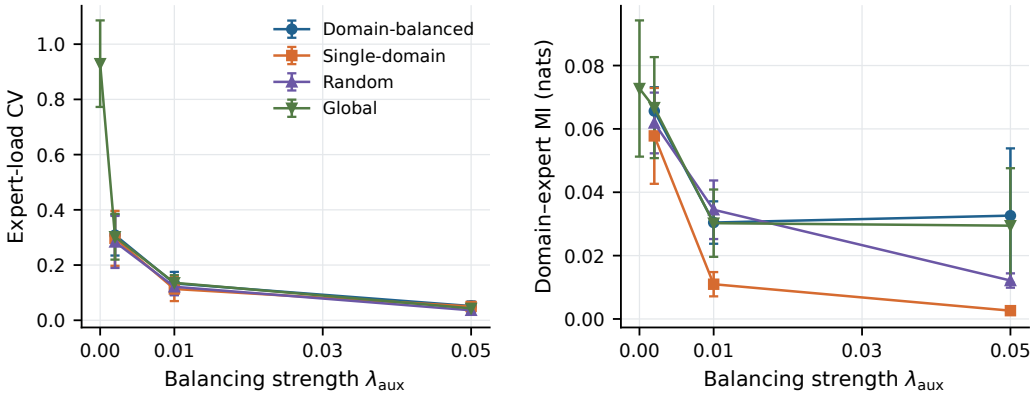


Figure 8: Balancing-strength sweep for the four auxiliary grouping rules. Points show five-seed means and bars show sample standard deviations. The single zero-strength run per seed is shown under global grouping because the auxiliary loss is inactive there.

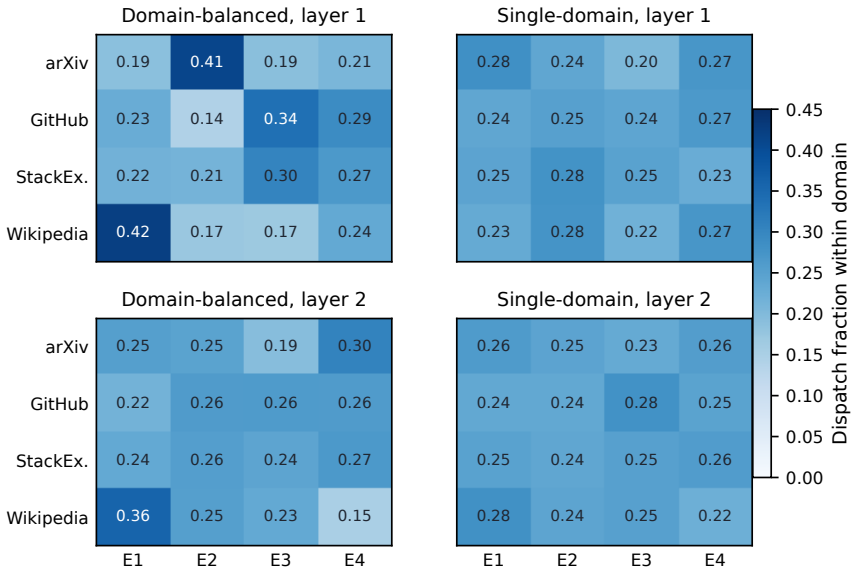


Figure 9: Illustrative domain-by-expert dispatch fractions for seed 2 at $\lambda_{\text{aux}} = 0.05$, shown separately for both MoE layers. Each row is normalized within a domain. These matrices illustrate one paired run; Figure 3 and Table 5 summarize all five seeds.

J EXPERIMENTAL SETTINGS

Appendices H–I define the experimental comparisons and metrics. We provide the corresponding numerical settings here.

J.1 POPULATION INTEGRATION

We integrate the joint flow using expert weights $w_\pm$ and router logits $z_\pm = (\beta \pm \alpha)/T$. For the transition panel, $q = 0.5$, $g = 0.8$, $T = 1$, $\chi = 1$, and $k = \lambda\chi$. The ratio grid is

$$k/g^2 \in \{0.25, 0.35, 0.45, 0.6, 0.75, 1, 1.5, 1.85, 2.15, 2.5, 3\}.$$

We pair 16 independently sampled initial states across all 11 settings; each state has four coordinates drawn uniformly from $[-1, 1]$. DOP853 integrates to $t = 10^4$ with relative and absolute tolerances of $2 \cdot 10^{-8}$ and $2 \cdot 10^{-10}$.

For fixed-temperature recovery, $T = 1$, $\lambda = 0.2$, $\chi = 1/512$, $r = 0.025$, and $M \in \{5, 6, 7, 8, 9, 10\}$. The initial router has $\alpha_0 = 0$ and $\beta_0 = -M$, and the initial experts are either $(1, 0)$ or $(0, 0)$. LSODA stops when $x + y = 2r$. For temperature scaling, $g = 0.8$, $\lambda = 0.1$, $\chi = 1$, $r = 0.1$, $T \in \{0.5, 1, 2\}$, and $M/T \in \{3, 4, 5, 6, 7, 8\}$. We compare initial experts $(0.6, -0.3)$ and $(0, 0)$ with the same router initialization. DOP853 integrates until $x + y = 2r$.

J.2 SAMPLED RECOVERY

For the bounded teacher in Appendix H.2, Z is balanced and agrees with the cue S with probability 0.9. A count-based sampler generates batches with the same update distribution as example-wise sampling; the code provides implementation details. Prediction and balancing use the same assignments and the state before each update.

The theorem-covered regime caps each trajectory at $5 \cdot 10^7$ iterations, and the larger-step regime caps each trajectory at $5 \cdot 10^6$. First-passage times include the update that crosses the target. Appendix H.2 specifies the batch sizes, step sizes, targets, and seed counts.

J.3 MULTI-EXPERT OPTIMIZATION AND EVALUATION

For four-task regression, the teacher vectors and input means are independent normalized Gaussian draws held fixed across conditions. Training batches contain 256 examples with the correlated task composition specified in Appendix I.1, and the router temperature is one. We run Adam for 2,000 updates with learning rate 0.003, decay rates $(0.9, 0.999)$, epsilon 10^{-8}, no weight decay, and global gradient-norm clipping at 5. Across all seven values of λ_{quad}, five paired seeds share initial parameters and sampled batches. The shared evaluation set contains 20,000 independent examples with uniformly distributed tasks. Within-batch CV is averaged over 200 correlated batches of size 256. Reported uncertainty is the sample standard deviation across the five seeds.

For the controlled MNIST study and the broader batch sweep in the supplementary numerical results, the encoder, router, expert hidden layer, and expert output layer use centered Gaussian initial weights with standard deviations $\sqrt{2/784}$, $0.1\sqrt{1/64}$, $\sqrt{2/64}$, and $\sqrt{1/32}$, respectively. All biases are initialized to zero. Paired seeds fix initial parameters and training examples across conditions. The random balancing partitions in the controlled experiment are also paired by seed. The controlled study and broader sweep use 1,000 and 5,000 SGD updates, respectively, at learning rate 0.05 without momentum or weight decay.