HIGH-DIMENSIONAL DYNAMICS OF MIXTURE-OF-EXPERTS: PHASE TRANSITIONS, ROUTING, AND SPECIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) architectures have been pivotal in scaling modern language models, yet the coupled evolution of their routing mechanisms and expert parameters remains underexplored. We give a high-dimensional analysis of the training dynamics of an MoE layer with E experts in the feature-learning regime, using dynamical mean-field theory (DMFT) to derive a closed system of integro-differential equations for the router and expert order parameters. The routing nonlinearity yields two differences from single-pathway DMFT: a non-Gaussian effective action, which we close by treating the routing fields as Gaussian, and a router-specific Onsager correction, needed because the router is trained on the examples it routes. We validate the equations against gradient flow at $E = 2$, 4 and 8 and show that their error grows as the routing temperature falls. For softmax routing we identify the gate with a Random Energy Model in a signal-dependent field, and use it to separate two distinct processes: a *dynamical* onset, where a symmetric saddle amplifies the initialization bias, and a *thermodynamic* condensation of the router. In gradient flow, cold routers condense onto poorly aligned experts while warm routers stay diffuse and align, and the specialization reached grows with the initialization bias, which in our longest runs is amplified by a finite gain.

1 INTRODUCTION

Mixture-of-Experts (MoE) layers are a standard primitive for scaling neural networks: many large language models released since 2020 (Fedus et al., 2022; Lepikhin et al., 2020; Du et al., 2022; Zoph et al., 2022; Jiang et al., 2024; Dai et al., 2024; Liu et al., 2024; team, 2024), and vision counterparts (Riquelme et al., 2021), conditionally activate a subset of experts per token through a learned router. What decides whether experts specialize to coherent subsets of the data or collapse onto a cluster-averaged solution is not understood from first principles. Training is also stabilized by heuristics such as auxiliary load-balancing losses, hard top-k routing (each token sent only to its k highest-scoring experts), and careful initialization, whose calibration is empirical.

Three observations about MoE training lack a common explanation. *Router collapse*: the router concentrates its weight on experts that do not reflect the input's content, and often the same few experts receive most of the load (Shazeer et al., 2017; Chi et al., 2022; Wu et al., 2024). *Temperature and initialization sensitivity*: training stability and routing quality depend on the router's initialization scale and temperature (Fedus et al., 2022; Chi et al., 2022), with no principled rule for either. *Uneven returns from load balancing*: the auxiliary losses used to prevent collapse are heuristic, applied uniformly, and known to trade off against the specialization they are meant to enable (Zhou et al., 2022; Liu et al., 2024). A theory should say which of these outcomes a given setting of routing temperature T, initialization bias ε, expert count E and sample ratio $\bar{\alpha}$ (training examples per input dimension) leads to. We find three answers. First, router collapse is the freezing transition of Derrida's Random Energy Model (REM) (Derrida, 1981): for router logits of unit variance, routing freezes below $T_c^{(0)} = 1/\sqrt{2 \ln E}$, and each input is then routed confidently to experts picked by noise. In a logit model with components shared across inputs, the load imbalance freezes at the same temperature. Second, initialization sensitivity comes from how training amplifies a small initial bias: in our longest runs the amplification stops at a finite gain, so how far experts specialize grows with the bias. Third, specialization and router condensation are separate processes. The first grows from the initialization

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.

bias, the second is set by the temperature and the signal, and experts specialize even when the router cannot condense onto them.

What makes MoE dynamics hard. The router and the experts co-evolve: each expert's gradient is gated by the probability $g_e(x)$ that the router assigns input x to expert e, and the router itself is trained on the residuals each expert produces. Two coupled subsystems learn from the *same* data, with one re-weighting the gradient signal the other sees. Standard high-dimensional learning theory (DMFT (Sompolinsky et al., 1988; Mignacco et al., 2020; Bordelon & Pehlevan, 2022; Montanari & Urbani, 2025; Bordelon & Pehlevan, 2026), and online-SGD analyses (Saad & Solla, 1995; Goldt et al., 2019)) replaces the weights by a few summary statistics, their overlaps with the teacher and with each other, and derives closed equations for how those evolve. For a two-layer network, the DMFT we build on treats each example's residual as a Gaussian process with the same mean and covariance (Montanari & Urbani, 2025), and a single Onsager-type correction, the residual response, accounts for the model being evaluated on the same examples it was trained on. Routing complicates both steps. The routing weights are a nonlinear function of the router preactivations, so the Gaussian treatment has to be extended to the routing fields. And the router is trained on the same examples whose routing it evaluates, so it routes its own training data more confidently than fresh data. The residual response corrects the experts' bias toward their training data, but not this bias in *which expert sees which example*. Without a correction for it, the DMFT predicts that each expert's alignment with the clusters it should not serve grows steadily, while gradient flow (GF) shows it rising and then falling (Figure 1).

We ask when and why these pathologies appear in a two-layer MoE trained on a Gaussian-mixture teacher, with n training examples in input dimension d, in the proportional limit $n, d \to \infty$ at fixed $\bar{\alpha} = n/d$. Prior analyses of MoE training prove guarantees for gradient-based training on specific architectures and data (Chen et al., 2022; Kawata et al., 2025). We instead derive closed equations for the coupled router and expert dynamics, and map their outcomes onto a phase diagram.

Contributions.

1. **A DMFT for jointly trained routers and experts** (Section 4). We derive closed equations for the coupled training dynamics. They extend the Gaussian treatment of the residual in single-pathway DMFT to the routing fields and add a router Onsager correction for the router's bias toward its own training examples. At $E = 2, 4, 8$ they track the expert alignment to within a couple percent error.
2. **Softmax routing as a Random Energy Model** (Section 5). Routing freezes below $T_c^{(0)} = 1/\sqrt{2 \ln E}$, and it can lock onto the correct expert only if the cluster signal-to-noise ratio exceeds $\sqrt{2 \ln(E-1)}$, a threshold that sharpens as E grows. Training amplifies a small initial bias, and in our longest runs the amplification stops at a finite gain, so how far experts specialize grows with the initial bias.
3. **Where the closure holds** (Section 4.4). At high temperature the DMFT of a softmax router reduces to that of a router linear in its logits. On the router contrast, the error of this linear-router DMFT comes mainly from linearizing the gate. With the gate means and drives evaluated exactly, the remaining error is below 1% well above $T_c^{(0)}$ and grows as the temperature falls.

2 Related Work

Theory of Mixture-of-Experts. Prior MoE theory characterizes universal approximation (Nguyen et al., 2016), parameter estimation for softmax-gated MoEs (Nguyen et al., 2024), *representation collapse* (Chi et al., 2022), and low expert activation (Wu et al., 2024). Chen et al. (2022) prove that gradient descent on a router and nonlinear experts learns the cluster structure, for a specific architecture beyond the kernel regime. Kawata et al. (2025) show that an MoE trained by SGD detects and learns a latent cluster structure of single-index tasks, whose information exponent is low within each cluster but high for the task as a whole. Two lines of work are closest to ours. Kang & Oh (1996) give an equilibrium statistical-mechanics analysis of the classical mixture of experts, in the annealed approximation, already exhibiting a symmetric phase and a specialization transition.

That analysis is static, so the question of when the dynamics *reach* such a state does not arise. More recently DMFT has been applied to MoE layers for hyperparameter transfer and scaling (Jiang et al., 2026; Vankadara et al., 2026), but those target parameterization and carry no specialization order parameter or routing phase diagram.

DMFT for neural networks. DMFT for high-dimensional learning dynamics descends from the spin-glass and teacher–student traditions (Crisanti & Sompolinsky, 1987; Sompolinsky et al., 1988; Saad & Solla, 1995; Engel & Van den Broeck, 2001) and has been applied to SGD on Gaussian mixtures, deep networks, attention, and the generalization/overfitting decoupling (Mignacco et al., 2020; Bordelon & Pehlevan, 2022; Bordelon et al., 2024b; Montanari & Urbani, 2025; Bordelon & Pehlevan, 2026). Online-SGD analyses of teacher–student networks, which use fresh samples, are a related line (Goldt et al., 2019). For first-order methods on Gaussian data the DMFT description is rigorous (Celentano et al., 2021). All of these treat single-pathway architectures with no learned gate routing inputs to subnetworks. We extend the framework to a coupled router-and-experts system, which requires a closure for the routing fields and a separate router Onsager correction. Extended discussion and full citation lists are in Appendix A.

3 Model and Setup

We study a two-layer MoE student trained by gradient flow on a Gaussian-mixture teacher, in the proportional limit $n, d \to \infty$ at fixed $\bar{\alpha} = n/d$, with m hidden units per expert. We keep m finite, because the experts' alignments change on a time scale of order $m/\bar{\alpha}$ (Section 4.4). The setup and much of the notation follow Montanari & Urbani (2025), from which we take the order-parameter conventions and the response and correlator notation.

3.1 Data Model

Inputs follow a K-cluster Gaussian mixture with $K = E$, so each expert can specialize to one cluster. Each cluster carries its own teacher direction, so there is a correct routing to learn: a router that sends each cluster to its own expert reduces every expert's task to a single-index problem. For sample μ a label $c^\mu \in \{1, \dots, E\}$ is drawn with prior $\pi_c = 1/E$, and

$$x^\mu \;=\; \kappa\, U_{c^\mu} \;+\; \xi^\mu, \qquad \xi^\mu \sim \mathcal{N}(0, I_d), \qquad y^\mu \;=\; \sigma\big(U_{c^\mu} \cdot x^\mu\big) \;+\; \zeta^\mu, \tag{1}$$

where the unit vector $U_c \in \mathbb{R}^d$ is both cluster mean and teacher direction, so $\mathbb{E}[x \mid c] = \kappa U_c$ and the target depends on x only through $U_c \cdot x$. $\kappa > 0$ controls the signal-to-noise ratio and $\zeta^\mu \sim \mathcal{N}(0, \tau^2)$ is i.i.d. label noise. The activation $\sigma : \mathbb{R} \to \mathbb{R}$ is odd in all our experiments, $\sigma(z) = \mathrm{erf}(z/\sqrt{2})$. The cluster geometry enters via the overlap matrix $\Omega^{UU}_{cc'} = U_c \cdot U_{c'}$. We take orthogonal U_c, $\Omega^{UU} = I_E$.

$K = E$ is a presentational assumption: no equation of motion changes if $K \neq E$, only the *reduction* to two overlap scalars in Section 3.3. The uniform prior matters more, since it makes the clusters *exchangeable* and so supplies the permutation symmetry the reduction rests on. A non-uniform prior leaves the DMFT intact but requires per-cluster order parameters.

3.2 MoE Model and Training

The student is a two-layer MoE with E experts and a single router. Expert e has m first-layer hidden units with weights $w_{l,e}$ on the unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ and second-layer scalars $a_{l,e} \in \mathbb{R}$. The router has weights $v_e \in \mathbb{S}^{d-1}$. The expert outputs and routing weights are

$$f_e(x) \;=\; \frac{1}{m} \sum_{l=1}^{m} a_{l,e}\, \sigma(\langle w_{l,e}, x \rangle), \qquad g_e(x) \;=\; r\big(V^\top x; T\big)_e, \qquad f(x;\theta) \;=\; \sum_{e=1}^{E} g_e(x)\, f_e(x), \tag{2}$$

where θ collects all trainable weights, $V = [v_1, \dots, v_E] \in \mathbb{R}^{d \times E}$, and $r(\,\cdot\,; T)$ is a routing function with temperature T. We treat r generally. The routers we analyze are the soft-discriminant router $r_e = a_0 + a_1\, \mathrm{disc}_e(x)$ with constants a_0, a_1, affine in the centered logit $\mathrm{disc}_e(x) = s_E(v_e - \bar{v}) \cdot x$ with $\bar{v} = \frac{1}{E}\sum_j v_j$ and $s_E = \sqrt{E/(E-1)}$ (numerical validation), softmax $r_e = e^{v_e \cdot x/T} / \sum_j e^{v_j \cdot x/T}$ (REM analysis, Section 5), and quadratic-probit / hard top-k (Appendix H). Both σ and r enter the

closed equations only through Gaussian averages, so the closure is defined for any router. We derive the router Onsager correction in closed form for the soft-discriminant router.

Gradient flow. Training minimizes the squared-loss empirical risk $\mathcal{L}(\theta) = \frac{1}{2n}\sum_\mu (y^\mu - f(x^\mu;\theta))^2$ under the spherical constraints $\|w_{l,e}\| = \|v_e\| = 1$, with a single proportional-asymptotic learning rate $1/d$ applied to the summed loss (equivalently, learning rate $\bar{\alpha}$ for every parameter on the mean loss), and with the second-layer scalars replaced by their mean, $a_{l,e} \to \bar{a}$, as under the symmetric initialization of Montanari & Urbani (2025). The full equations $\dot{w}_{l,e}, \dot{v}_e, \dot{a}_{l,e}$ are in Appendix B.2. This coupling structure drives the DMFT below: $\dot{w}_{l,e}$ is gated by $g_e(x^\mu)$, and $\dot{v}_e$ is sourced by f_e.

3.3 Symmetric Initialization and Order Parameters

The loss, gradient flow and data model are jointly invariant under independently permuting cluster labels and expert indices ($S_E \times S_E$), so unless that symmetry is broken the dynamics cannot distinguish relabeling of states. A *symmetric initialization* breaks the product symmetry to its diagonal subgroup, giving each expert a small bias toward its cluster:

$$\mathbb{E}[w_{l,e}(0)\cdot U_c] \;=\; \varepsilon\,\delta_{ec} + o_d(1) \quad \text{as } d \to \infty, \tag{3}$$

with router overlaps initialized analogously at scale ε^v. Under the residual diagonal S_E, every expert–cluster bilinear collapses to two scalars, distinguished only by whether the expert and cluster indices match. At training time t,

$$P_{\text{self}}(t) \;=\; \mathbb{E}[w_{l,e}(t)\cdot U_e], \qquad P_{\text{cross}}(t) \;=\; \mathbb{E}[w_{l,e}(t)\cdot U_{e'}]\big|_{e'\neq e}, \tag{4}$$

and similarly $P^v_{\text{self}}, P^v_{\text{cross}}$ for the router, reducing all E^2 overlaps to two scalars each for the experts and the router. The second layer is described by the single scalar $\bar{a}(t)$ (Appendix B.1). We define the *specialization order parameter*

$$\Delta(t) \;:=\; P_{\text{self}}(t) - P_{\text{cross}}(t), \tag{5}$$

which measures how much more each expert aligns with its own cluster than with the others. It is positive when experts have aligned with their own clusters and zero when they have collapsed onto a cluster-averaged solution. The reduction extends to all two-time correlators and responses, leaving the finite-dimensional Volterra system derived in Section 4.

4 Core DMFT

Following every weight of a large network is intractable, but in high dimension a few summary statistics determine the training curve: how strongly each expert's weights and the router's weights overlap with the teacher directions, and with each other. As $n, d \to \infty$ at fixed $\bar{\alpha} = n/d$ these overlaps concentrate, and the gradient flow of Section 3.2 closes on them, with DMFT supplying the closed equations (Sompolinsky et al., 1988; Mignacco et al., 2020; Bordelon & Pehlevan, 2022). Because the same n examples are reused throughout training, every past gradient step has already partly fit the examples the model sees now, so the equations carry memory: they are ODEs coupled to Volterra integral equations whose kernels integrate over the past. We derive them with the standard Martin–Siggia–Rose–De Dominicis–Janssen (MSRDJ) path integral (Appendix C), use the S_E symmetry to reduce the E^2 expert–cluster overlaps to a matched and a mismatched value, and integrate the result by forward Euler.

4.1 What the Theory Tracks

The state of training is summarized by single-time and two-time quantities. The single-time ones are the alignments of Section 3, $P_{\text{self}}, P_{\text{cross}}$ for the experts and $P^v_{\text{self/cross}}$ for the router, the output scale $\bar{a}(t)$, and the router's preference for the matched expert, the router contrast $\Delta_v(t) := P^v_{\text{self}} - P^v_{\text{cross}}$, or its normalized form $r_D(t) = \sqrt{(E-1)/E}\,\Delta_v(t)$. The two-time ones record how the state at time t is correlated with the state at an earlier time s and how it responds to perturbations made at s, and they carry the memory. The central one is the residual response $R_A(t,s)$, the change in the residual

at time t caused by a perturbation of the network output at time s (Appendix B.1 lists the rest). Each single-time quantity obeys an ODE of the form

$$\frac{dP_{\text{cross}}}{dt} = \underbrace{\nu_w(t)\,P_{\text{cross}}}_{\text{constraint drift}} + \underbrace{\frac{\bar{\alpha}\,\bar{a}(t)}{mE}\left[\tilde{\Phi}_{\text{cross}}(t) - \Psi_{\text{cross}}(t)\right]}_{\text{teacher signal} - \text{self-interaction}} - \underbrace{\text{mem}_{P_c}(t)}_{\text{memory backreaction}} . \tag{6}$$

The first term, with ν_w the Lagrange multiplier of the constraint $\|w\| = 1$, keeps the weights on the unit sphere and acts as a norm-controlling decay. The second is the gradient signal: the pull toward the teacher, $\tilde{\Phi}$, minus the part of the target the network already fits, Ψ. The third is the memory of past steps on the reused data. The analogous equations for P_{self}, $P^v_{\text{self/cross}}$ and $\bar{a}$, with every kernel written out, are in Appendix D.5. The system is closed: every term is an order parameter or a fixed function of the model, so no simulation enters as input. Several modeling choices, listed in Appendix D, were made by comparison with gradient flow at $E = 2$.

4.2 Why Routing Needs an Approximation

The DMFT we build on treats each example's residual as a Gaussian process with the same mean and covariance, an approximation whose error vanishes on time scales of order one in the wide-network limit and is small in simulations at later times (Montanari & Urbani, 2025). We use it at finite width. In an MoE the residual $y - \sum_e g_e f_e$ contains gates that are nonlinear functions of the router preactivations, so averaging over the data leaves a non-Gaussian term in the effective action. We close it by treating the routing fields as Gaussian at their tracked mean and covariance, which we call the second-cumulant closure. This is a closure rather than a controlled expansion: the neglected cumulants carry no small prefactor, and even an affine gate produces them through the log-determinant that the data average leaves. We expect it to be accurate in the regime

$$|P^v_{\text{self}}|,\ |P^v_{\text{cross}}| \ll 1 \quad \text{(weak router–teacher alignment)}, \qquad T \gg T_c^{(0)} \quad \text{(unfrozen routing)}, \tag{7}$$

where the router is still weakly aligned with the teacher and routing has not frozen (Section 5). At $E = 2$ it stays accurate up to $r_D = 0.6$ (Figure 3). Section 4.4 measures its error directly, and Appendices K and C.4 give the details. With the closure, the memory kernels in (6) become products of $R_A(t, s)$ with Gaussian expectations of the gate and activation at the current order parameters (Appendix D.2). Our integration keeps the R_A-based memory of the expert and router kernels only at first order in the Euler step, a closure choice rather than a derived result. Halving the step changes Δ and P_{self} at $t = 50$ by under 2% of their change during training (Appendix D.4).

4.3 The Router Onsager Correction

The theory so far still misses the shape of training. It predicts that the mismatched overlap P_{cross} grows steadily, while gradient flow shows it rising, peaking and then decaying (Figure 1). The missing piece is overfitting in the router. The router is trained on the same examples whose routing we evaluate, so it routes its training data more confidently than fresh data. In softmax gradient flow at $E = 4$, for example, a trained router gives the matched expert a weight of 0.53 on its training inputs and only 0.39 on fresh ones (Appendix F.8).

The mechanism is a self-reaction. Each training example μ pushes the router by $\delta V \propto \text{res}^\mu\, \delta F^\mu_e\, x^\mu/d$, where $\text{res}^\mu = y^\mu - f(x^\mu;\theta)$ is its residual and $\delta F^\mu_e = f_e(x^\mu) - \frac{1}{E}\sum_j f_j(x^\mu)$ is expert e's output relative to the mean. Because $\|x^\mu\|^2/d \to 1$, this push moves the example's own routing score disc^μ, its centered logit for the matched expert, by an $\mathcal{O}(1)$ amount. Relative to the population drive the effect is $\mathcal{O}(1/\bar{\alpha})$, so it disappears only when every example is fresh. Single-pathway DMFT corrects the analogous bias of the expert weights through R_A, but R_A does not capture a bias in which expert sees which example. For the soft-discriminant router, accumulating the pushes gives each training example the routing shift

$$\Delta\text{disc}^\mu(t) = a_1 \tfrac{E}{E-1} \int_0^t R^v_-(t, s)\, \text{res}^\mu(s)\, \delta F^\mu_{c^\mu}(s)\, ds, \tag{8}$$

where $R^v_-(t, s)$ is the router's response to a push along an expert's own contrast direction (Appendix E.3). The shift changes the example's gate and, through the part of the fit the example has

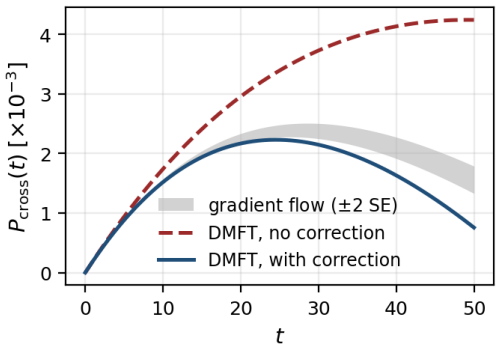


Figure 1: **The router Onsager correction.** $P_{\text{cross}}(t)$ at $E = 2$ (soft-discriminant router, settings of Figure 5) for gradient flow (gray, ± 2 SE over 16 seeds) and for the DMFT with and without the router self-reaction (the arm without it also sets $\delta_{\text{disc}} = 0$). The self-reaction produces the rise and decay and lowers the trajectory RMSE significantly. The corrected curve reaches 93% of the gradient-flow peak three time units early.

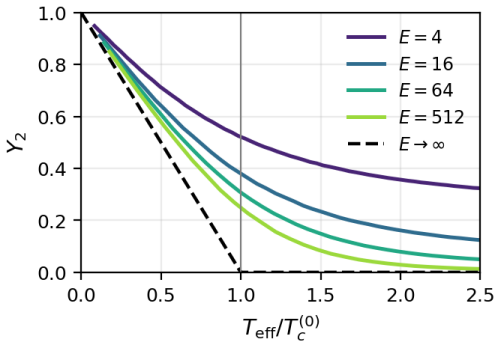


Figure 2: **Freezing of softmax routing.** Participation ratio $Y_2 = \sum_e g_e^2$ for $g_e \propto \exp(\eta_e/T_{\text{eff}})$ with i.i.d. $\eta_e \sim \mathcal{N}(0, 1)$, at effective temperature T_{eff} (Section 5.1). Dashed: the $E \to \infty$ REM result $Y_2 = 1 - T_{\text{eff}}/T_c^{(0)}$ below $T_c^{(0)}$ and 0 above (Derrida & Toulouse, 1985). At finite E the transition is a crossover that sharpens slowly as E grows.

done for itself, its residual. Because it is largest for the examples the router was pushed on hardest, it is correlated with their residuals and expert outputs, and this correlation reshapes expert learning. Adding it to the expert drive produces the rise and decay of P_{cross} (Figure 1, Appendix E.5). The cluster mean of the shift, δ_{disc}, is the measurable bias in routing. It obeys a closed Volterra equation whose source has no fitted constant and matches a direct gradient-flow measurement to within 3% at $E = 2, 4, 8$ (Appendix E.4). Fed alone into the expert drive, this mean produces about half of the self-reaction's effect (Appendix E.6).

4.4 Validation Against Gradient Flow

The theory tracks training across widths. Figure 3 compares the equations with gradient flow at $E = 2$ over a $12\times$ range of widths. We measure errors as the time-averaged RMSE divided by how much the gradient-flow curve changes during training, or by its peak for P_{cross}, which starts at zero. The DMFT tracks Δ and P_{self} (not plotted) to within 6.7% and 2.4% at $m = 100$, and 1.5% and 0.8% at $m = 1200$. P_{cross} is the least accurate quantity: the DMFT peaks three to five time units early and then decays further, so its error falls only from 22% of the small peak at $m = 100$ to 13% at $m = 1200$. It captures the height of the peak to within 12% and its $1/m$ shrinkage (12.4-fold in gradient flow, 11.8-fold in the DMFT). Width matters because the expert alignments move by $\mathcal{O}(1/m)$ at fixed time: in the DMFT $m\,(\Delta(t) - \varepsilon)$ is the same at every width (4.9 at $t = 50$), and gradient flow approaches this as m grows (4.1 at $m = 100$, 4.7 at $m = 1200$). The router's $r_D(t)$, by contrast, has a width-independent limit.

Where the closure holds. At high temperature softmax is close to uniform, and to first order in the router logits $z_e^v = v_e \cdot x$ it is exactly the soft-discriminant gate, $g_e = 1/E + (z_e^v - \bar{z}^v)/(ET) + \mathcal{O}(((z_e^v - \bar{z}^v)/T)^2)$ with $\bar{z}^v$ their mean, at the coupling

$$a_1(T, E) = \sqrt{E - 1}/(T\,E^{3/2}). \tag{9}$$

The soft-discriminant DMFT at this coupling is therefore the leading high-temperature approximation to the softmax DMFT, and comparing it with softmax gradient flow measures two errors at once: the closure and the linearization of the gate. At $E = 2$, over nine temperatures, the error on Δ_v rises smoothly as T falls. Evaluating the gate means and the expert and router drives with the exact softmax removes most of this error (Appendix I). The rest, from the closure and the router-side terms we keep linearized, is below 1% of the gradient-flow peak of Δ_v at $2.5\,T_c^{(0)}$ and grows as T falls, to

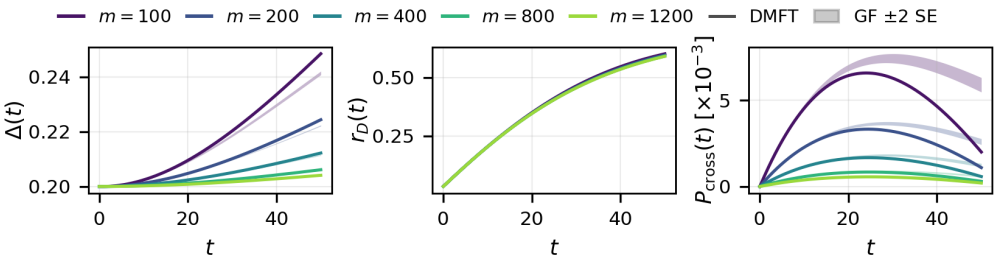


Figure 3: **DMFT–GF agreement across expert width m.** Lines: DMFT. Bands: GF mean ± 2 SE over 8 seeds. Left to right: $\Delta(t)$, $r_D(t)$, $P_{\text{cross}}(t)$. Settings: $E = 2$, $\bar{\alpha} = \kappa = 1$, $a_1 = 0.3$, $\sigma(z) = \text{erf}(z/\sqrt{2})$, $d = 6000$, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, integration step 0.1.

4.9% at $0.7\,T_c^{(0)}$, where the fully linearized DMFT's error is 11 times larger. Softmax runs at $E = 4$ and 8 show the same trends.

5 The REM Correspondence and Phase Diagram

A softmax router must choose among E options whose scores are partly signal and partly noise. At high temperature it spreads weight over all of them. At low temperature it concentrates on the highest score, and when the noise is larger than the signal, the highest score belongs to a random expert. Statistical physics has an exactly solvable model of this choice, Derrida's Random Energy Model (REM) (Derrida, 1981): a Boltzmann distribution over many states whose energies are independent Gaussians. A Boltzmann distribution is a softmax: each state gets weight proportional to $e^{-\text{energy}/T}$. The central prediction of the REM is a sharp temperature below which the distribution *freezes* onto a few of the lowest-energy states. This section shows that softmax routing is a REM, derives when the router freezes onto the wrong expert, and combines this with the linearized DMFT and gradient flow to describe the regimes of training.

5.1 Softmax Routing as a Random Energy Model

For an input from cluster c, the router's logit for expert e is $z_e^v = v_e \cdot x$. It splits into a signal $\kappa P_{e,c}^v(t)$, with $P_{e,c}^v = v_e \cdot U_c$ the router's alignment with that cluster, and noise $v_e \cdot \xi$. Different experts' router vectors overlap by $C_o^v(t,t)$, so their noises are correlated. Up to a shift common to all logits, which softmax ignores, the noise equals $\sqrt{1 - C_o^v(t,t)}\,\eta_e(t)$ with i.i.d. standard Gaussian η_e, for any $C_o^v \geq -1/(E-1)$ (Appendix F.1). Dividing the temperature by this noise scale, $T_{\text{eff}}(t) = T/\sqrt{1 - C_o^v(t,t)}$, gives the temperature of an equivalent problem with unit-variance noise, and turns the routing weights exactly into REM Boltzmann weights over E i.i.d. Gaussian energies, shifted by the signal.

Freezing. Before the router aligns with the clusters, the signal is nearly the same for every expert and drops out, so each expert's logit is pure noise. At temperature T_{eff} an expert with noise η receives weight proportional to $e^{\eta/T_{\text{eff}}}$. Experts with large noise are rare, since about $E\,e^{-\eta^2/2}$ of the E experts have noise near η, but each of them receives more weight, and the total weight comes mostly from experts with noise near $\eta^* = 1/T_{\text{eff}}$, where these two effects balance. As T_{eff} falls, η^* rises, until it reaches the largest noise present among E experts, about $\sqrt{2\ln E}$. This happens at

$$T_c^{(0)} = \frac{1}{\sqrt{2\ln E}}. \tag{10}$$

Below this temperature the router *freezes*: the weight no longer spreads over a large group of experts and goes to the few with the largest noise. Which experts win is decided by the noise, not by which expert is right, and cooling further concentrates the weight on fewer of them. We measure the concentration by the participation ratio $Y_2 = \sum_e g_e^2$, which equals $1/E$ when an input's weight is spread evenly over all E experts and 1 when it all goes to one, so $1/Y_2$ is the effective number of

experts an input uses. As $E \to \infty$ the average of Y_2 tends to 0 above $T_c^{(0)}$ and to $1 - T_{\text{eff}}/T_c^{(0)}$ below it (Derrida & Toulouse, 1985; Mézard et al., 1987), so below $T_c^{(0)}$ each input uses a few experts chosen by the noise rather than by its cluster (Figure 2), the router collapse of Section 1. At finite E the change from diffuse to concentrated routing is spread over temperatures of relative width $\mathcal{O}(1/\sqrt{\ln E})$, already of order one at $E = 2$, so the regimes below are asymptotic in $\ln E$.

Freezing onto the right expert. Once the router aligns, the matched expert's logit has a head start $\kappa\,\Delta_v(t)$. The router *condenses* onto the right expert, giving it more weight than all other experts together, when that head start outweighs the combined weight of the other $E - 1$ experts, whose largest noise is set by $\ln(E-1)$. Averaging that combined weight over the noise gives its logarithm as $\ln(E-1) + \beta_{\text{eff}}^2/2$ with $\beta_{\text{eff}} = 1/T_{\text{eff}}$ (Appendix F.3), and equating it with the matched expert's weight gives the condensation temperature

$$T_c^{\text{eff},\pm}(t) \;=\; \frac{s \pm \sqrt{s^2 - 2\ln(E-1)}}{2\ln(E-1)}, \qquad s := \frac{\kappa\,\Delta_v(t)}{\sqrt{1 - C_o^v(t,t)}}, \tag{11}$$

where s is the head start in units of the noise. Only the upper root is meaningful: the noise average used here is reliable only above $T_c^{(0)}$, and the lower root falls below $T_c^{(0)}$ just past threshold. A real root requires a minimum head start,

$$\kappa\,\Delta_v(t) \;\geq\; \sqrt{1 - C_o^v(t,t)}\,\sqrt{2\ln(E-1)} \;=:\; \kappa\,\Delta_v^*. \tag{12}$$

Below this signal the router cannot condense onto the matched expert, so if it freezes, it freezes onto a random one. The data bound the signal. With orthonormal cluster directions and unit-norm routers, $(\Delta_v)^2 \leq 1 - C_o^v$, so $s \leq \kappa$ (Appendix F.8). Condensation onto the matched expert therefore needs a cluster signal-to-noise ratio of at least $\sqrt{2\ln(E-1)}$, however training proceeds. At $E = 4$, where this threshold is 1.48, it is a gradual crossover, and it sharpens as E grows. In gradient flow at $E = 4$ and $\kappa \in \{1, 2, 3\}$, the REM at each trained router's measured s and T_{eff} predicts its mean matched-expert weight on fresh inputs to within 0.004 (Appendix F.8). Above the threshold, specialization and routing can reinforce each other through the loop below, whose last step we do not derive:

$$\Delta_v(t) \nearrow \implies T_c^{\text{eff}}(t) \nearrow \implies \text{condensation deepens} \implies \text{experts specialize} \implies \Delta_v(t) \nearrow . \tag{13}$$

The bare REM, with i.i.d. energies, has no such loop.

5.2 Three Regimes of Training

The REM divides the router's state into three regimes (Figure 4, left). In the *democratic* regime routing stays spread over many experts. In the *specialization* regime the router condenses onto the matched expert and the loop (13) can engage. In the *frozen-collapse* regime, which requires $T_{\text{eff}} < T_c^{(0)}$, the router locks on before its signal has grown, onto experts poorly aligned with the clusters it routes. Because condensation onto the matched expert needs $\kappa \geq \sqrt{2\ln(E-1)}$, only the democratic and frozen-collapse regimes are reachable at $\kappa = 1$. The experts nevertheless specialize in both, through the onset described below, so expert specialization and router condensation are distinct processes. Condensation onto the matched expert is *thermodynamic*: at a given head start and temperature it is a property of the softmax, the REM line (11). The onset of specialization is *dynamical*: the linearized DMFT sets its rate, it depends on how training grows from its initialization, and it does not need the minimum signal (12). To our knowledge the frozen-collapse regime is new to MoE training dynamics. For single-pathway two-layer networks, Veiga et al. (2022) give a phase diagram of SGD.

The onset amplifies the initial bias. The symmetric state, in which every expert is equally aligned with every cluster, is unstable. An expert slightly more aligned with its cluster receives slightly more of that cluster's inputs, which aligns it further, and this loop outweighs the forces that restore symmetry. In the linearized DMFT a small initial bias grows at a rate $\Lambda(t) > 0$ at every temperature we evaluate, and this rate decays as the loss falls (Appendix J). A decaying rate need not give a finite *gain* $\exp(\int \Lambda\, dt)$, but in gradient flow at $E = 4$, $T = 1$ and $\bar{\alpha} = 8$ the growth stops. With router bias $\varepsilon/4$, $\bar{\Delta}$ first reaches 0.2 at $t = 57$ and 172 for $\varepsilon = 0.1$ and 0.03. A constant rate fitted to these

times predicts that $\varepsilon = 0.01$ reaches it near $t = 280$, yet that run peaks at $\Delta = 0.17$ near $t = 1600$, a gain of 17 (one seed, to $t = 2400$). Larger biases are amplified less: Δ peaks at 0.39 and 0.61 for $\varepsilon = 0.03$ and 0.1, so runs approach no common specialized state. At $\bar{\alpha} = 1$ and 4 our runs end at $t = 800$ still growing, at $\Delta/\varepsilon \approx 3$ and 8. The learning rate on the mean loss is $\bar{\alpha}$, and at equal $\bar{\alpha}t$ the three gains are close (3.2, 4.0, 4.3).

Gradient flow at $E = 8$. The right two panels of Figure 4 show gradient flow at $E = 8$ and $\kappa = 1$, where both reachable regimes appear. At low temperature the router is strongly concentrated but poorly aligned ($Y_2^\infty \approx 0.99$, $\Delta_v^\infty \approx 0.12$ at $T = 0.10$, averaged over $\bar{\alpha}$), and at high temperature it is diffuse but better aligned ($Y_2^\infty \approx 0.28$, $\Delta_v^\infty \approx 0.47$ at $T = 1.50$). The experts specialize at every temperature, most at $T = 0.45$, just below $T_c^{(0)}$. At $\kappa = 3$, where the threshold $s_* \approx 1.97$ can be reached, the same grid also shows the specialization regime: in all eight cells the REM places there, the router gives the matched expert more weight on fresh inputs than all other experts together, and at $T = 1.5$, above the REM line, it gives it less (Appendix F.8). Below $T_c^{(0)}$ the router freezes first, and the experts specialize less the lower the temperature (Appendix J).

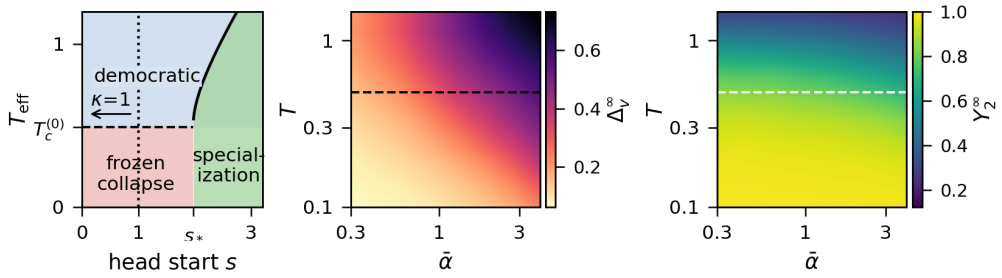


Figure 4: **Regimes of softmax routing, theory and gradient flow.** *Left*: regimes at $E = 8$ against the head start s and T_{eff}, with $T_c^{(0)}$ (dashed), the branch $T_c^{\text{eff},+}$ (11) (solid) from $s_* = \sqrt{2 \ln 7}$, and the bound $s \leq \kappa$ at $\kappa = 1$ (dotted). *Center, right*: gradient flow at $E = 8$, $\kappa = 1$ (one seed per cell, $d = 2500$, $m = 120$, training time 200), final Δ_v^∞, and Y_2^∞ measured on the training inputs, dashed $T_c^{(0)} \approx 0.49$. Because $s \leq \kappa$, every run lies left of the dotted line in the left panel: routing is concentrated but poorly aligned at low T and diffuse but better aligned at high T. The Y_2^∞ crossover is broad, drifts with $\bar{\alpha}$, and is measured on inputs the router was trained on, so these panels exhibit the two regimes rather than locate the REM line.

6 Discussion and Conclusion

Implications. For router logits of unit variance, the freezing scale $T_c^{(0)} = 1/\sqrt{2 \ln E}$ falls only slowly with E (0.49 at $E = 8$, 0.35 at $E = 64$). For $E \geq 3$ a router can lock onto the correct expert only if the cluster signal-to-noise ratio exceeds $\sqrt{2 \ln(E-1)}$, a threshold that sharpens as E grows. Below that threshold, a router run below $T_c^{(0)}$ condenses onto poorly aligned experts, and cooling it further only makes its errors more confident. From a near-symmetric start, training amplifies a small initial bias. In the runs long enough to see the amplification stop, at $\bar{\alpha} = 8$, the gain is finite, so the initial bias sets how far experts specialize. Finally, load-balancing losses act on the load, and frozen routing in our setting leaves the load balanced, so at leading order these losses exert no force against frozen collapse (Appendix F.7).

Limitations and future work. The theory assumes orthogonal, equally likely clusters, two-layer experts, a single global router, and gradient flow, and it derives the Onsager source explicitly only for the soft-discriminant router. Two caveats remain. First, Λ is small at our parameters and decays as the loss falls, so the onset is slow here and may be faster with a stronger signal. Second, the Gaussian-mixture teacher makes the freezing transition solvable, so for production MoE our claims concern mechanism, not magnitude. The prediction expected to transfer is that specialization and router condensation stay distinct wherever routing is hard.

REPRODUCIBILITY STATEMENT

Every equation in the main text is derived in the appendices or stated there with its assumptions, including the MSRDJ construction and the cumulant closure, the Volterra system as integrated, the router Onsager source, and the REM correspondence with its proofs. Closures that are not derived are labeled as such. All figure captions state the hyperparameters needed to reproduce the results. We plan to release the code at the end of the review process.

AI USE STATEMENT

Generative AI tools were used for polishing some of the writing, checking equation and notational consistency, double-checking derivations, and assisting with code. They were not used for generating research ideas, performing initial derivations. All derivations in the main text and appendices were performed and manually checked by the authors, and every numerical value reported in this paper was produced by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including all text, claims, and results.

REFERENCES

Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. SGD learning on neural networks: Leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory*, 2023.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional and mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layer networks. *arXiv preprint arXiv:2302.05882*, 2023.

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, and Jean Barbier. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22 (106):1–51, 2021.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in neural information processing systems*, 35, 2022.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

Blake Bordelon and Cengiz Pehlevan. Disordered dynamics in high dimensions: Connections to random matrices and machine learning. *arXiv preprint arXiv:2601.01010*, 2026.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, 2024a.

Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *Advances in Neural Information Processing Systems*, 37:35824–35878, 2024b.

Jean-Philippe Bouchaud and Marc Mézard. Universality classes for extreme-value statistics. *Journal of Physics A: Mathematical and General*, 30(23):7997–8015, 1997.

Anton Bovier. *Statistical Mechanics of Disordered Systems: A Mathematical Perspective*. Cambridge University Press, 2006.

Tommaso Castellani and Andrea Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, pp. P05012, 2005.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *arXiv preprint arXiv:2208.02813*, 2022.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. On the representation collapse of sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Andrea Crisanti and Haim Sompolinsky. Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Physical Review A*, 36(10):4922–4939, 1987.

Damai Dai, Chengqi Deng, Chenggang Zhao, R X Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

Alex Damian, Jason D Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. *arXiv preprint arXiv:2206.15144*, 2022.

Alex Damian, Loucas Pillaud-Vivien, Jason D Lee, and Joan Bruna. Computational-statistical gaps in Gaussian single-index models. *arXiv preprint arXiv:2403.05529*, 2024.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *International Conference on Machine Learning*, 2024.

C De Dominicis. Techniques of field renormalization and dynamics of critical phenomena. *Journal de Physique Colloques*, 37(C1):247–253, 1976.

B. Derrida and G. Toulouse. Sample to sample fluctuations in the random energy model. *Journal de Physique Lettres*, 46(6):223–228, 1985.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24(5):2613–2626, 1981.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2014.

Andreas Engel and Christian Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

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.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation, 2026.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in Neural Information Processing Systems*, 32, 2019.

David J Gross and Marc Mézard. The simplest spin glass. *Nuclear Physics B*, 240(4):431–452, 1984.

Xu Owen He. Mixture of a million experts. *arXiv preprint arXiv:2407.04153*, 2024.

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.

H K Janssen. On a Lagrangean for classical field dynamics and renormalization group calculations of dynamical critical properties. *Zeitschrift für Physik B Condensed Matter*, 23(4):377–380, 1976.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Tianze Jiang, Blake Bordelon, Cengiz Pehlevan, and Boris Hanin. Hyperparameter transfer with mixture-of-expert layers, 2026.

Kukjin Kang and Jong-Hoon Oh. Statistical mechanics of the mixture of experts. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

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, 2025.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. *arXiv preprint arXiv:2103.16716*, 2021.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34, 2021.

Carlo Lucibello and Marc Mézard. Exponential capacity of dense associative memories. *Physical Review Letters*, 132(7), February 2024. ISSN 1079-7114. doi: 10.1103/physrevlett.132.077301.

P C Martin, E D Siggia, and H A Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423–437, 1973.

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

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

Marc Mézard, Giorgio Parisi, and Miguel A Virasoro. *Spin glass theory and beyond*. World Scientific, 1987.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33, 2020.

Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

Hien D Nguyen, Luke R Lloyd-Jones, and Geoffrey J McLachlan. A universal approximation theorem for mixture-of-experts models. *Neural Computation*, 28(12):2585–2593, 2016.

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

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

Maria Refinetti and Sebastian Goldt. The dynamics of representation learning in shallow, non-linear autoencoders. In *International Conference on Machine Learning*, pp. 18499–18519. PMLR, 2022.

Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional Gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pp. 8936–8947. PMLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8583–8595, 2021.

Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. *Advances in Neural Information Processing Systems*, 34, 2021.

Grant M Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75 (9):1889–1935, 2022. Preprint 2018, arXiv:1805.00915.

David Ruelle. A mathematical reformulation of Derrida's REM and GREM. *Communications in Mathematical Physics*, 108(2):225–239, 1987.

David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52 (4):4225–4243, 1995.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

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. *arXiv preprint arXiv:1701.06538*, 2017.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Haim Sompolinsky, Andrea Crisanti, and H J Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262, 1988.

Qwen team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Leena Chennuru Vankadara, Moritz Haas, Luke Hayward, Sebastian Bordt, and Alessandro Breccia. How to scale mixture-of-experts: From mup to the maximally scale-stable parameterization, 2026.

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *Advances in Neural Information Processing Systems*, 35, 2022.

Xun Wu, Shaohan Huang, Wenhui Wang, and Furu Wei. Multi-head mixture-of-experts. *arXiv preprint arXiv:2404.15045*, 2024.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

A Extended Related Work

This appendix expands the discussion of related literature beyond the main-text Section 2, providing full citation lists and historical context.

Mixture-of-Experts in practice. The MoE layer dates back to Jacobs et al. (1991), was extended to deep networks by Eigen et al. (2014), and was revived as a sparse, learnable routing primitive by Shazeer et al. (2017) (see Yuksel et al. (2012); Masoudnia & Ebrahimpour (2014) for surveys). Sparse-gating variants have since become standard in large language models (Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022; Zoph et al., 2022; Jiang et al., 2024; Dai et al., 2024; Liu et al., 2024; team, 2024), vision (Riquelme et al., 2021), and the many-expert limit (He, 2024), with dense soft-gating variants such as Puigcerver et al. (2023). Training these architectures requires auxiliary load-balancing losses or hard assignment tricks (Lewis et al., 2021; Roller et al., 2021; Zhou et al., 2022) whose justification has remained largely empirical, and suffers from representation collapse (Chi et al., 2022) and low expert activation (Wu et al., 2024). We do not model these interventions, but our phase diagram separates the democratic and frozen-collapse regimes, which respond to different controls.

DMFT for neural networks (extended). Online-SGD analyses have compared two-layer networks with kernel methods on Gaussian mixtures (Refinetti et al., 2021) and studied autoencoders (Refinetti & Goldt, 2022). DMFT has been used to analyze kernel evolution in wide networks (Bordelon & Pehlevan, 2022), scaling laws (Bordelon et al., 2024a), and batch reuse (Dandi et al., 2024). The mean-field perspective on overparametrized two-layer networks (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2022; Sirignano & Spiliopoulos, 2020) and committee-machine and exactly solvable analyses (Aubin et al., 2018; Saxe et al., 2014; Veiga et al., 2022; Loureiro et al., 2021) provide related tools, all for single-pathway networks.

Information exponent and single-index learning. The information exponent k^* governing SGD timescales for single-index models was introduced by Ben Arous et al. (2021) and refined for neural-network feature learning by Damian et al. (2022); Arnaboldi et al. (2023); Abbe et al. (2023); Bietti et al. (2022); Berthier et al. (2023); Damian et al. (2024). Kawata et al. (2025) relate the benefit of MoE on clustered single-index tasks to the information exponent, which is low within each cluster but high for the task as a whole. We define a joint teacher–gate exponent and discuss, as a heuristic, how it could set the speed of the onset (Appendix H).

Random Energy Model. The REM (Derrida, 1981; Gross & Mézard, 1984; Ruelle, 1987; Bouchaud & Mézard, 1997) is the simplest mean-field model of a glassy free-energy landscape and underlies much of replica theory (Mézard et al., 1987; Castellani & Cavagna, 2005; Bovier, 2006). We exploit a correspondence between softmax routing and the REM Boltzmann measure to convert the MoE training-dynamics question into a phase diagram organized by the freezing scale $T_c^{(0)}$. The correspondence between softmax weights and the REM Boltzmann measure was identified by Lucibello & Mézard (2024) for dense associative memories and applied to attention at initialization by Giorlandino & Goldt (2026). To our knowledge it has not been applied to a *learned router*.

B Notation and Background

This appendix collects the symbols used throughout the paper and states the technical facts on which the derivations rely. Readers who know DMFT can skip it. The technical appendices that follow assume this notation without re-introduction.

B.1 Symbol table

Architecture and data	
E	number of experts (= number of clusters K)
m	first-layer hidden units per expert

continued on next page

Symbol table, continued	
d	input dimension
n	number of training examples
$\bar{\alpha} = n/d$	sample complexity (proportional asymptotic ratio)
κ	cluster signal-to-noise ratio: $x = \kappa\, U_c + \xi$ (with a subscript, κ_ℓ is an ℓ-th cumulant)
τ^2	label-noise variance
T	router temperature, and its rescaled form $T_{\text{eff}} = T/\sqrt{1-C_o^v(t,t)}$
h	forward-Euler step size of the DMFT integration
η_e	i.i.d. standard Gaussian REM noise of expert e (Section 5)
ε	symmetric-initialization bias scale, eq. (3)
ε^v	router initialization bias scale, below eq. (3)
$U_c \in \mathbb{S}^{d-1}$	teacher direction for cluster c
$\Omega^{UU}_{cc'} = U_c \cdot U_{c'}$	cluster-overlap matrix
$w_{l,e}, v_e \in \mathbb{S}^{d-1}$	expert and router weight vectors
$a_{l,e} \in \mathbb{R}$	second-layer scalar, with $\bar{a}$ its mean
$z^w_{l,e} = w_{l,e} \cdot x$	expert preactivation
$z^v_e = v_e \cdot x$	router preactivation
$\lambda_c = U_c \cdot x$	teacher-direction projection, with $\lambda_{\text{tot}} = \sum_c \lambda_c$
σ	teacher / student activation (centered)
$r(\cdot; T)$	routing function, with gate $g_e(x) = r(V^\top x; T)_e$
a_0, a_1	constant ($1/E$) and coupling of the soft-discriminant gate (a_2, a_3: quadratic and cubic gate coefficients)
$\text{res}^\mu = y^\mu - f(x^\mu; \theta)$	residual on training example μ
Single-time order parameters (scalar functions of t)	
$P_{\text{self}}, P_{\text{cross}}$	expert–teacher overlap (matched / mismatched cluster)
$\Delta = P_{\text{self}} - P_{\text{cross}}$	specialization order parameter
$P^v_{\text{self}}, P^v_{\text{cross}}$	router–teacher overlap (matched / mismatched cluster)
$\Delta_v = P^v_{\text{self}} - P^v_{\text{cross}}$	router contrast
$r_D = \sqrt{(E-1)/E}\,\Delta_v$	normalized router contrast
$\bar{a}$	mean second-layer amplitude
δ_{disc}	mean routing shift on training examples, eq. (103)
ν_w, ν_v	spherical-constraint Lagrange multipliers
Two-time correlators and responses	
C_d, C_o, C_x	expert-weight correlators: same-neuron / off-neuron same-expert / cross-expert
C_{dv}, C_{vo} (also C^v_d, C^v_o)	router correlators: intra-expert / cross-expert
$Q^v_{\text{self/cross}}(t,s)$	expert–router cross-covariance (two-time, distinct from the single-time router–teacher overlap $P^v_{\text{self/cross}}(t)$ above)
C_A, R_A (also $\mathcal{C}_A, \mathcal{R}_A$)	residual correlator and response (full propagator)
$R_d, R_o, R_x, R_{dv}, R_{vo}$	response (retarded Green's) functions matching the C-subscript convention
Memory and signal kernels	
$M^d_R, M^o_R, M^{vd}_R, M^{vo}_R$	response-driven self-energy memory kernels (Appendix D.2)
F_e	expert output averaged over its hidden units at the current order parameters (Appendix D.6)
$\delta F_e = F_e - \frac{1}{E}\sum_j F_j$	expert-output contrast (drives the router gradient)
$R^v_- = R_{dv} - R_{vo}$	router contrast response (Appendix E.3)
$s_E = \sqrt{E/(E-1)}$	discriminant normalization, $\text{disc}_e = s_E(v_e - \bar{v})\cdot x$
$M^\bullet_C$	correlator-driven counterparts (Appendix D.2)
$\tilde{\Phi}_{\text{self/cross}}$	expert-weight signal-source kernels (Appendix D.3)
$\tilde{\Phi}^{\text{self/cross}}_v$	router signal-source kernels
$\hat{\Phi}_{\text{self/cross}}$	amplitude ($\bar{a}$) signal kernels
$H_\mu, H'_\mu, H^{\text{lam}}_\mu, \partial_{m_i} H_\mu, \partial_{m_i} H^{\text{lam}}_\mu$	two-dimensional Gaussian-expectation building blocks (Appendix D.1)
Phase-diagram quantities	
$T_c^{(0)} = 1/\sqrt{2\ln E}$	bare REM freezing temperature

continued on next page

Symbol table, continued	
$T_c^{\text{eff}}(t)$	matched-condensation temperature, eq. (11)
$\Lambda(t)$	linearized growth rate of δP, eq. (133)
$s = \kappa\Delta_v/\sqrt{1-C_o^v}$	head start of the matched expert, eq. (11)
$Y_2 = \sum_e g_e^2$	participation ratio
Hermite expansion	
He_k	probabilist's Hermite polynomial of order k
$\hat{\sigma}_k = \mathbb{E}[\sigma(G)\mathrm{He}_k(G)]$	teacher-activation Hermite coefficient
$\hat{\rho}_k$	router-function Hermite coefficient (diagonal slot)
$k^* = \min\{k \geq 1 : \hat{\sigma}_k\hat{\rho}_k \neq 0\}$	joint information exponent, eq. (126)

B.2 Key technical facts

The following standard facts are invoked throughout the paper. We state them in the conventions used here. References are given for proofs.

Hermite expansion (probabilist's convention). Any L^2 function $f : \mathbb{R} \to \mathbb{R}$ against the standard Gaussian measure admits the orthogonal-polynomial expansion $f(z) = \sum_{k\geq 0} \hat{f}_k \,\mathrm{He}_k(z)/k!$ with $\hat{f}_k = \mathbb{E}[f(G)\mathrm{He}_k(G)]$ for $G \sim \mathcal{N}(0,1)$, and $\mathbb{E}[\mathrm{He}_j(G)\mathrm{He}_k(G)] = j!\,\delta_{jk}$.

Mehler's formula. For centered jointly Gaussian (G_1, G_2) with $\mathrm{Var}(G_i) = 1$ and $\mathrm{Cov}(G_1, G_2) = c$,

$$\mathbb{E}\big[\mathrm{He}_j(G_1)\mathrm{He}_k(G_2)\big] \;=\; j!\,\delta_{jk}\,c^j.$$

This identity evaluates Gaussian expectations of products of Hermite polynomials in correlated arguments and is used in the linearized growth-rate argument of Appendix H.

Stein's lemma (multivariate). For centered jointly Gaussian $(G_1, \ldots, G_n)$ and a smooth f, $\mathbb{E}[G_i\, f(G)] = \sum_j \mathrm{Cov}(G_i, G_j)\,\mathbb{E}[\partial_j f(G)]$. We use it to evaluate Gaussian expectations of the form $\mathbb{E}[G\, f(G)]$ in the drive kernels (Appendix D.6).

Derrida's Random Energy Model freezing. For E states with i.i.d. Gaussian energies $\mathcal{E}_e \sim \mathcal{N}(0,1)$ and partition function $Z(\beta) = \sum_e e^{-\beta\mathcal{E}_e}$, the annealed entropy $\ln E - \beta^2/2$ vanishes at $\beta_c = \sqrt{2\ln E}$, equivalently

$$T_c^{(0)} \;=\; 1/\sqrt{2\ln E},$$

which separates a paramagnetic (diffuse) regime from a frozen regime where Z is dominated by a few states (Derrida, 1981). Section 5 extends this to the signal-modified setting.

Proportional asymptotic limit. We take $n, d \to \infty$ with $\bar{\alpha} = n/d$ fixed, at finite width m. Parameter–data correlations vanish when every example is fresh ($n/d \to \infty$) but produce $\mathcal{O}(1)$ corrections under the proportional limit. These are the Onsager-type contributions corrected by the residual response R_A in the expert sector and by the router self-reaction in the router sector (Section 4.3).

MSRDJ generating functional. Gradient flow $\dot{\theta} = -\bar{\alpha}\nabla\mathcal{L}(\theta)$ is encoded as a path integral via response fields and Grassmann variables (Appendix C.1). Disorder averaging the data-dependent action and saddle-pointing in the proportional limit produces the closed Volterra system in terms of a small set of order parameters. The present paper applies this construction to a coupled router–experts gradient flow.

Gradient-flow equations of motion (full form). Gradient flow on the squared-loss empirical risk $\mathcal{L}(\theta) = \frac{1}{2n}\sum_\mu (y^\mu - f(x^\mu;\theta))^2$ with the spherical constraints $\|w_{l,e}\| = \|v_e\| = 1$ enforced by

Lagrange multipliers ν_w, ν_v reads

$$\dot{w}_{l,e} \;=\; \frac{\bar{a}}{md}\sum_{\mu} g_e(x^\mu)\,\mathrm{res}^\mu\,\sigma'(z^w_{l,e,\mu})\,x^\mu \;+\; \nu_w\,w_{l,e}, \tag{14}$$

$$\dot{v}_e \;=\; \frac{1}{d}\sum_{\mu}\mathrm{res}^\mu\,\Big(\sum_{j=1}^{E}\big(\partial_{v_e} g_j(x^\mu)\big)f_j(x^\mu)\Big) \;-\; \nu_v\,v_e, \tag{15}$$

$$\dot{a}_{l,e} \;=\; \frac{1}{md}\sum_{\mu}\mathrm{res}^\mu\,g_e(x^\mu)\,\sigma(z^w_{l,e,\mu}), \tag{16}$$

where $z^w_{l,e,\mu} = \langle w_{l,e}, x^\mu\rangle$ is the expert preactivation, $\mathrm{res}^\mu = y^\mu - f(x^\mu;\theta)$ is the residual, and $\bar{a}$ is the mean of the second-layer scalars. Lagrange multipliers are determined by $(\dot{w}_{l,e}\cdot w_{l,e}) = 0$ and likewise for v_e. All three updates carry the same $1/d$ prefactor: the dynamics are gradient flow on the *summed* squared loss $\frac{1}{2}\sum_\mu (y^\mu - f(x^\mu))^2$ with a single learning rate $1/d$, equivalently mean-loss gradient flow with learning rate $\bar{\alpha}$ for every parameter. This is the standard proportional-asymptotic convention: it makes the data term $\Theta(1)$ as $n, d \to \infty$ at fixed $\bar{\alpha} = n/d$, and it produces the $\bar{\alpha}$ prefactors of the order-parameter equations (87)–(90). The structural factors $1/m$ in (14) and (16) come from $\partial f_e/\partial w_{l,e}$ and $\partial f_e/\partial a_{l,e}$ respectively, not from the learning rate. The router gradient (15) carries a sum over *all* experts j, because a shift-invariant gate depends on v_e through the logit differences and so every g_j responds to v_e. For the $E = 2$ soft-discriminant router the sum collapses to a term proportional to $\mathrm{res}^\mu\,\big(f_e(x^\mu) - f_{e'}(x^\mu)\big)x^\mu$, which is the form used in both the simulator and the DMFT derivation.

C DMFT Derivation via MSRDJ

This appendix sketches the path-integral derivation that justifies the closed Volterra system of Section 4 and Appendix D. The derivation follows the Martin–Siggia–Rose–De Dominicis–Janssen (MSRDJ) formalism (Martin et al., 1973; De Dominicis, 1976; Janssen, 1976) specialized to coupled router–experts gradient flow on a Gaussian-mixture teacher in the proportional asymptotic limit $n, d \to \infty$ at $\bar{\alpha} = n/d$ fixed. We give the new ingredients that the MoE architecture requires (the routing function generates a non-Gaussian effective action, and the expert–router overlap appears as a new order parameter) and refer to standard references (Sompolinsky et al., 1988; Crisanti & Sompolinsky, 1987; Mignacco et al., 2020; Bordelon & Pehlevan, 2026) for the parts of the formalism shared with single-pathway DMFT.

C.1 MSRDJ generating functional and Grassmann variables

The MSRDJ formalism encodes gradient flow (14)–(16) as a path integral over response fields. We introduce response fields $\tilde{a}_{l,e}(t) \in \mathbb{R}$, $\tilde{w}_{l,e}(t) \in \mathbb{R}^d$, $\tilde{v}_e(t) \in \mathbb{R}^d$ and combine each weight with its response into a super-field indexed by the super-time coordinate $\hat{a} := (t_a, \vartheta_a)$ with Grassmann variable ϑ_a:

$$w_{l,e}(\hat{a}) \;:=\; w_{l,e}(t_a) + i\vartheta_a\,\tilde{w}_{l,e}(t_a), \qquad v_e(\hat{a}) \;:=\; v_e(t_a) + i\vartheta_a\,\tilde{v}_e(t_a), \tag{17}$$

and analogously for $a_{l,e}(\hat{a})$. The MSRDJ generating functional $Z_{\mathrm{dyn}} = \int \mathcal{D}[\theta]\exp(S_{\mathrm{dyn}})$ then has action $S_{\mathrm{dyn}} = S_{\mathrm{kin},W} + S_{\mathrm{kin},v} + S_{\mathrm{kin},a} + S_{\mathrm{int}}$ with kinetic pieces

$$\begin{aligned} S_{\mathrm{kin},W} &= \frac{d}{2}\sum_{l,e}\int d\hat{a}\,d\hat{b}\;K^W_{(l,e)}(\hat{a},\hat{b})\,\langle w_{l,e}(\hat{a}),\,w_{l,e}(\hat{b})\rangle,\\ S_{\mathrm{kin},v} &= \frac{d}{2}\sum_{e}\int d\hat{a}\,d\hat{b}\;K^v_{e}(\hat{a},\hat{b})\,\langle v_e(\hat{a}),\,v_e(\hat{b})\rangle,\\ S_{\mathrm{kin},a} &= \frac{d}{2}\sum_{l,e}\int d\hat{a}\,d\hat{b}\;\bar{K}_{(l,e)}(\hat{a},\hat{b})\,a_{l,e}(\hat{a})\,a_{l,e}(\hat{b}), \end{aligned} \tag{18}$$

where the kernels $K^W, K^v, \bar{K}$ encode the gradient-flow operator and the spherical-constraint Lagrange multipliers. The interaction piece is the empirical risk evaluated on the path,

$$S_{\mathrm{int}} \;=\; -n\int d\hat{a}\;\hat{\mathcal{R}}_n\big(\theta(\hat{a})\big) \;=\; -\bar{\alpha}d\int d\hat{a}\;\hat{\mathcal{R}}_n\big(\theta(\hat{a})\big), \tag{19}$$

with $\hat{\mathcal{R}}_n$ the empirical-risk functional of the super-fields. Relative to the kinetic terms, which are of order d, the factor $\bar{\alpha} = n/d$ sets the strength of the data coupling and plays the role of an inverse temperature.

C.2 Order parameters and saddle point

Disorder averaging the path integral against the data distribution $\{x^\mu, c^\mu\}$ generates a four-point coupling among the super-fields that we resolve by introducing six order-parameter super-fields:

$$\begin{aligned}
&\text{OP1: } Q_{(l,e),(l',e')}(\hat{a},\hat{b}) = \langle w_{l,e}(\hat{a}),\, w_{l',e'}(\hat{b})\rangle && \text{(expert autocorrelator)}\\
&\text{OP2: } Q^{vv}_{e,e'}(\hat{a},\hat{b}) = \langle v_e(\hat{a}),\, v_{e'}(\hat{b})\rangle && \text{(router autocorrelator)}\\
&\text{OP3: } P_{(l,e),c}(\hat{a}) = \langle w_{l,e}(\hat{a}),\, U_c\rangle && \text{(expert–teacher overlap)}\\
&\text{OP4: } P^{v}_{e,c}(\hat{a}) = \langle v_e(\hat{a}),\, U_c\rangle && \text{(router–teacher overlap)}\\
&\text{OP5: } Q^{v}_{(l,e),e'}(\hat{a},\hat{b}) = \langle w_{l,e}(\hat{a}),\, v_{e'}(\hat{b})\rangle && \text{(expert–router overlap)}\\
&\text{OP6: } a_e(\hat{a}) = m^{-1}\sum_l a_{l,e}(\hat{a}) && \text{(amplitude)}
\end{aligned} \tag{20}$$

The expert–router overlap (OP5) is the new order parameter introduced by the MoE architecture. It has no analog in single-pathway DMFT and gives rise to the expert–router covariance kernel $Q^{v}_{\mathrm{self/cross}}(t,s)$ of Appendix B.1.

At the dynamical saddle point $(\vartheta_a, \vartheta_b)$-dependent expansions of the order parameters yield two-time correlator and response functions:

$$Q_{ee'}(\hat{a},\hat{b}) \;=\; C_{ee'}(t_a,t_b) \;+\; \vartheta_a\, R_{e'e}(t_b,t_a) \;+\; \vartheta_b\, R_{ee'}(t_a,t_b), \tag{21}$$

with the causal constraint $R(t,s) = 0$ for $t < s$. The same decomposition holds for $Q^{vv}_{ee'}$, yielding the router correlator $C^{v}_{ee'}$ and response $R^{v}_{ee'}$. The inverse propagator is the 2×2 block

$$G^{-1}(\hat{a},\hat{b}) \;=\; \begin{pmatrix} K^W - \tilde{Q} & -\tilde{Q}^{v} \\ -(\tilde{Q}^{v})^\top & K^{v} - \tilde{Q}^{vv} \end{pmatrix}, \tag{22}$$

mixing expert and router super-fields through $\tilde{Q}^v$. Setting $\delta S_{\mathrm{dyn}}/\delta(\mathrm{OP}) = 0$ gives the Dyson equations whose two-time component reduces to the Volterra system of Appendix D.4.

Residual self-energies and Dyson equations. The residual response $R_A(t,s)$ and correlator $C_A(t,s)$ used as kernels throughout Appendix D are determined by saddle-point Dyson equations driven by the residual self-energies Σ_R and Σ_C:

$$\begin{aligned}
\Sigma_C(t,s) = \tau^2 + \mathbb{E}[\sigma(\lambda_c)^2] - \sum_{c,e} \pi_c\, \bar{a}_e(t)\, \hat{\Phi}_{c,e}(t) - \sum_{c,e} \pi_c\, \bar{a}_e(s)\, \hat{\Phi}_{c,e}(s) \\
+ \sum_{c,e,e'} \pi_c\, \bar{a}_e(t)\, \bar{a}_{e'}(s)\, H_c^{(e,e')}(t,s) \;+\; \Sigma_C^{\mathrm{ons}}(t,s),
\end{aligned} \tag{23}$$

$$\begin{aligned}
\Sigma_R(t,s) = -\sum_{c,e,e'} \pi_c\, \bar{a}_e(t)\, \bar{a}_{e'}(s) \Big[\tfrac{1}{m}\partial_{C_d} H \cdot R_d + \big(1 - \tfrac{1}{m}\big)\partial_{C_o} H \cdot R_o \\
+ \partial_{C_x^{(e,e')}} H \cdot R_x^{(e,e')} + \partial_{C_{dv}} H \cdot R_{dv} \\
+ \partial_{C_o^v} H \cdot R_o^v + \partial_{Q^v_{e,e'}} H \cdot R_{Q^v_{e,e'}} \Big],
\end{aligned} \tag{24}$$

where $H_c^{(e,e')}$ and $\hat{\Phi}_{c,e}$ are the effective kernels of (30)–(31) and the partial derivatives are taken with respect to the indicated entry of the underlying covariance matrix (30). The sum in (24) runs over *all* entries of that covariance: the same-expert weight blocks C_d (same hidden unit, weight $1/m$) and C_o (distinct units, weight $1 - 1/m$), the cross-expert block C_x, the same- and cross-expert router blocks C_{dv} and C_o^v, and the weight–router block Q^v. Evaluating (24) on the diagonal-S_E saddle, keeping

the same-hidden-unit channel of the expert weights and the router channel, and setting R_d and R^v_- to their equal-time value 1 gives the form our solver integrates,

$$\begin{aligned}\Sigma_R(t,s) \;=\; &-\; \frac{\bar a(t)\bar a(s)}{m}\Big[G_s\,H'_{ss}(C_d) + (E-1)\,G_c\,H'_{cc}(C_d)\Big] \\ &-\; a_1^2\,s_E^2\sum_e \frac{\mathbb{E}_\lambda\big[\delta F_e(t)\,\delta F_e(s)\,\mathrm{res}_\lambda(s)\big]}{\mathbb{E}_\lambda\big[\mathrm{res}_\lambda(s)\big]} \;-\; V_n(t,s),\end{aligned} \tag{25}$$

with $\delta F_e = F_e - \frac{1}{E}\sum_j F_j$, $s_E \equiv \sqrt{E/(E-1)}$ the discriminant normalization, G_s, G_c the two-time gate moments of (54), H'_{uv} the kernel shorthand of Appendix D.1, and all kernels evaluated at (t,s). The first term is the expert-weight channel, and the second the router channel, whose earlier-time factor is weighted by the in-plane residual res_λ of (59). The last, $V_n = a_1^2 s_E^2\,\frac{E-1}{mE}\,\bar a(t)\bar a(s)\big[H_{ss}(C_d) - H_{ss}(C_o) + H_{cc}(C_d) - H_{cc}(C_o)\big]$, in the shorthand of Appendix D.1, is the same-neuron fluctuation of δF_e at finite m. The R_o, R_x and R_{Q^v} channels of (24) are dropped. Here $R^v_- = R_{dv} - R_{vo}$ is the contrast response of Appendix E.3. Which of the first two channels dominates depends mainly on the expert width, because only the expert-weight channel carries the explicit $1/m$. Sweeping $E \in \{2,4,8\}$ against $T \in \{0.6, 1.0, 1.5\}$ at matched a_1, the router channel supplies 94 to 99% of Σ_R at $m = 1200$ but only 61 to 90% at $m = 120$, its share falling with both temperature and E. It therefore dominates at large width, whereas at the widths reachable in simulation the split is regime-dependent and both channels matter. The term Σ_C^{ons} is the Onsager contribution $\mathrm{Cov}(\eta_e, F_e)$: the closure treats the router's out-of-plane component η_e as independent of the expert output F_e, but example μ's own gradient moves v_e, giving η_e a self-reaction $\int_0^t R^v_-(t,\tau)\,c^v_e(\tau)d\tau$ that is correlated with F_e on the same example. With $\delta g_e = a_1 s_E\,\mathrm{acc}_e$, where acc_e is this self-reaction, this contributes

$$\begin{aligned}\Sigma_C^{\mathrm{ons}}(t,s) = &-\,a_1^2 s_E^2 E\sum_{u\in\{t,s\}}\int_0^u R^v_-(u,\tau)\,\mathbb{E}_\lambda\big[y\,F_e(u)\,K_e(\tau)\big]d\tau \\ &+\,2a_1^2 s_E^2 E\int_0^t R^v_-(t,\tau)\,\mathbb{E}_\lambda\big[f_{\mathrm{mf}}\,F_e(t)\,K_e(\tau)\big]d\tau,\end{aligned} \tag{26}$$

where $K_e = \delta F_e\,\mathrm{res}$ is the per-example router force and $f_{\mathrm{mf}} = \sum_e g_e F_e$ the mean-field student. The first term corrects $\hat\Phi$, entering Σ_C once at each time argument. The second is the cross term of $(f_{\mathrm{mf}} + \delta f)^2$ at equal times, hence the factor 2, and our solver adds it to the whole row $\Sigma_C(t,\cdot)$. The solver uses R_{dv} in place of R^v_- here, which differs from it by at most 3×10^{-4} at the settings of Figure 5, because R_{vo} has no δ-source and is driven only through the memory term $M^{vo}_R R_{dv}$, and it uses the rescaled force of (59). Without this term the modeled student output is 8.5% smaller and Σ_C 1.14% larger. In the student–student term of Σ_C the solver evaluates the out-of-plane router part with the equal-time router overlap, which is exact at $t = s$.

The Dyson integral equations

$$R_A(t,s) \;=\; \delta(t-s) \;+\; \int_s^t \Sigma_R(t,\tau)\,R_A(\tau,s)\,d\tau, \tag{27}$$

$$C_A(t,s) \;=\; \int_0^t\!\!\int_0^s R_A(t,\tau)\,\Sigma_C(\tau,\tau')\,R_A(s,\tau')\,d\tau'\,d\tau, \tag{28}$$

have the same retarded structure as the expert and router two-time Volterra equations of Appendix D.4 and are integrated by the same forward-Euler scheme. Under the diagonal-S_E reduction (Appendix C.7), Σ_R and Σ_C collapse to scalar two-time functions via the counting lemma (47), so R_A, C_A become scalar two-time objects governed by closed Volterra equations.

C.3 Preactivation statistics and effective kernels

For a training input $x^\mu = \kappa U_{c^\mu} + \xi^\mu$ from cluster c, the preactivations $z^w_e(t) = \langle w_{l,e}(t), x\rangle$, $z^v_e(t) = \langle v_e(t), x\rangle$, and the teacher-direction projection $\lambda_c = U_c^\top x$ are jointly Gaussian under the saddle-point ansatz. The means depend linearly on the overlaps,

$$\bar m^w_{e,c}(t) \;=\; \kappa\,P_{e,c}(t), \qquad \bar m^v_{e,c}(t) \;=\; \kappa\,P^v_{e,c}(t), \tag{29}$$

and the joint covariance of the expert preactivations $z_e^w(t), z_{e'}^w(s)$ and of the router preactivations $z^v(t), z^v(s)$ of all E experts is the matrix $\Sigma^{(e,e')}(t,s)$ whose entries are the order-parameter values at (t,s) (Appendix B.1). Under the diagonal-S_E reduction it takes one form per orbit of the counting lemma (47).

The effective student–student interaction kernel of the Dyson equation is

$$H_c^{(e,e')}(t,s) \;=\; \mathbb{E}_{\mathcal{N}(\bar{m},\,\Sigma^{(e,e')}(t,s))}\Big[\, g_e(z^v(t))\,\sigma(z_e^w(t))\; g_{e'}(z^v(s))\,\sigma(z_{e'}^w(s))\,\Big], \tag{30}$$

and the student–teacher kernel is the analogous Gaussian expectation against $\sigma(\lambda_c)$,

$$\hat{\Phi}_{c,e}(t) \;=\; \mathbb{E}_{\mathcal{N}(\bar{m},\,\Sigma^{\Phi}_{e,c}(t))}\Big[\, g_e(z^v(t))\,\sigma(z_e^w(t))\,\sigma(\lambda_c)\,\Big], \tag{31}$$

with the covariance built from the same overlaps together with the teacher overlaps $P_{e,c}$ and $P^v_{e,c}$. For the soft-discriminant gate, g_e depends on z^v only through disc_e, whose covariances are contrasts of the router overlaps. These are the building blocks of the memory kernels in Appendix D.2 and the signal sources in Appendix D.3. Under the diagonal-S_E reduction, $H_c^{(e,e')}$ takes one value per orbit of (47) and $\hat{\Phi}_{c,e}$ takes the two values $\hat{\Phi}_{\mathrm{self}}, \hat{\Phi}_{\mathrm{cross}}$ of Appendix D.3.

C.4 Cumulant expansion of the routing function

The MoE output $f(x) = \sum_e g_e(V^\top x)\, f_e(x)$ is non-Gaussian because the routing weights are nonlinear functions of the router preactivations. Following Montanari & Urbani (2025), the DMFT treats each example's residual as a Gaussian process with its exact mean and covariance. Conditioned on the router preactivations z^v of an example, we apply this treatment to the expert outputs. With the $1/m$ normalization of (2), f_e converges as $m \to \infty$ to a deterministic function of the Gaussian fields, $F_e = \bar{a}\,\sigma(u_e/q)$ of Appendix D.6, with Gaussian fluctuations of order $m^{-1/2}$, so this step is an approximation, accurate when F_e is close to linear in the fields. For an example μ with loss ℓ_μ, integrating out the conditional Gaussian residual yields, for fixed z^v,

$$\mathbb{E}_{F|z^v}\big[e^{-\int d\hat{a}\,\ell_\mu(\hat{a})}\big] \;=\; \exp\big(\,X_1[z^v]\,\big), \qquad X_1[z^v] \;=\; -\frac{1}{2}\,\mathrm{tr}\,\ln\big(I + \Sigma_{\mathrm{res}|z^v}\big), \tag{32}$$

where $\Sigma_{\mathrm{res}|z^v}(t,s)$ is the conditional covariance of the example's residual under the routing trajectory z^v, teacher and label-noise parts included. Its expert-pair part is $\sum_{e,e'} \bar{a}_e(t)\bar{a}_{e'}(s)\, g_e(t)\, g_{e'}(s)\, \mathbb{E}[\sigma(z_e^w(t))\,\sigma(z_{e'}^w(s)) \mid z^v]$. The examples are i.i.d., each with its own routing trajectory, so the data average factorizes into n identical factors, and the routing randomness is integrated by the per-example cumulant generating function

$$\mathcal{F} \;:=\; \ln \mathbb{E}_{z^v}\big[e^{X_1[z^v]}\big] \;=\; \sum_{\ell \geq 1} \frac{\kappa_\ell(X_1)}{\ell!}, \tag{33}$$

which enters the effective action as $n\mathcal{F} = \bar{\alpha} d\,\mathcal{F}$, with $\kappa_\ell(X_1)$ the ℓ-th cumulant. The first cumulant $\kappa_1(X_1) = \mathbb{E}_{z^v}[X_1] = -\frac{1}{2}\,\mathbb{E}_{z^v}[\mathrm{tr}\,\ln(I + \Sigma_{\mathrm{res}|z^v})]$ is *not* the same as $-\frac{1}{2}\,\mathrm{tr}\,\ln(I + \mathbb{E}_{z^v}[\Sigma])$: by Jensen's inequality applied to the concave $\ln\det$,

$$\mathbb{E}_{z^v}[\mathrm{tr}\,\ln(I+\Sigma)] - \mathrm{tr}\,\ln(I + \mathbb{E}_{z^v}[\Sigma]) \;=\; -\tfrac{1}{2}\,\mathbb{E}_{z^v}\mathrm{tr}\big[\big((I+\bar{\Sigma})^{-1}\delta\Sigma\big)^2\big] \;+\; \mathcal{O}(\delta\Sigma^3), \tag{34}$$

where $\bar{\Sigma} = \mathbb{E}_{z^v}[\Sigma]$ and $\delta\Sigma = \Sigma - \bar{\Sigma}$. This follows from $\mathrm{tr}\ln(I+\Sigma) = \mathrm{tr}\ln(I+\bar{\Sigma}) + \mathrm{tr}\ln(I+M)$ with $M = (I+\bar{\Sigma})^{-1}\delta\Sigma$, expanding $\mathrm{tr}\ln(I+M) = \mathrm{tr}\,M - \frac{1}{2}\,\mathrm{tr}\,M^2 + \ldots$ and using $\mathbb{E}[M] = 0$. The gap is negative because $\mathrm{tr}\ln$ is concave, consistent with Jensen. It is the trace of a *squared matrix*, not the variance of the trace: the two agree when $\delta\Sigma$ has rank one. It is $\mathcal{O}(1)$ in m and sits inside $\kappa_1 = \mathbb{E}_{z^v}[X_1]$. The DMFT evaluates the log-determinant at the mean covariance, $-(n/2)\,\mathrm{tr}\ln(I+\bar{\Sigma})$, together with Gaussian averages over the routing fields at the tracked two-point functions. We call this the second-cumulant closure because it keeps the mean and covariance of those fields. It is a closure and not a truncation controlled by a small parameter: every cumulant $\kappa_\ell(X_1)$ of a single example is of order one, and the accuracy of the closure is established by comparison with gradient flow (Section 4.4).

Truncation validity. Two further parameters affect how strongly the routing fluctuates, and so how much the neglected structure matters. They are heuristics, not controlled small parameters: the error that remains with the gate means and drives evaluated exactly grows as T falls (Section 4.4).

1. **Inverse number of experts $1/E$.** As E grows the routing distribution self-averages. For simplex-valued routers ($\sum_e g_e = 1$, $g_e \geq 0$: softmax, hard top-k), the participation ratio $Y_2 = \sum_e g_e^2$ concentrates above $T_c^{(0)}$, where $\mathrm{Var}_x[Y_2] \to 0$ as E grows. Direct sampling of the softmax router gives $\mathrm{Var}_x[Y_2] = \mathcal{O}(1/E^3)$ at fixed T above $T_c^{(0)}$: the ratio per doubling of E approaches $1/8$, by $E \approx 256$ at $T = 1.5$ and 2 and more slowly at $T = 1$. Below $T_c^{(0)}$, Y_2 stays of order one and its variance does not vanish. For the linear-disc surrogate $g_e = 1/E + a_1 \mathrm{disc}_e$ used in the solvable model, Y_2 is not a participation ratio and its *absolute* variance grows with E. There the corresponding statement is that the relative variance $\mathrm{Var}[Y_2]/\mathbb{E}[Y_2]^2$ is $\mathcal{O}(1/E)$, which we observe for $E \gtrsim 16$.
2. **Inverse router temperature $1/T$.** For softmax routing, $\mathrm{Var}_z[g(z/T)] = \mathcal{O}(1/T^2)$, so the gate fluctuations that feed the neglected structure shrink as T grows.

Validity during specialization. The weak-alignment condition of (7) fails during specialization, as P^v grows from $\mathcal{O}(\varepsilon^v)$ to $\mathcal{O}(1)$, so the worst case for the closure is the regime the paper aims to describe. In the specialization regime of Section 5.2 the router condenses onto the matched expert, the participation ratio approaches 1, and the routing average collapses onto an effectively deterministic gate, which can make the closure accurate again although the router is no longer weakly aligned. Sampling supports this for the gate itself: at fixed E and T_{eff}, the variance of Y_2 over inputs decays as

$$\mathrm{Var}_x(Y_2) \;\propto\; \exp\!\big(-c\, s(t)^2\big) \qquad \text{at large } s, \tag{35}$$

where $s = \kappa\,\Delta_v/\sqrt{1 - C_o^v}$ is the matched expert's head start of (11), with c between 0.08 and 0.21 at $E = 4, 8$ and $T_{\text{eff}} \in \{0.3, 0.7\}$. Matched condensation needs $\kappa \geq \sqrt{2\ln(E-1)}$, so this mechanism is absent from our $\kappa = 1$ validation runs at $E \geq 3$. The regime $\Delta(t) \sim \mathcal{O}(1)$ without matched condensation is where the closure can fail. Two router classes avoid part of this failure mode. For a terminating gate Hermite expansion (linear and quadratic-probit, Appendix H) the gate's own non-Gaussianity is at most a finite polynomial, evaluated exactly. Hard top-k carries an infinite odd-order tail (Figure 10), but its gate is an indicator, so its participation ratio equals $1/k$ for every input. The covariance Σ still fluctuates with which experts are selected, so this removes one source of fluctuation, not all of them. Only for softmax can the gate's own fluctuations grow. The residual higher cumulants of the effective action remain, and are bounded numerically rather than analytically (Section 4.4).

Catalog of routers and their gate moments. Table 2 summarizes the four routing functions analyzed in the paper and how their gate moments enter the closure.

Routing function r	Gate Hermite series	Gate moments	Notes
Linear: $a_0 + a_1\,\mathrm{disc}_e(x)$	terminates at $k = 1$	closed-form Gaussian	used for the DMFT–GF validation
Quadratic-probit: $a_0 + a_1 z + a_2(z^2{-}1)$	terminates at $k = 2$	closed-form Gaussian	no normalized measure over experts, so no REM freezing
Softmax-T: $\mathrm{smax}_T(V^\top x)_e$	infinite series	exact reduction (Appendix C.5)	DMFT error grows as T falls (Section 4.4)
Hard top-k: $\mathbf{1}\{e \in \mathrm{top}_k\}/k$	discrete	exact finite mixture	$Y_2 = 1/k$ for every input

Table 2: Routing functions and how their gate moments are evaluated: as finite polynomials in Gaussian fields when the gate Hermite series terminates (linear and quadratic-probit), as an exact finite mixture over selected sets when each gate takes only the values 0 and $1/k$ (hard top-k), and by the reduction of Appendix C.5 for softmax. In every case the effective action keeps the non-polynomial $\mathrm{tr}\log$ of the data average, so its cumulant series does not terminate and all four routers use the same second-cumulant closure, expressed in the same order parameters.

Quadratic-probit and REM freezing. REM freezing is a property of a normalized Boltzmann measure over experts: the softmax weights $g_e \propto e^{z_e^v/T}$ compete through their normalizer, which is the REM partition function. The quadratic-probit gate $r_e = a_0 + a_1 z_e^v + a_2((z_e^v)^2 - 1)$ is a polynomial in each expert's own logit with no normalization across experts, so it has no partition function over experts and no freezing transition. Its effective action still contains the non-polynomial $\mathrm{tr}\log$ of the data average, so the quadratic-probit DMFT uses the same second-cumulant closure as the other routers.

C.5 Exact softmax gate reduction at arbitrary E

The cumulant truncation above is a statement about the routing-induced effective action. The gate *moments* that enter every kernel admit a separate, exact treatment at any temperature, which lets the Volterra system be evaluated with softmax gate moments rather than those of its high-temperature surrogate.

For data from cluster c the router preactivation is $z_e^v = \kappa P_{e,c}^v + v_e \cdot \xi$. Up to a shift common to all experts, which cancels in the translation-invariant softmax, the exchangeable noise equals $\sqrt{1 - C_o^v}\,\eta_e$ with η i.i.d. standard normal. This holds for any $C_o^v \geq -1/(E-1)$: the form $\sqrt{(1 + (E-1)C_o^v)/E}\,\omega_0 + \sqrt{1 - C_o^v}\,(\eta_e - \bar{\eta})$, with ω_0 standard normal and $\bar{\eta} = \frac{1}{E}\sum_j \eta_j$, has the correct covariance, and both ω_0 and $\bar{\eta}$ are common to all experts. With

$$\beta \;=\; \frac{\sqrt{1 - C_o^v}}{T}, \qquad s \;=\; \frac{\kappa\,\Delta_v}{\sqrt{1 - C_o^v}}, \qquad \Delta_v = P_{\mathrm{self}}^v - P_{\mathrm{cross}}^v, \tag{36}$$

the gate is exactly a Random Energy Model in a field, $g_e = \mathrm{softmax}_e\big(\beta(s\,\delta_{ec} + \eta_e)\big)$. Writing $A = e^{\beta(s+\eta_c)}$ for the matched expert and $S = \sum_{j \neq c} e^{\beta\eta_j}$ for the mismatched bulk, A and S are independent, and

$$\frac{1}{(A+S)^p} \;=\; \frac{1}{(p-1)!}\int_0^\infty u^{p-1} e^{-u(A+S)}\,du, \qquad \mathbb{E}\big[e^{-uS}\big] \;=\; \psi(u)^{E-1}, \tag{37}$$

since S is a sum of i.i.d. lognormals. Every gate moment therefore reduces to two nested one-dimensional integrals in which E *appears only as an exponent*: for instance $\mathbb{E}[g_c] = \int \chi_A(u)\,\psi(u)^{E-1} du$ with $\chi_A(u) = \mathbb{E}[A e^{-uA}]$, and $\mathbb{E}[g_c^2] = \int u\,q_A(u)\,\psi(u)^{E-1} du$ with $q_A(u) = \mathbb{E}[A^2 e^{-uA}]$. The representation is exact for every β, so it holds *below* the freezing temperature as well as above. Only the u-grid must be widened as T falls. Higher moments follow from (37) at $p = 3, 4$, and each numerator factor consumes one expert while every remaining expert contributes its ψ.

Two-time gate correlators reduce the same way, with one auxiliary variable per time. Because the matched expert's noise is independent of the bulk's and the bulk experts are i.i.d.,

$$\begin{aligned}\mathbb{E}\big[g_c(t)g_c(s)\big] &= \iint \chi_{AA}(u,v)\,\psi_B(u,v)^{E-1}\,du\,dv, \\ \mathbb{E}\big[g_m(t)g_m(s)\big] &= \iint \chi_{BB}(u,v)\,\psi_A(u,v)\,\psi_B(u,v)^{E-2}\,du\,dv,\end{aligned} \tag{38}$$

and the mixed pair $\mathbb{E}[g_c(t)g_m(s)]$ by the same construction with one-sided numerators. Each kernel is a bivariate Gaussian expectation over one expert's noise at the two times. Writing $\eta' = r\eta + \sqrt{1-r^2}\,\omega$, with ω an independent standard normal and $r = (C_d^v - C_o^v)/\sqrt{(1 - C_o^v(t,t))(1 - C_o^v(s,s))}$ turns each into nested one-dimensional sums. The two-time sector needs five S_E orbits: a matched gate paired with itself, a mismatched gate paired with itself, the two mixed pairs, and, for $E \geq 3$, two distinct mismatched gates. They are not interchangeable: $\mathbb{E}[g_c^2]$ and $\mathbb{E}[g_m^2]$ differ by orders of magnitude once routing condenses.

The linear-discriminant gate is softmax's leading high-temperature term. Expanding at small z_e/T gives $g_e \simeq 1/E + (z_e - \bar{z})/(ET)$, which is the linear-discriminant gate $g_e = 1/E + a_1 s_E (z_e - \bar{z})$ at

$$a_1 \;=\; \frac{1}{E\,T\,s_E} \;=\; \frac{\sqrt{E-1}}{T\,E^{3/2}}, \qquad s_E = \sqrt{\tfrac{E}{E-1}}. \tag{39}$$

This is the matched coupling used throughout Section 4.4: the linear router is not a separate model but the first order of a controlled expansion of softmax, with the $\mathcal{O}(1/T^2)$ term as the bridge error, $\mathcal{O}(1/T^3)$ at $E = 2$, where softmax is $\frac{1}{2} + \frac{1}{2}\tanh(\Delta z/2T)$ and the first correction cancels by oddness.

The router gradient differs in form, not only in coefficient. The two gates give

$$\text{linear:}\quad \partial_{v_e} f = a_1 s_E\,(F_e - \bar{F})\,x, \qquad \text{softmax:}\quad \partial_{v_e} f = \tfrac{1}{T}\,g_e\,(F_e - f)\,x, \tag{40}$$

with $f = \sum_j g_j F_j$. The softmax form carries a state-dependent coefficient g_e and subtracts the *gate-weighted* mean f rather than the uniform mean $\bar{F}$. Both differences grow as routing condenses, and neither can be absorbed into a rescaling of a_1. Expanding the residual and f inside $\mathbb{E}[\mathrm{res}\, g_e(F_e - f)\lambda]$ produces terms carrying gate moments up to third order, each multiplying an expert-only kernel of the H_μ family. The gate factorizes from the expert sector up to $\mathcal{O}(P^v, Q^v)$, and the leading correction is the Gaussian cross-covariance (Price) term. Both gates satisfy the shift-invariance sum rule $\sum_e \partial_{v_e} f = 0$ pointwise in x, and hence $\sum_e \mathbb{E}[\partial_{v_e} f\,\lambda_j] = 0$ for every j, which makes the router multiplier (96) well defined at any E.

C.6 Exact hard top-k gate reduction

Write the router preactivations in the reduced form used for the softmax gate, $\tilde{z}_e = s\,\delta_{ec} + \eta_e$ with $\eta \sim \mathcal{N}(0, I_E)$ and $s = \kappa\Delta_v/\sqrt{1 - C_o^v}$, and take $g_e = \frac{1}{k}\mathbb{1}[e \in \text{top-}k]$ so that $\sum_e g_e = 1$, as for the softmax, soft-discriminant and cubic gates (the soft-discriminant gate can leave $[0, 1]$, and the cubic gate is clipped to it). The quadratic-probit gate as written is not normalized. Since argmax is invariant under a common rescaling of $\tilde{z}$, no temperature appears.

Single-time moments. Conditioning on the candidate's own noise makes the rivals conditionally independent, so "e is among the k largest" is a binomial tail:

$$\begin{aligned}\mathbb{P}(c \in \text{top-}k \mid \eta_c = u) &= \mathbb{P}\big(\mathrm{Bin}(E-1,\,1-\Phi(s+u)) \le k-1\big), \\ \mathbb{P}(m \in \text{top-}k \mid \eta_m = u) &= \Phi(u-s)\,\mathbb{P}\big(\mathrm{Bin}(E-2,\,1-\Phi(u)) \le k-1\big) \\ &\quad + \big(1-\Phi(u-s)\big)\,\mathbb{P}\big(\mathrm{Bin}(E-2,\,1-\Phi(u)) \le k-2\big),\end{aligned} \tag{41}$$

the mismatched case splitting on whether the matched expert outranks it. At $k = 1$ these reduce to $\mathbb{E}[g_c] = \int \phi(u)\Phi(s+u)^{E-1}du$ and $\mathbb{E}[g_m] = \int \phi(u)\Phi(u-s)\Phi(u)^{E-2}du$. Because the hard gate is a scaled indicator, $g_e^2 = g_e/k$, so the participation ratio is pinned:

$$Y_2 \;=\; \sum_e \mathbb{E}[g_e^2] \;=\; \frac{1}{k}\sum_e \mathbb{E}[g_e] \;=\; \frac{1}{k} \tag{42}$$

identically, for every E, every signal strength and every router state. The sum rule $\mathbb{E}[g_c] + (E-1)\mathbb{E}[g_m] = 1$ is the statement that some expert wins, and $g_e^2 = \frac{1}{k}g_e$ gives (42) directly.

Two-time orbits. For $k = 1$, with $\eta_j(t), \eta_j(s)$ jointly Gaussian of correlation r and independent across j, and signals s_1, s_2 at the two times, conditioning on the winner's noise and requiring every rival to lose at *both* times gives, with $\Phi_2(\cdot,\cdot;r)$ the bivariate normal CDF,

$$\begin{aligned}\mathrm{cc} &= \iint \phi_r(u,v)\,\Phi_2(s_1{+}u,\; s_2{+}v; r)^{E-1}\,du\,dv, \\ \mathrm{mm} &= \iint \phi_r(u,v)\,\Phi_2(u{-}s_1,\; v{-}s_2; r)\,\Phi_2(u,v;r)^{E-2}\,du\,dv.\end{aligned} \tag{43}$$

The mixed orbits involve two distinct correlated pairs and appear four-dimensional, but each loser enters through a single inequality (the matched expert's second-time noise only through $v < q - s_2$, the mismatched expert's first-time noise only through $p < s_1 + u$), and both integrate out in closed form, leaving

$$\mathrm{cm} = \iint \phi(u)\phi(q)\,\Phi\Big(\tfrac{s_1+u-rq}{\sqrt{1-r^2}}\Big)\Phi\Big(\tfrac{q-s_2-ru}{\sqrt{1-r^2}}\Big)\Phi_2(s_1{+}u,\; q; r)^{E-2}\,du\,dq, \tag{44}$$

at the same two-dimensional cost as the others. A fifth orbit, in which a mismatched expert wins at t and a different one at s, has the same structure and is required for the sum rules $\mathrm{cc}{+}(E{-}1)\mathrm{cm} = \mathbb{E}[g_c]$ and $\mathrm{mc} + \mathrm{mm} + (E-2)\mathrm{mx} = \mathbb{E}[g_m]$ to close. Those two identities, holding to machine precision between integrands of different structure, serve as our check on this sector.

Boundary density and the self-reaction. For the argmax the response is $\partial g_c/\partial \tilde{z}_c = \sum_{j\neq c} \delta(\tilde{z}_c - \tilde{z}_j) \prod_{k\neq c,j} \mathbb{1}[\tilde{z}_c > \tilde{z}_k]$ and $\partial g_c/\partial \tilde{z}_m = -\delta(\tilde{z}_c - \tilde{z}_m) \prod_{k\neq c,m} \mathbb{1}[\tilde{z}_c > \tilde{z}_k]$. Against the Gaussian noise the delta collapses onto the boundary, giving one-dimensional integrals such as

$$\chi_{cc} = (E-1) \int \phi(u)\, \phi(s+u)\, \Phi(s+u)^{E-2}\, du, \tag{45}$$

with the remaining orbits obtained the same way. Two identities fix these: the row sums vanish, $\chi_{cc} + (E-1)\chi_{cm} = 0$, which expresses shift invariance and makes the Onsager term well posed, and the diagonal entries are the derivatives of the single-time moments with respect to the signal, $\chi_{cc} = d\,\mathbb{E}[g_c]/ds$, tying this sector to (41).

Expert-drive grid. Inside the expert-drive quadrature the gate is needed pointwise in $(\lambda_0, \lambda_1, \lambda_x)$. Expanding $z_e^v = P^v_{\text{cross}}\lambda_{\text{tot}} + \Delta_v \lambda_e + \text{perp}_e$, the common $P^v_{\text{cross}}\lambda_{\text{tot}}$ cancels in the argmax exactly as it does in the softmax, leaving independent per-expert noise of variance $\varsigma^2 = 1 - C_o^v - (\Delta_v)^2$. Scale invariance then divides through by ς: with $b_e = \Delta_v \lambda_e/\varsigma$, the top-1 gate is

$$\mathbb{E}[g_0 \mid \lambda] = \int \phi(u)\, \Phi(u + b_0 - b_1)\, \Phi(u + b_0 - b_x)^{E-2}\, du, \tag{46}$$

which is exact for $E \leq 3$. At $E > 3$ we evaluate the $E-3$ unresolved experts at the resolved node λ_x in place of averaging over their independent fields, an approximation shared by the expert-drive quadrature of every router at $E > 3$. At $k = 1$ the second moments needed by the quadratic part of the drive are free: $\mathbb{E}[g_0^2 \mid \lambda] = \mathbb{E}[g_0 \mid \lambda]$ and $\mathbb{E}[g_0 g_e \mid \lambda] = 0$ for $e \neq 0$, where the softmax requires a separate reduction for the same objects.

C.7 Diagonal-S_E reduction

The loss, gradient flow, and data distribution are jointly invariant under the product symmetry group $S_E^{\text{experts}} \times S_E^{\text{clusters}}$ acting independently on expert and cluster labels. Under this symmetry, all E^2 expert–cluster overlap entries $P_{e,c}$ are equivalent. The symmetric initialization (3) explicitly breaks $S_E^{\text{experts}} \times S_E^{\text{clusters}}$ down to the diagonal subgroup S_E^{diag} that simultaneously permutes expert e and its assigned cluster e. The $\mathcal{O}(\varepsilon)$ initial bias is the seed from which $\Delta(t) = P_{\text{self}}(t) - P_{\text{cross}}(t)$ is amplified. Because in our longest runs the amplification stops at a finite gain (Appendix J), the bias decides how far specialization proceeds, as well as which expert takes which cluster. The growth rate $\Lambda(t)$ sets how far the seed is lifted, and below $T_c^{(0)}$ router freezing competes with it (Appendix H.3).

Under the diagonal S_E^{diag} symmetry, every $E \times E$ correlator matrix collapses to two scalars, distinguished only by whether the expert and cluster indices match (Appendix B.1):

$$\begin{aligned} &\{P_{e,c},\, P^v_{e,c},\, Q^v_{e,e'}(t,s),\, C_x^{(e,e')}(t,s),\, \dots\} \\ &\qquad \longrightarrow \; \{P_{\text{self}}, P_{\text{cross}};\; P^v_{\text{self}}, P^v_{\text{cross}};\; Q^v_{\text{self}}, Q^v_{\text{cross}};\; C_x;\; \dots\}. \end{aligned}$$

The triple sum $\sum_{c,e,e'} \pi_c(\dots)$ in the saddle-point Dyson equation for the residual self-energies decomposes via the counting lemma

$$\begin{aligned} \sum_{c,e,e'} \pi_c\, (\dots)_{c,e,e'} \;=\; \pi \Big[& E\, (\dots)_{c=e=e'} \;+\; E(E{-}1) \big[(\dots)_{c=e\neq e'} + (\dots)_{c=e'\neq e} + (\dots)_{e=e'\neq c} \big] \\ & +\; E(E{-}1)(E{-}2)\, (\dots)_{c\neq e\neq e'\neq c} \Big], \end{aligned} \tag{47}$$

with equal cluster priors $\pi = 1/E$. The five terms are the orbits of the index triple (c, e, e') under the diagonal S_E action, and their multiplicities sum to $E + 3E(E{-}1) + E(E{-}1)(E{-}2) = E^3$ as required. The all-distinct orbit is empty for $E \leq 2$, which is why the $E = 2$ reduction retains only the first four. The general-E solver evaluates all five. Under the further symmetric reduction used for the numerics, the orbits collapse to two independent scalar channels (matched and mismatched), with explicit prefactors. The E^2-dimensional correlator system therefore reduces to an $\mathcal{O}(1)$ scalar system with the same dimensionality as standard single-expert DMFT, with the cluster sum contributing only an outer loop over E terms in the prefactors of Appendix D.5. This reduction makes the DMFT tractable at any E and underlies all DMFT–GF comparisons in Section 4.4.

D COMPLETE VOLTERRA SYSTEM AND EQUAL-TIME CONSTRAINTS

This appendix gives the equations of motion that our solver integrates for the order parameters of Appendix B.1. We write the system at general E and give the $E = 2$ form where it simplifies. Numerical comparisons span $E \in \{2, 4, 8\}$. Several kernels, mostly in the router sector, are closures rather than derivations, and we say so where they appear. The equations are displayed as derived, but our integration keeps the $\mathcal{R}_A$-based memory of the expert and router kernels only at order h in the Euler step h, a closure choice (Appendix D.4). Several further choices were made by comparison with gradient flow at $E = 2$: the router memory kernels (57)–(58) together with the step $h = 0.1$ at which they act (Appendix D.2), the pairing and sign of the noise terms in the two-time equations (Appendix D.7), the closed form of ν_w (Appendix D.6), and leaving the amplitude kernels without the mean routing shift (Appendix E.5). The rescaling $\lambda(t)$ of the router force in (59), with its floor, is a heuristic of the closure. Appendix D.8 lists the approximations of our implementation at $E > 2$. Unless stated otherwise the equations use the soft-discriminant router $r_e(x) = 1/E + a_1 \operatorname{disc}_e(x)$ with $\operatorname{disc}_e(x) = s_E(v_e - \bar{v})\cdot x$ (equal to $(v_e - v_{e'})\cdot x/\sqrt{2}$ at $E = 2$) and the activation $\sigma(z) = \operatorname{erf}(z/\sqrt{2})$ for both teacher and student.

D.1 GAUSSIAN KERNELS

The signal and memory kernels are built from two-dimensional Gaussian expectations. Let (G_1, G_2) be jointly standard Gaussian with $\operatorname{Var}(G_i) = 1$ and $\operatorname{Cov}(G_1, G_2) = c$. For mean shifts m_1, m_2, define

$$H_\mu(c, m_1, m_2) := \mathbb{E}\big[\sigma(m_1 + G_1)\,\sigma(m_2 + G_2)\big], \tag{48}$$

$$H^{\mathrm{lam}}_\mu(c, m_1, m_2) := \mathbb{E}\big[\sigma(m_1 + G_1)\,\sigma(m_2 + G_2)\,(m_2 + G_2)\big], \tag{49}$$

$$H'_\mu(c, m_1, m_2) := \mathbb{E}\big[\sigma'(m_1 + G_1)\,\sigma'(m_2 + G_2)\big], \tag{50}$$

$$\partial_{m_1} H_\mu(c, m_1, m_2) = \mathbb{E}\big[\sigma'(m_1 + G_1)\,\sigma(m_2 + G_2)\big], \tag{51}$$

$$\partial_{m_1} H^{\mathrm{lam}}_\mu(c, m_1, m_2) = \mathbb{E}\big[\sigma'(m_1 + G_1)\,\sigma(m_2 + G_2)\,(m_2 + G_2)\big], \tag{52}$$

and analogously $\partial_{m_2} H_\mu$. The first argument c is the cross-correlation of the underlying Gaussians. In the kernels of subsequent subsections it is one of the two-time correlators C_d, C_o, C_x from Appendix B.1 or an expert–teacher overlap $P_{\mathrm{self/cross}}$. The mean-shift arguments take three canonical values:

$$m_s(t) := \kappa\, P_{\mathrm{self}}(t), \qquad m_c(t) := \kappa\, P_{\mathrm{cross}}(t), \qquad m_\lambda := \kappa. \tag{53}$$

For two-time kernels we write $H_{uv}(C) := H_\mu\big(C(t,s),\, m_u(t),\, m_v(s)\big)$ for $u, v \in \{s, c\}$, and likewise $H'_{uv}(C)$ and $\partial_{m_1} H_{uv}(C)$. We evaluate (48)–(52) by 64-node Gauss–Hermite quadrature over G_1, with G_2 given G_1 integrated in closed form. The expert drive is evaluated on tensor Gauss–Hermite grids (20 nodes per dimension) over the teacher coordinates of the matched cluster, one mismatched cluster and a third expert's cluster, and for $E \geq 4$ the sum of the remaining coordinates, with the clusters matched to neither expert treated as in Appendix D.8. Its self-reaction uses the same grids with 10 nodes at $E \leq 3$, and at $E \geq 4$ a fixed sample of 2^{13} scrambled-Sobol draws of all E teacher coordinates, which reproduces the $E = 2$ and 3 quadratures to 0.2%. The δ_{disc} source uses a 24-node grid of the same kind. The mean-field part of Σ_C uses a 24×24 Gauss–Hermite grid. The router drive, the population part of ν_v and the router-force Gram entering ν_v and M^{vd}_C are Monte Carlo averages over fixed samples of 4×10^5 and 4×10^4 antithetic Gaussian draws. For the soft-discriminant and quadratic-probit routers the gate enters polynomially, so the router-variable integration is exact at finite order.

The expert-sector kernels carry the cluster prior $1/E$ and a gate-weighted sum over the matched and mismatched cluster channels, with the two-time gate moments

$$G_s(t,s) = \bar{g}_s(t)\bar{g}_s(s) + a_1^2\,(C_{dv} - C_{vo})(t,s), \qquad G_c(t,s) = \bar{g}_c(t)\bar{g}_c(s) + a_1^2\,(C_{dv} - C_{vo})(t,s), \tag{54}$$

where $\bar{g}_s = 1/E + a_1\kappa r_D$ and $\bar{g}_c = 1/E - a_1\kappa r_D/(E-1)$ are the gate means of the matched and a mismatched expert (62), and m_s or m_c of (53) is the corresponding preactivation mean. Both time arguments are evaluated in the *same* cluster channel. The kernel is $H'_\mu = \mathbb{E}[\sigma'\sigma']$, not H_μ: the self-energy is generated by $\partial\,\mathrm{res}(s)/\partial z^w(s) = -(\bar{a}/m) g_e \sigma'(z^w(s))$ acting on a force already proportional to $g_e\sigma'(z^w(t))$, so both slots carry a derivative. The correlator partners M^d_C, M^o_C

are obtained by replacing $\mathcal{R}_A \to \mathcal{C}_A$ throughout. The explicit $1/m$ of the expert sector appears in (70)–(71), not inside these kernels.

Extended-covariance partial derivatives. Several memory kernels in the next subsection involve derivatives like $\partial_{C_{vo}} H_\mu^{sc,o}$ and $\partial_{Q^v_{\text{self/cross}}(t,s)} H_\mu(\dots)$. These are taken with respect to the corresponding entry of the joint covariance of the expert and router preactivations used in the saddle-point evaluation (Appendix C.3, eq. (30)). When the only relevant cross-correlation is the single scalar c of (48)–(52) we suppress the larger covariance argument. Equivalently, $\partial_{(\bullet)}$ is the partial derivative with respect to the labeled entry of the full covariance $\Sigma^{(e,e')}(t,s)$.

D.2 Memory kernels

The subscripts on the kernels and on the two-time correlators of Appendix B.1 follow a uniform convention: *d* (diagonal) for self-correlation between a hidden unit and itself, *o* (off-diagonal) for correlation between two distinct hidden units within the same expert, and *x* (cross) for correlation between hidden units in different experts. The expert-sector self-energies are products of the residual response $\mathcal{R}_A(t,s)$ (or correlator $\mathcal{C}_A(t,s)$) with Hessian slices of the gate-weighted activation kernel,

$$M_R^d(t,s) = \frac{\bar\alpha}{E}\,\mathcal{R}_A(t,s)\big[G_s(t,s)\,H'_{ss}(C_d) + G_c(t,s)\,H'_{cc}(C_d)\big], \tag{55}$$

$$M_R^o(t,s) = \frac{\bar\alpha}{E}\,\mathcal{R}_A(t,s)\big[G_s(t,s)\,H'_{ss}(C_o) + G_c(t,s)\,H'_{cc}(C_o)\big], \tag{56}$$

with $m_s = \kappa P_{\text{self}}$ and $m_c = \kappa P_{\text{cross}}$. At $E = 2$ the two terms are the full cluster sum. The noise partners M_C^d, M_C^o replace $\mathcal{R}_A$ by $\mathcal{C}_A$.

The router kernels are closures:

$$M_R^{vd}(t,s) = \bar\alpha\,\frac{a_1^2}{E^2}\,\mathcal{R}_A(t,s)\,\big[H_{ss}(C_d) + H_{cc}(C_d)\big], \tag{57}$$

$$M_R^{vo}(t,s) = \bar\alpha\,\frac{a_1^2}{E^2(E-1)}\,\mathcal{R}_A(t,s)\,\big[H_{sc}(C_x) + H_{cs}(C_x)\big]. \tag{58}$$

(57) uses the same-neuron C_d and two cluster types, and (58) pairs a matched and a mismatched arm of two distinct experts. Each kernel refers to one expert or one pair of experts, and the explicit multiplicities where they enter, such as the $(E-1)$ in (78), count the partners. Both enter our integration through the cell-integral form of $\mathcal{R}_A$ (85) and so contribute at order h. At $h = 0.1$ the resulting contrast response agrees with gradient flow on $1 - R_-^v$ at $E = 2$ to within 2%, an agreement that depends on the step, since this contribution vanishes as $h \to 0$.

The router noise kernel of a single expert is the router-force correlator,

$$M_C^{vd}(t,s) \;=\; \bar\alpha\,\lambda(t)\lambda(s)\,\mathbb{E}_\mu\big[c_e^v(t)\,c_e^v(s)\big], \qquad c_e^v = a_1 s_E\,\delta F_e\,\text{res}_\lambda, \tag{59}$$

a four-point expectation that does *not* factorize into $\mathbb{E}[\delta F\,\delta F]\,\mathbb{E}[\text{res}\,\text{res}]$. Here res_λ is the residual with the router's out-of-plane noise dropped, and the factor $\lambda(t)^2 = \max\big(1 - 2E\,\Delta\nu_v(t)/(\bar\alpha\,\mathbb{E}[\text{res}^2](t)),\,10^{-6}\big)$, with $\Delta\nu_v$ the Onsager part of ν_v in (96), rescales the force. Both are part of the closure. The noise kernel of a pair of distinct experts is (58) with $\mathcal{R}_A$ replaced by $\mathcal{C}_A$.

Two further kernels couple the expert overlaps to the router through the expert–router covariance Q^v. We evaluate them by Price's theorem with the gates replaced by their mean, which gives

$$\begin{aligned} M^{Q^v,\text{self}}(t,s) &= c_Q\,\bar a(s)\int_s^t \mathcal{R}_A(t,\tau)\,\bar a(\tau)\big[\partial_{m_1} H_{ss}(C_d) + \partial_{m_1} H_{cc}(C_d)\big]_{(\tau,s)}\,d\tau, \\ M^{Q^v,\text{cross}}(t,s) &= c_Q\,\bar a(s)\int_s^t \mathcal{R}_A(t,\tau)\,\bar a(\tau)\big[\phi_c(s)\,\phi'_s(\tau) + \phi_s(s)\,\phi'_c(\tau)\big]d\tau, \end{aligned} \tag{60}$$

with $c_Q = \bar\alpha\,\frac{a_1}{E}\sqrt{(E-1)/E}$, the bracket of the first line evaluated at the time pair (τ, s), and $\phi_u = \text{erf}(\kappa P_u/2)$ and $\phi'_u = \pi^{-1/2} e^{-\kappa^2 P_u^2/4}$ for $u \in \{s, c\}$ (with $P_s = P_{\text{self}}$, $P_c = P_{\text{cross}}$), the factorized form of the cross-expert kernel at $C_x = 0$. These kernels are closures and are not local in

time. The $1/m$ of the expert sector enters where they are used, in (93)–(94), and their noise partner $M_C^{Q^v}$ replaces the τ-integral by $\mathcal{C}_A(t,s)$ times the bracket at $\tau = t$. The residual response $\mathcal{R}_A(t,s)$ and correlator $\mathcal{C}_A(t,s)$ are determined by the saddle-point Dyson equation of Appendix C and, under the diagonal-S_E reduction, by scalar Volterra equations driven by the same Hessian slices.

D.3 SIGNAL-SOURCE KERNELS

The expert drive of (87)–(86) is the population expectation

$$\tilde{\Phi} - \Psi \;=\; E\,\mathbb{E}_\lambda\big[g_0\,\mathrm{res}\,\sigma'(z_0)\,\lambda\big], \tag{61}$$

with $\lambda = \lambda_0$ for the self kernel and λ_1 for the cross kernel. We evaluate it directly by quadrature on the grid of Appendix D.1, summing over the E cluster conditions with multiplicities 1, 1 and $E-2$ for the matched cluster, the observed mismatched cluster and the remaining ones. The expectation runs over the teacher coordinates and the out-of-plane router noise, whose covariance with the gates contributes a term proportional to $V_b = 1 - C_o^v - E\,r_D^2/(E-1)$. Evaluated this way the drive reproduces the gradient-flow signal source measured on *held-out* samples, which carries no training-set self-reaction and so isolates the population kernel. The two agree to 8×10^{-4} at every measured time, over 8 seeds at $E = 2$, $m = 300$, $d = 3000$. Appendix D.6 gives the kernel in closed form.

The gate means entering the drive follow from the router discriminant. Writing $\mathrm{disc}_e = s_E(v_e - \bar{v})\cdot x$ with $s_E = \sqrt{E/(E-1)}$, its conditional mean is $\rho_D\,(\lambda_e - \lambda_{\mathrm{tot}}/E)$ with $\lambda_{\mathrm{tot}} = \sum_j \lambda_j$ and $\rho_D = s_E\,\Delta_v$, so

$$\mathrm{Cov}(\mathrm{disc}_e, \lambda_j) \;=\; \rho_D\Big(\delta_{ej} - \tfrac{1}{E}\Big) \;=\; \begin{cases} +r_D, & j = e, \\ -\dfrac{r_D}{E-1}, & j \neq e, \end{cases} \tag{62}$$

using $r_D = \Delta_v\sqrt{(E-1)/E} = \rho_D(E-1)/E$. The matched and mismatched coefficients are therefore *not* equal and opposite except at $E = 2$, and the gate means are $\bar{g}_s = \frac{1}{E} + a_1\kappa r_D$ and $\bar{g}_c = \frac{1}{E} - \frac{a_1\kappa r_D}{E-1}$.

To this population drive the router self-reaction adds a two-time contribution. The gate is linear in the discriminant, $g_e = \frac{1}{E} + a_1\,\mathrm{disc}_e$, so the per-example shift $\Delta\mathrm{disc}_0$ of (105) propagates directly into the gate-weighted drive (61), giving

$$\Delta\tilde{\Phi}_{\mathrm{self/cross}}(t) \;=\; a_1^2\,s_E^2\,E\int_0^t d\tau\; R_-^v(t,\tau)\;\mathbb{E}_\lambda\Big[\mathrm{res}(t)\,\sigma'\big(z_0(t)\big)\,\lambda_{0/1}\;\cdot\;\delta F_0(\tau)\,\mathrm{res}(\tau)\Big], \tag{63}$$

with λ_0 for the self kernel and λ_1 for the cross kernel. The expectation is a *joint* average over the quenched teacher fields at both times and does not factorize into $\mathbb{E}[\cdot](t)\,\mathbb{E}[\cdot](\tau)$. The factor E converts the cluster average into the cluster sum, so that all E cluster conditions contribute with multiplicities $1, 1, E-2$ for the matched, mismatched-observed and remaining experts respectively.

A second self-reaction acts on the *residual* rather than the gate. On the training set, sample μ's residual is reduced by its own contribution to the fit: propagating the router update through the contrast response and using $\|x^\mu\|^2/d \to 1$,

$$\delta f^\mu(t) \;=\; a_1^2 s_E^2\int_0^t R_-^v(t,s)\,Q^\mu(t,s)\,\mathrm{res}^\mu(s)\,ds, \qquad Q^\mu(t,s) = \sum_e \delta F_e^\mu(t)\,\delta F_e^\mu(s), \tag{64}$$

so that $\mathrm{res}_{\mathrm{train}} = \mathrm{res}_{\mathrm{pop}} - \delta f$. The same contrast propagator appears because $\sum_{e'} R^v_{e,e'}\delta F_{e'} = R_-^v\,\delta F_e$. The resulting drive correction is

$$\Delta\tilde{\Phi}^{\mathrm{res}}_{\mathrm{self/cross}}(t) \;=\; -\,a_1^2 s_E^2\,E\sum_k \mathbb{E}_\lambda\Big[A^k(t)\int_0^t R_-^v(t,s)\,B^k(s)\,ds\Big], \tag{65}$$

with $A^k(t) = g_0\,\sigma'(z_0)\,\lambda_{0/1}\,\delta F_k$ evaluated at t, $B^k(s) = \delta F_k\,\mathrm{res}$ evaluated at s, and k running over the three expert classes with multiplicities $1, 1, E-2$. The expert channel of the same self-reaction is $O(1/m)$ and the amplitude channel $O(1/d)$. Both are subleading and dropped.

These two contributions are not independent corrections to be added to the population drive: they are the first-order terms of a single product. On the training set *both* factors of $g_0\,\mathrm{res}$ carry a self-reaction, $g_0^{\mathrm{tr}} = g_0 + a_1\Delta\mathrm{disc}_0$ and $\mathrm{res}^{\mathrm{tr}} = \mathrm{res} - \delta f$, and the drive is the expectation of their product. Expanding it exactly,

$$\begin{aligned}\tilde{\Phi}^{\mathrm{tr}}(t) &= E\,\mathbb{E}_\lambda\big[(g_0 + a_1\Delta\mathrm{disc}_0)\,(\mathrm{res} - \delta f)\,\sigma'(z_0)\,\lambda\big] \\ &= \underbrace{E\,\mathbb{E}_\lambda[g_0\,\mathrm{res}\,\sigma'\lambda]}_{\text{population}} + \underbrace{a_1 E\,\mathbb{E}_\lambda[\Delta\mathrm{disc}_0\,\mathrm{res}\,\sigma'\lambda]}_{(63)} \\ &\quad - \underbrace{E\,\mathbb{E}_\lambda[g_0\,\delta f\,\sigma'\lambda]}_{(65)} - \underbrace{a_1 E\,\mathbb{E}_\lambda[\Delta\mathrm{disc}_0\,\delta f\,\sigma'\lambda]}_{\text{cross}}. \end{aligned} \tag{66}$$

The expansion terminates: a product of two perturbed factors has exactly four terms. The cross term is second order in the self-reaction. It is not negligible in the regime of interest, where the router is strongly aligned (Δ_v reaches 0.58 by $t = 25$ at $\bar{\alpha} = 1$). The reported results include it, with $\Delta\mathrm{disc}_0$ evaluated with the training-set residual $\mathrm{res}^{\mathrm{tr}}$. We evaluate these terms on the 10-node grid of Appendix D.1 with the in-plane residual res_λ of (59), accumulating the contrast response R^v_- at every grid node.

The router drive is the population expectation of the router force against the teacher coordinates, over the full mixture,

$$D_{\mathrm{self}}(t) \;=\; a_1 s_E\,\mathbb{E}_x\big[\delta F_e\,\mathrm{res}_\lambda\,\lambda_e\big], \qquad D_{\mathrm{cross}}(t) \;=\; a_1 s_E\,\mathbb{E}_x\Big[\delta F_e\,\mathrm{res}_\lambda\,\tfrac{\lambda_{\mathrm{tot}}-\lambda_e}{E-1}\Big], \tag{67}$$

evaluated by Monte Carlo at the current order parameters. Shift invariance gives $D_{\mathrm{self}} + (E-1)D_{\mathrm{cross}} = 0$, which the implementation satisfies to machine precision. The training-set residual enters through the factor $1 + \rho(t)$ of (88), with $\rho = -\mathbb{E}_\lambda[\lambda_0\,\delta F_0\,\delta f]/\mathbb{E}_\lambda[\lambda_0\,\delta F_0\,\mathrm{res}]$ the relative size of the residual self-reaction (64) in the router force, which is the two-term expansion (111) of Appendix E.5.

The amplitude-source kernels driving the $\bar{a}$ ODE are

$$\begin{aligned}\hat{\Phi}_{\mathrm{self}}(t) &= \big(\tfrac{1}{E} + a_1\,\kappa\,r_D\big)\,H_\mu(P_{\mathrm{self}}, m_s, m_\lambda) \\ &\quad + a_1\,r_D\,\partial_{m_2}H_\mu(P_{\mathrm{self}}, m_s, m_\lambda) + \delta_{P^v}(t)\,\partial_{m_1}H_\mu(P_{\mathrm{self}}, m_s, m_\lambda), \end{aligned} \tag{68}$$

$$\begin{aligned}\hat{\Phi}_{\mathrm{cross}}(t) &= \big(\tfrac{1}{E} - \tfrac{a_1\,\kappa\,r_D}{E-1}\big)\,H_\mu(P_{\mathrm{cross}}, m_c, m_\lambda) \\ &\quad - \tfrac{a_1\,r_D}{E-1}\,\partial_{m_2}H_\mu(P_{\mathrm{cross}}, m_c, m_\lambda) + \delta_{P^v}(t)\,\partial_{m_1}H_\mu(P_{\mathrm{cross}}, m_c, m_\lambda), \end{aligned} \tag{69}$$

where $\delta_{P^v}(t) := a_1\sqrt{(E-1)/E}\,\big(Q^v_{\mathrm{self}}(t,t) - Q^v_{\mathrm{cross}}(t,t)\big) = a_1\,r_D(t)\,\Delta(t)$, because the solver pins $Q^v(t,t)$ to its teacher-plane value (Appendix D.7). These amplitude kernels use the uncorrected r_D: applying the mean shift δ_{disc} to them inflates the $\bar{a}$ source by 22.8% and makes $\bar{a}$ grow faster than in gradient flow (Appendix E.5). For the softmax router, $\bar{g}_s$ and $\bar{g}_c$ in these kernels are the exact softmax gate means, while the terms proportional to a_1 keep their linear-gate form.

D.4 Two-time Volterra equations

The expert-weight intra-unit correlators and responses obey

$$\begin{aligned}\partial_t C_d(t,s) &= +\nu_w(t)\,C_d(t,s) - \frac{1}{m}\int_0^t \big[M^d_R(t,\tau)\,C_d(\tau,s) + (m-1)\,M^o_R(t,\tau)\,C_o(\tau,s)\big]\,d\tau \\ &\quad - \frac{1}{m}\int_0^s \big[M^d_C(t,\tau)\,R_d(t,\tau) + (m-1)\,M^o_C(t,\tau)\,R_o(t,\tau)\big]\,d\tau, \end{aligned} \tag{70}$$

$$\begin{aligned}\partial_t R_d(t,s) &= \delta(t-s) + \nu_w(t)\,R_d(t,s) \\ &\quad - \frac{1}{m}\int_s^t \big[M^d_R(t,\tau)\,R_d(\tau,s) + (m-1)\,M^o_R(t,\tau)\,R_o(\tau,s)\big]\,d\tau. \end{aligned} \tag{71}$$

The noise term enters (70) with the sign and the response pairing $R(t,\tau)$ of our solver, which differ from the MSRDJ form (Appendix D.7). The off-diagonal hidden-unit pair (C_o, R_o) does *not* follow from a bare $M^d \leftrightarrow M^o$ exchange, and C_o carries no memory term, because its two-time signal drive

already contains the out-of-plane Onsager self-reaction, which a memory integral would count a second time. The equations are

$$\partial_t C_o(t,s) = +\nu_w(t)\, C_o(t,s) + \Xi_o(t,s), \tag{72}$$

$$\partial_t R_o(t,s) = +\nu_w(t)\, R_o(t,s) - \frac{1}{m}\int_s^t \big[M_R^d\, R_o(\tau,s) + M_R^o\, R_d(\tau,s) + (m-2)\, M_R^o\, R_o(\tau,s)\big]\, d\tau, \tag{73}$$

where

$$\Xi_o(t,s) = \frac{\bar\alpha\, \bar a(t)}{mE}\Big[\big(P_{\rm self} - P_{\rm cross}\big)(s)\, \tilde\Phi_{\rm self}(t) + P_{\rm cross}(s)\, \big(\tilde\Phi_{\rm self} + (E-1)\tilde\Phi_{\rm cross}\big)(t)\Big]. \tag{74}$$

Here $\tilde\Phi$ denotes the full drive $\tilde\Phi - \Psi$ of (6), including the self-reaction of Appendix D.3, and the equal-time value $C_o(t,t)$ is pinned by (101). The $(m-2)$ multiplicity in (73) counts the remaining neurons of the same expert and is the term a bare kernel exchange omits. R_o carries no $\delta(t-s)$ source, since distinct neurons have no instantaneous self-response. Equation (72) reproduces the simulated $C_o(t,s)$ to 0.12% at every measured t. The cross-expert pair has no drive of the form Ξ_o and couples to the router through the Q^v kernels of (60):

$$\begin{aligned}\partial_t C_x(t,s) = \nu_w(t)\, C_x(t,s) - \frac{1}{m}\int_0^t \big[&M_R^d\, C_x + M^{Q^v,\rm cross}\, C_{dv} \\ &+ \big(M^{Q^v,\rm self} + M^{Q^v,\rm cross}\big) C_{vo}\big](\tau,s)\, d\tau \\ - \int_0^s \big[&M_C^{vd}(t,\tau)\, Q^v_{\rm cross}(\tau,t) + M_C^{Q^v}(t,\tau)\, R_d(t,\tau)\big]\, d\tau,\end{aligned} \tag{75}$$

with R_x carrying the memory $\frac{1}{m}\big[M_R^d R_x + M^{Q^v,\rm cross} R_{dv}\big]$ and no $\delta(t-s)$ source. The expert–router covariance $Q^v_{\rm self}(t,s)$ obeys the same form with memory $\frac{1}{m}[M_R^d Q^v + M^{Q^v,\rm self} C_{dv} + M^{Q^v,\rm cross} C_{vo}]$ and noise $-\int_0^s M_C^{Q^v} R_d(t,\tau)\, d\tau$, and $Q^v_{\rm cross}$ with C_{dv} and C_{vo} exchanged. The equal-time values of C_o, C_x and Q^v are pinned to their teacher-plane values (Appendix D.7).

The router correlators are *not* propagated as a single Volterra equation. Because $\|v_e\| = 1$, the two-time router correlator splits exactly into a deterministic teacher-plane part, which is algebraic in the single-time overlaps already solved for, and an out-of-plane part carrying all of the MSRDJ memory:

$$C_{dv}(t,s) = \underbrace{P^v_{\rm self}(t) P^v_{\rm self}(s) + (E-1)\, P^v_{\rm cross}(t) P^v_{\rm cross}(s)}_{\text{teacher plane, algebraic}} + C_d^\perp(t,s), \tag{76}$$

$$C_{vo}(t,s) = P^v_{\rm self}(t) P^v_{\rm cross}(s) + P^v_{\rm cross}(t) P^v_{\rm self}(s) + (E-2)\, P^v_{\rm cross}(t) P^v_{\rm cross}(s) \;+\; C_o^\perp(t,s), \tag{77}$$

with the out-of-plane pair obeying

$$\begin{aligned}\partial_t C_d^\perp(t,s) = -\mu(t)\, C_d^\perp(t,s) &- \int_0^t \big[M_R^{vd}\, C_d^\perp(\tau,s) + (E-1)\, M_R^{vo}\, C_o^\perp(\tau,s)\big]\, d\tau \\ &+ \int_0^s \big[M_C^{vd}\, R_{dv}(t,\tau) + (E-1)\, M_C^{vo}\, R_{vo}(t,\tau)\big]\, d\tau,\end{aligned} \tag{78}$$

$$\begin{aligned}\partial_t C_o^\perp(t,s) = -\mu(t)\, C_o^\perp(t,s) &- \int_0^t \big[M_R^{vd}\, C_o^\perp(\tau,s) + M_R^{vo}\, C_d^\perp(\tau,s) + (E-2)\, M_R^{vo}\, C_o^\perp(\tau,s)\big]\, d\tau \\ &+ \int_0^s \big[M_C^{vd}\, R_{vo}(t,\tau) + M_C^{vo}\, R_{dv}(t,\tau) + (E-2)\, M_C^{vo}\, R_{vo}(t,\tau)\big]\, d\tau,\end{aligned} \tag{79}$$

all kernels evaluated at (t,τ). The equal-time values are fixed by the constraint and by the equal-time C_{vo} equation,

$$\begin{aligned} C_d^\perp(t,t) &= 1 - \big[P^v_{\rm self}(t)^2 + (E-1) P^v_{\rm cross}(t)^2\big], \\ C_o^\perp(t,t) &= C_{vo}(t,t) - \big[2 P^v_{\rm self} P^v_{\rm cross} + (E-2)(P^v_{\rm cross})^2\big](t),\end{aligned} \tag{80}$$

and $\mu(t)$ is a Lagrange multiplier, not a free parameter: as in spherical models it is *defined* by requiring (78) to reproduce the exact diagonal derivative $\partial_t C_d^\perp(t,t)$ implied by (80). No parameter

is fitted, so $C^{\perp}(t,s)$ at $s<t$ follows from the equations, given the pairing choice of Appendix D.7. In the exact dynamics $\mu(t)$ equals $\nu_v(t)$. Fixing it by the diagonal instead keeps the equal-time pin exact under the closures and absorbs their error into μ. The out-of-plane source enters (78)–(79) with a *positive* sign, and it is the only term not proportional to $C^{\perp}$, so it alone produces the decorrelation.

The router response obeys an ordinary Volterra equation,

$$\partial_t R_{dv}(t,s) = \delta(t-s) - \nu_v(t)\, R_{dv}(t,s) - \int_s^t \big[M_R^{vd}(t,\tau)\, R_{dv}(\tau,s) + (E-1)\, M_R^{vo}(t,\tau)\, R_{vo}(\tau,s)\big]\, d\tau, \tag{81}$$

while the off-diagonal response carries no $\delta(t-s)$ source and has memory $M_R^{vd}\, R_{vo} + M_R^{vo}\big[R_{dv} + (E-2)\, R_{vo}\big]$, the $(E-2)$ counting the remaining experts. Subtracting the two leaves the contrast equation (107), in which the multiplicities cancel identically.

Residual response and correlator. The residual response $\mathcal{R}_A(t,s)$ and correlator $\mathcal{C}_A(t,s)$ that drive every memory and signal kernel above are determined by their own scalar Volterra equations. Reusing the residual self-energies Σ_R, Σ_C derived in Appendix C (eqs. (23)–(24)) and specialized to the diagonal-S_E saddle, these read

$$\mathcal{R}_A(t,s) = \delta(t-s) + \int_s^t \Sigma_R(t,\tau)\, \mathcal{R}_A(\tau,s)\, d\tau, \tag{82}$$

$$\mathcal{C}_A(t,s) = \int_0^t \int_0^s \mathcal{R}_A(t,\tau)\, \Sigma_C(\tau,\tau')\, \mathcal{R}_A(s,\tau')\, d\tau\, d\tau'. \tag{83}$$

Writing $\mathcal{R}_A = \delta + r$, the leading term of this contraction is $\Sigma_C(t,s)$ on its own, which the solver retains:

$$\mathcal{C}_A(t,s) \;=\; \Sigma_C(t,s) \;+\; \mathcal{O}(r). \tag{84}$$

At $E = 2$ it matches the measured $\mathcal{C}_A(t,t) = \mathbb{E}[\mathrm{res}^2]$ on the diagonal, and off the diagonal it preserves the persistence of Σ_C, as the simulated lag^2 decay requires. The solver sets $\mathcal{C}_A(t_{n+1},\cdot) = \Sigma_C(t_n,\cdot)$, with a lag of one step. The subleading terms are $\mathcal{O}(r)$ and are not retained: evaluated with the present $\mathcal{R}_A$ they change the measured diagonal by only 0.4 to 0.5%, so the leading term suffices. Here Σ_R, Σ_C are functions of the same order parameters $(P_{\mathrm{self/cross}}, P^{v}_{\mathrm{self/cross}}, \bar{a}, C_d, R_d, C_o, R_o, C_x, R_x, C_{dv}, R_{dv}, C_{vo}, R_{vo})$ tracked by (70)–(81) and the cross-expert and expert–router companions, evaluated through the kernels $H, \hat{\Phi}$ of Appendix C.3. The solver stores $\mathcal{R}_A$ in cell-integral form, $\hat{\mathcal{R}}_A[n,j] \approx h\, \mathcal{R}_A(t_n,t_j)$ for the forward-Euler step h, with the δ placed on the subdiagonal:

$$\hat{\mathcal{R}}_A[n+1,j] \;=\; \mathbf{1}[j=n] \;+\; h \sum_{k<n} \Sigma_R[n,k]\, \hat{\mathcal{R}}_A[k,j]. \tag{85}$$

Memory integrals are formed as $h\sum_j M[n,j]\, X[j]$. The kernels M_R^d, M_R^o, M_R^{vd} and M_R^{vo} are built directly from $\hat{\mathcal{R}}_A$, so the memory terms they carry in the one-time, response and δ_{disc} equations, the δ part of $\mathcal{R}_A$ included, enter at order h and vanish as $h \to 0$. The Q^v kernels (60) are the exception: their τ-integral uses the step carried by $\hat{\mathcal{R}}_A$, so they enter at order one. The system we integrate is therefore, up to order h, the one displayed with the memory terms of M_R^d, M_R^o, M_R^{vd} and M_R^{vo} removed. In the MSRDJ equations the δ part of $\mathcal{R}_A$ enters these terms at full weight, and in the equation for P_{self} it would be about a third of dP_{self}/dt as integrated. We have not derived its removal. It is a closure choice, and the agreement with gradient flow in Section 4.4 and Appendix E is obtained with it. The memory that remains as $h \to 0$ is the router self-reaction (Appendix E), the Onsager part of ν_v, the noise terms built from $\mathcal{C}_A$, and the Q^v kernels. At $m = 300$, halving h from 0.1 moves the P_{cross} peak by 0.2%, and Δ and P_{self} at $t = 50$ by 1.7% and 1.5% of their change during training. The quantities most sensitive to h are the router's contrast response (Appendix D.2) and δ_{disc}, which at $t = 50$ is 0.535 at $h = 0.1$ and about 0.55 as $h \to 0$ (Appendix E.3). In our runs the entries of $\hat{\mathcal{R}}_A$ off the subdiagonal stay below 3×10^{-4}. The residual sector adds two scalar two-time arrays of the same $\mathcal{O}(N_t^2)$ memory cost as the expert and router two-time correlators above. Under the diagonal-S_E reduction, the cluster and expert sums in Σ_R, Σ_C collapse to weighted scalar combinations of the symmetric-mode kernels of Appendix D.2 via the counting lemma (47).

D.5 Single-time ODEs and Lagrange multipliers

The single-time scalar order parameters obey ODEs of the form

$$\frac{dP_{\text{cross}}}{dt} = \nu_w(t)\, P_{\text{cross}}(t) + \frac{\bar{\alpha}\,\bar{a}(t)}{mE}\big[\tilde{\Phi}_{\text{cross}}(t) - \Psi_{\text{cross}}(t)\big] - \text{mem}_{P_c}(t), \tag{86}$$

$$\frac{dP_{\text{self}}}{dt} = \nu_w(t)\, P_{\text{self}}(t) + \frac{\bar{\alpha}\,\bar{a}(t)}{mE}\big[\tilde{\Phi}_{\text{self}}(t) - \Psi_{\text{self}}(t)\big] - \text{mem}_{P_s}(t), \tag{87}$$

$$\frac{dP^v_{\text{self}}}{dt} = -\nu_v(t)\, P^v_{\text{self}}(t) + \bar{\alpha}\,\big(1+\rho(t)\big)\, D_{\text{self}}(t), \tag{88}$$

$$\frac{dP^v_{\text{cross}}}{dt} = -\nu_v(t)\, P^v_{\text{cross}}(t) + \bar{\alpha}\,\big(1+\rho(t)\big)\, D_{\text{cross}}(t), \tag{89}$$

$$\frac{d\bar{a}}{dt} = \frac{\bar{\alpha}}{m}\,\big[\psi_{\bar{a}}(t) - \bar{a}(t)\, H_{\bar{a}}(t)\big], \tag{90}$$

where the amplitude drive and its equal-time self-energy are *cluster averages* over the E equiprobable clusters,

$$\psi_{\bar{a}}(t) \;=\; \frac{1}{E}\big[\hat{\Phi}_{\text{self}}(t) + (E-1)\,\hat{\Phi}_{\text{cross}}(t)\big], \tag{91}$$

$$\begin{aligned} H_{\bar{a}}(t) \;=\; \frac{1}{E}\Big[&\sum_{u\in\{s,c\}} n_u\,\langle g^2\rangle_u \Big(\tfrac{1}{m} H_\mu\big(1, m_u, m_u\big) + \big(1-\tfrac{1}{m}\big) H_\mu\big(C_o(t,t), m_u, m_u\big)\Big) \\ &+\; (E-1)\,\langle g_0 g_e\rangle \big[H_\mu\big(C_x(t,t), m_s, m_c\big) + H_\mu\big(C_x(t,t), m_c, m_s\big)\big] \\ &+\; (E-1)(E-2)\,\langle g_e g_{e'}\rangle_{\text{mm}}\, H_\mu\big(C_x(t,t), m_c, m_c\big)\Big], \end{aligned} \tag{92}$$

with $n_s = 1$, $n_c = E-1$, $\langle g^2\rangle_{s/c}$ and $\langle g_0 g_e\rangle$ the equal-time gate second moments of (98), and $\langle g_e g_{e'}\rangle_{\text{mm}} = (\bar{g}_c - \langle g^2\rangle_c - \langle g_0 g_e\rangle)/(E-2)$ the moment of two mismatched gates, which the sum rule $\sum_e g_e = 1$ fixes. The multiplicities follow from averaging over the E equiprobable clusters: expert e sees one matched cluster and $E-1$ mismatched ones, and its pairings with the other experts split into $E-1$ matched–mismatched, $E-1$ mismatched–matched, and $(E-1)(E-2)$ mismatched–mismatched orbits. At $E=2$ every multiplicity is one and the last orbit is empty. The $1/E$ is an average over clusters rather than a per-expert convention and does not cancel against the expert sum, and within an expert the same-neuron and off-neuron kernels carry weights $1/m$ and $1-1/m$. The memory corrections couple back through the response functions:

$$\begin{aligned} \text{mem}_{P_s}(t) = \frac{1}{m}\int_0^t \big[&M^d_R(t,\tau)\, P_{\text{self}}(\tau) + M^{Q^v,\text{self}}_R(t,\tau)\, P^v_{\text{self}}(\tau) \\ &+ (E-1)\, M^{Q^v,\text{cross}}_R(t,\tau)\, P^v_{\text{cross}}(\tau)\big]\, d\tau, \end{aligned} \tag{93}$$

$$\begin{aligned} \text{mem}_{P_c}(t) = \frac{1}{m}\int_0^t \big[&M^d_R(t,\tau)\, P_{\text{cross}}(\tau) + M^{Q^v,\text{self}}_R(t,\tau)\, P^v_{\text{cross}}(\tau) \\ &+ M^{Q^v,\text{cross}}_R(t,\tau)\, P^v_{\text{self}}(\tau) + (E-2)\, M^{Q^v,\text{cross}}_R(t,\tau)\, P^v_{\text{cross}}(\tau)\big]\, d\tau, \end{aligned} \tag{94}$$

The router equations use the drives $D_{\text{self/cross}}$ and the factor $1+\rho$ of (67), both instantaneous, and carry no $\mathcal{R}_A$-based memory. The amplitude equation is instantaneous in the continuum. The solver evaluates it as a retarded convolution with the discretized $\hat{\mathcal{R}}_A$ of (85), whose δ sits on the subdiagonal, so the step adds h times the previous integrand, and at $h = 0.1$ the $\bar{a}$ drive is 1.1 times the instantaneous form. In (93)–(94) the M^d_R terms enter at order h and the Q^v terms at order $1/m$. In each memory the sum runs over the mediating expert e'', so the multiplicities total E. The kernel is $M^{Q^v,\text{self}}$ when $e'' = e$ and $M^{Q^v,\text{cross}}$ otherwise, and the router overlap is P^v_{self} when e'' is the expert of the tracked cluster and P^v_{cross} otherwise. For $P_{\text{self}} = w_e \cdot U_e$ this gives the self kernel with P^v_{self} once and the cross kernel with P^v_{cross} for the $E-1$ other experts. For $P_{\text{cross}} = w_e \cdot U_{e'}$ it gives the self kernel with P^v_{cross} ($e'' = e$), the cross kernel with P^v_{self} ($e'' = e'$), and the cross kernel with P^v_{cross} for the $E-2$ others.

The spherical-constraint Lagrange multipliers are

$$\nu_w(t) = -\frac{\bar{\alpha}\,\bar{a}(t)}{mE}\big[\tilde{\Phi}^z(t) - \Psi^z(t)\big], \tag{95}$$

$$\nu_v(t) = \bar{\alpha}\,\mathbb{E}_\lambda\big[c_e^v(t)\,z_e^v(t)\big] \;+\; \bar{\alpha}\int_0^t R_{dv}(t,\tau)\,\mathbb{E}_\lambda\big[c_e^v(t)\,c_e^v(\tau)\big]\,d\tau. \tag{96}$$

The first term of ν_v is the population average of the router force c_e^v against the router preactivation z_e^v, with $c_e^v = a_1 s_E\,\delta F_e\,\mathrm{res}$ for the soft-discriminant gate and $c_e^v = T^{-1} g_e(F_e - f)\,\mathrm{res}$ for softmax (40). For the softmax router our solver keeps its linear-gate form. The second term is the Onsager self-reaction, the R_{dv}-propagated force correlator of (59), with the same rescaling $\lambda(t)\lambda(\tau)$. The exact propagator of an example's own kick is the contrast response $R_-^v = R_{dv} - R_{vo}$ (Appendix E.3), which differs from R_{dv} by at most 3×10^{-4} at the settings of Figure 5, and (102) makes the same substitution. Both terms are evaluated *instantaneously*. The expert multiplier ν_w is likewise instantaneous: it is fixed algebraically by the equal-time condition $\frac{d}{dt}\|w_{l,e}\|^2 = 0$, which on (14) gives $\nu_w = -\frac{\bar{a}}{md}\sum_\mu g_e\,\mathrm{res}^\mu\,\sigma'(z^w_{l,e,\mu})\,(x^\mu\cdot w_{l,e})$. Since $x\cdot w = z^w$, this is the same expectation as the P drive with the teacher direction λ replaced by the neuron's own pre-activation, which is the content of (95). In the reported results the kernels of (95) are evaluated at the shifted discrimination $r_D + \delta_{\mathrm{disc}}/\kappa$ (Appendix E.5). The kernels are given in closed form in Appendix D.6.

D.6 Explicit form of the drive and constraint kernels

The four kernels appearing in (87) and (95) are the same Gaussian expectation evaluated in four combinations: teacher versus student residual, and teacher-direction versus own-pre-activation slot. Write, for data from cluster c, $\lambda_j = U_j\cdot x \sim \mathcal{N}(\kappa\delta_{jc}, 1)$ independent across j, and

$$u_e \;=\; \sum_j P_{e,j}\,\lambda_j, \qquad \mathcal{P}_e \;=\; \textstyle\sum_j P_{e,j}^2, \qquad q \;=\; \sqrt{2-\mathcal{P}_e}, \qquad D_e \;=\; \frac{\sigma'(u_e/q)}{q}, \tag{97}$$

with $P_{e,j} = P_{\mathrm{self}}$ for $j = e$ and P_{cross} otherwise. Three integrals close exactly, reducing these objects to low-dimensional quadratures over the teacher coordinates:

1. *Neuron average.* At the closure used here, $C_o(t,t) = \mathcal{P}_e$ and $C_x(t,t) = 2P_{\mathrm{self}}P_{\mathrm{cross}} + (E-2)P_{\mathrm{cross}}^2$, so the neuron mean direction has no component outside the teacher plane and $z_{l,e} = u_e + \sqrt{1-\mathcal{P}_e}\,\xi_l$ with ξ_l independent. For $\sigma(z) = \mathrm{erf}(z/\sqrt{2})$ the average over neurons is then exact, $\langle\sigma(z_{l,e})\rangle_l = \sigma(u_e/q)$, giving $F_e = \bar{a}\,\sigma(u_e/q)$.
2. *The ξ integral.* $\mathbb{E}_\xi[\sigma'(u + \sqrt{1-\mathcal{P}_e}\,\xi)] = D_e$ and, in the own-pre-activation slot, $\mathbb{E}_\xi\big[\sigma'(z_{l,e})\,z_{l,e}\big] = (u_e/q^2)\,D_e$. The second follows from $\mathbb{E}[e^{-w^2/2}w] = -\partial_u\mathbb{E}[e^{-w^2/2}]$.
3. *The router integral.* The gate is affine in the discriminant, so with $\mu_{d,e} = \frac{E}{E-1}\,r_D\big(\lambda_e - \frac{1}{E}\sum_j\lambda_j\big)$ and $V_b = (1-C_o^v) - \frac{E}{E-1}r_D^2$ (at $E=2$: $\mu_d = r_D(\lambda_0-\lambda_1)$, $V_b = (1-C_o^v) - 2r_D^2$),

$$\begin{aligned}\langle g_0\rangle &= \tfrac{1}{E} + a_1\mu_d, \qquad \langle g_0^2\rangle = \tfrac{1}{E^2} + \tfrac{2a_1\mu_d}{E} + a_1^2(\mu_d^2 + V_b),\\ \langle g_0 g_e\rangle &= \tfrac{1}{E^2} + \tfrac{a_1(\mu_{d,0}+\mu_{d,e})}{E} + a_1^2\Big(\mu_{d,0}\mu_{d,e} - \tfrac{V_b}{E-1}\Big), \quad e\neq 0,\end{aligned} \tag{98}$$

Writing $\mathbb{E}_\lambda$ for the remaining Gaussian average over the teacher coordinates and $S \equiv \sum_e\langle g_0 g_e\rangle\,\sigma(u_e/q)$ (which at $E=2$ is $\langle g_0^2\rangle\sigma(u_0/q) + \langle g_0 g_1\rangle\sigma(u_1/q)$), the four kernels are

$$\tilde{\Phi}_{\mathrm{self}}(t) = \sum_c \mathbb{E}_\lambda\big[\langle g_0\rangle\,\sigma(\lambda_c)\,D_0\,\lambda_0\big], \qquad \Psi_{\mathrm{self}}(t) = \bar{a}(t)\sum_c \mathbb{E}_\lambda\big[S\,D_0\,\lambda_0\big], \tag{99}$$

$$\tilde{\Phi}^z(t) = \sum_c \mathbb{E}_\lambda\Big[\langle g_0\rangle\,\sigma(\lambda_c)\,D_0\,\tfrac{u_0}{q^2}\Big], \qquad \Psi^z(t) = \bar{a}(t)\sum_c \mathbb{E}_\lambda\Big[S\,D_0\,\tfrac{u_0}{q^2}\Big], \tag{100}$$

and $\tilde{\Phi}_{\mathrm{cross}}, \Psi_{\mathrm{cross}}$ are (99) with λ_1 in the final slot. The $\tilde{\Phi}$ are the teacher contributions. The Ψ are the student self-interaction, i.e. the part of the residual $\mathrm{res} = y - f$ coming from the network's own output.

Why the drive carries Ψ and an instantaneous kernel. Two features of (87) deserve comment. First, the drive is teacher *minus* student: the mean-field student term Ψ is not a small correction, and it grows over training: in the validation regime $\Psi_{\rm self}/\tilde{\Phi}_{\rm self}$ runs from 11% at initialization to 21% by the end of the run, and $\Psi_{\rm cross}/\tilde{\Phi}_{\rm cross}$ from 6% to 15%. A teacher-only drive is inflated by that amount throughout. Second, the kernel is instantaneous. A retarded convolution $\int_0^t R_A(t,s)\tilde{\Phi}(s)\,ds$ of the bare teacher signal would resum the student back-reaction through R_A, and combining it with Ψ would count that back-reaction twice. Since R_A is close to $\delta(t-s)$ in our runs, the retarded teacher-only form differs from (87) mainly by the missing Ψ, and its trajectory RMSE is 19 times larger. Section 4.4 reports the accuracy of (87). The amplitude equation (90) carries its student term in the same way. The term ν_w is small, about 2% of $dP_{\rm self}/dt$, so the $1/m$ scaling of (95) matters more than its precise magnitude.

Relation to the self-energy expansion. We use the closed form (95) for ν_w instead of a memory-kernel expression of the form $-\int[M_R^d C_d + (m-1)M_R^o C_o]$. In that expression the student self-interaction enters through $F_e = (\bar{a}/m)\sum_{l'}\sigma(z_{l',e})$, an *average* over neurons, so the same-neuron and other-neuron parts must carry weights $\frac{1}{m}$ and $1-\frac{1}{m}$, not 1 and $m-1$. With the weights 1 and $m-1$ the expression is too large by a factor of m: the ratio of the expansion to the directly measured student projection, divided by m, is 0.55, 0.53 and 0.53 at $m = 100$, 400 and 1200. With the correct weights it still differs from the measured projection by a factor of about two, which we trace to M_R^o, so we use the closed form (100). The kernel M_R^o enters the two-time equations (70)–(73) only through the order-h memory of (85).

Why ν_v is instantaneous. The router Lagrange multiplier ν_v is determined algebraically by the spherical constraint $\frac{d}{dt}(v_e\cdot v_e) = 0$, which is an *equal-time* condition: differentiating $\|v_e(t)\|^2 = 1$ gives $v_e\cdot\dot{v}_e = 0$ at every t, and substituting (15) produces a single algebraic equation for $\nu_v(t)$ in terms of quantities all evaluated at time t. The first term of (96) is this equal-time projection of the population router force, and the second is the Onsager self-reaction of the same projection. ν_v therefore inherits the equal-time structure of the spherical constraint.

D.7 Equal-time pins, initial conditions, and integration

The spherical constraints $\|w_{l,e}\| = \|v_e\| = 1$ pin $C_d(t,t) = C_{dv}(t,t) = 1$. The solver also pins the equal-time expert correlators and the expert–router covariance to their teacher-plane values,

$$\begin{aligned} C_o(t,t) = P_{\rm self}^2 + (E-1)P_{\rm cross}^2, \qquad C_x(t,t) = 2P_{\rm self}P_{\rm cross} + (E-2)P_{\rm cross}^2, \\ Q^v_{\rm self}(t,t) = P_{\rm self}P^v_{\rm self} + (E-1)P_{\rm cross}P^v_{\rm cross}, \\ Q^v_{\rm cross}(t,t) = P_{\rm self}P^v_{\rm cross} + P_{\rm cross}P^v_{\rm self} + (E-2)P_{\rm cross}P^v_{\rm cross}, \end{aligned} \tag{101}$$

which neglect the out-of-plane parts of these overlaps, while the off-diagonal router correlator at equal times obeys its own equation, for $e \neq e'$,

$$\partial_t C_{vo}(t,t) \;=\; 2\bar{\alpha}\,\mathbb{E}_\lambda\big[c_e^v(t)\,z_{e'}^v(t)\big] \;-\; \frac{2\bar{\alpha}}{E-1}\int_0^t R_{dv}(t,\tau)\,\mathbb{E}_\lambda\big[c_e^v(t)\,c_e^v(\tau)\big]\,d\tau \;-\; 2\nu_v(t)\,C_{vo}(t,t). \tag{102}$$

The middle term is the Onsager self-reaction of ν_v in (96), shared among the $E-1$ other experts because $\sum_e c_e^v = 0$, with the rescaling of (59). If the out-of-plane router components were uncorrelated, $C_{vo}(t,t)$ would equal its teacher-plane value $2P^v_{\rm self}P^v_{\rm cross} + (E-2)(P^v_{\rm cross})^2$. They are anti-correlated, because the router gradients of different experts oppose each other, so the solver integrates (102) instead.

Initial conditions. At $t = 0$ under the initialization (3):

$$P_{\rm self}(0) = \varepsilon, \quad P_{\rm cross}(0) = 0, \quad P^v_{\rm self}(0) = \varepsilon^v, \quad P^v_{\rm cross}(0) = 0, \quad \bar{a}(0) = \bar{a}_0, \quad \delta_{\rm disc}(0) = 0,$$

$$C_d(0,0) = 1, \quad C_{dv}(0,0) = 1, \quad C_o(0,0) = \varepsilon^2, \quad C_x(0,0) = 0,$$
$$C_{vo}(0,0) = 0, \quad Q^v_{\rm self}(0,0) = \varepsilon\,\varepsilon^v, \quad Q^v_{\rm cross}(0,0) = 0,$$

consistent with (101). Here ε and ε^v are the expert- and router-side biases and $\bar{a}_0$ the initial second-layer amplitude, which is distinct from the gate constant $1/E$. The runs reported here use $\varepsilon = 0.2$,

$\varepsilon^v = 0.05$ and $\bar{a}_0 = 0.5$ unless stated otherwise. Throughout the two-time equations the constraint drift enters as $+\nu_w$ in the expert sector and $-\nu_v$ in the router sector. The C-type memory terms are integrated to $\min(t, s)$ with the response evaluated at (t, τ), as in (70). The MSRDJ derivation pairs the noise kernel with $R(s, \tau)$ instead and enters the expert-sector noise term with a positive sign. The pairing $R(t, \tau)$ is a choice made by comparison with gradient flow: on $1 - C_{dv}$ at $E = 2$ it agrees within 1.4σ over 5 seeds, where the MSRDJ pairing is 5% low. The difference from the MSRDJ form, pairing and sign, affects only the two-time correlators at $s < t$. It is small in our runs, because C_d enters only through H near $C_d = 1$ and C_o carries no memory term. The response functions satisfy the equal-time conditions $R_d(t^+, t) = R_{dv}(t^+, t) = 1$ and $R_o(t^+, t) = R_x(t^+, t) = R_{vo}(t^+, t) = 0$ throughout the integration. The initial condition $C_{vo}(0, 0) = 0$ holds because distinct router vectors start with independent out-of-plane components and $P^v_{\text{cross}}(0) = 0$, and (102) then drives $C_{vo}(t, t)$ away from zero as the router trains.

Forward-Euler integration. The Volterra integrals have a strictly causal structure: the integrand at (t, τ) depends only on order parameters at times $\leq t$, so no self-consistency iteration is required at each step. We integrate by forward Euler with step size h and maximum time $T_{\max}$, storing two-time functions on the upper-triangular grid $0 \leq s \leq t \leq T_{\max}$. Memory storage is $\mathcal{O}(N_t^2)$ per two-time function with $N_t = T_{\max}/h$. The cost per step is dominated by the quadratures and Monte Carlo averages of Appendix D.1.

D.8 Implementation at $E > 2$

At $E = 2$ every multiplicity above is one, and the solver integrates the equations above with the treatment of the memory described in Appendix D.4. At $E > 2$ three points need care. First, in a cluster matched to neither of the two resolved experts, one of the remaining $E - 2$ experts is matched to the cluster and $E - 3$ are not. The drive evaluates the matched one at the cluster's coordinate and the others at a resolved unmatched coordinate, with the cluster's own coordinate absorbed into the sum of the rest, which is exact because the drive is linear in the student output. For softmax the unresolved experts enter the normalizer as independent factors with their own means. Their correlation with the sum of the rest, which reaches the expert fields only through P_{cross}, is neglected, a relative effect of order $P_{\text{cross}} \lesssim 10^{-2}$. Second, the self-reaction involves several experts' outputs at one example and is evaluated on the Sobol sample of Appendix D.1. Third, the kernels M^d_R, M^o_R, their noise partners, the Q^v kernels and Σ_C weight the mismatched clusters by $E - 1$, and the cross-expert Q^v kernel includes the $E - 2$ clusters matched to neither expert. The soft-discriminant run at $E = 4$ of Appendix I uses this implementation. The $E = 4$ and 8 curves of Figures 11 and 12 and the $E = 4$ and 8 Jacobians of Appendices H.2 and J were computed before these corrections, with the experts outside the matched and one mismatched cluster at a shared coordinate and the mismatched clusters weighted by one in the kernels above. In the softmax runs reported here, at any E, the population part of ν_v, the router-force Gram, Σ_R, Σ_C, the router memory kernels, the couplings of the Q^v kernels, ν_w, the δ_{disc} source and the self-reaction keep their linear-gate forms at the matched a_1.

E Router Onsager Correction: Source-Term Derivation

This appendix derives the router self-reaction at general E for the soft-discriminant router $r_e(x) = a_0 + a_1 \operatorname{disc}_e(x)$: its per-example form, the closed Volterra equation for its cluster mean $\delta_{\text{disc}}(t)$ with the source term, and how the correction enters the dynamics. The Volterra structure extends to any router whose gradient contains a router–residual self-coupling. The explicit source and self-reaction terms are specific to the soft-discriminant router.

E.1 Physical motivation

In the uncorrected DMFT, the centered logit $\operatorname{disc}^\mu = \operatorname{disc}_{c^\mu}(x^\mu)$ of a training example for its matched expert is assumed Gaussian with the population mean $\kappa\, r_D(t)$, where $r_D = \sqrt{(E-1)/E}\,\Delta_v$ is the normalized router contrast of Appendix B.1.

This population assumption is correct for the experts but *fails* for the router. The router weights V have been trained on the same examples whose logits we evaluate: every training example contributes

a self-reaction $\delta V_e \propto \mathrm{res}^\mu \delta F_e^\mu x^\mu/d$ that pushes its own disc^μ above the population prediction, so training examples are routed to their matched expert more confidently than fresh ones. The standard Onsager term carried by the residual response R_A does not capture this, because R_A corrects the expert-weight self-reaction, not the conditional routing distribution on which the gate-weighted expectations are evaluated.

E.2 The order parameter δ_{disc}

We define the gap between the conditional means of the matched-expert discriminant on the training examples and on fresh examples of the same cluster as

$$\delta_{\mathrm{disc}}(t) \;:=\; \mathbb{E}_{\mathrm{train}}\big[\mathrm{disc}^\mu(t)\,\big|\,c^\mu=1\big] \;-\; \mathbb{E}_{\mathrm{pop}}\big[\mathrm{disc}(t)\,\big|\,c=1\big], \tag{103}$$

where the right-hand expectation is the population statistic, which the uncorrected DMFT computes, and the left-hand expectation averages over the actual training examples. By the S_E symmetry every cluster gives the same gap for its own matched expert, so a single scalar suffices. The gap is positive in every run: training examples are routed to their matched expert more confidently than the population statistic predicts.

E.3 Derivation of the Volterra equation

Step 1: Onsager self-reaction. The router gradient flow (15) for the soft-discriminant gate gives, dropping the spherical projection at leading order,

$$\dot v_e(t) \;=\; -\nu_v(t)\,v_e(t) \;+\; \frac{a_1 s_E}{d}\sum_{\mu=1}^{n}\mathrm{res}^\mu(t)\,\delta F_e^\mu(t)\,x^\mu, \qquad \delta F_e^\mu \;=\; F_e^\mu - \tfrac{1}{E}\textstyle\sum_j F_j^\mu.$$

Here $s_E=\sqrt{E/(E-1)}$. At $E=2$ this is the familiar $(a_1/\sqrt{2})\,\Delta F_e$ with $\Delta F_e = F_0 - F_1$, since $\Delta F_e = 2\,\delta F_e$ and $s_E=\sqrt{2}$. When training example μ updates the router during $d\tau$ it moves *every* expert direction, $\delta v_j(\tau) = a_1 s_E\,\mathrm{res}^\mu(\tau)\,\delta F_j^\mu(\tau)\,x^\mu\,d\tau/d$. Router vector v_e therefore inherits its own kick through the diagonal router response $R_{dv}(t,\tau)$ and every other expert's kick through the off-diagonal $R_{vo}(t,\tau)$. Since $\sum_{j\neq e}\delta F_j = -\delta F_e$, the two combine into the *contrast* response

$$R_-^v(t,\tau) \;:=\; R_{dv}(t,\tau) - R_{vo}(t,\tau), \qquad \delta v_e(t) \;=\; R_-^v(t,\tau)\,a_1 s_E\,\mathrm{res}^\mu(\tau)\,\delta F_e^\mu(\tau)\,x^\mu\,d\tau/d. \tag{104}$$

The mean router direction is unmoved, $\delta\bar v(t)\propto\big(R_{dv}+(E-1)R_{vo}\big)\sum_j\delta F_j=0$, so the contrast picks up s_E a second time. Using $\|x^\mu\|^2/d\to 1$, the self-reaction on the same example's discriminant at time t is

$$\delta_\tau\big(\mathrm{disc}_e^\mu(t)\big) \;\approx\; R_-^v(t,\tau)\;a_1\,s_E^2\;\mathrm{res}^\mu(\tau)\,\delta F_e^\mu(\tau)\,d\tau, \qquad s_E^2=\frac{E}{E-1}. \tag{105}$$

Step 2: Cluster average. The kicks of the other examples reach disc^μ only through $x^\nu\cdot x^\mu/d$. Their mean is the population signal that the uncorrected DMFT already carries, and their fluctuation has zero conditional mean, so only the example's own kick contributes to δ_{disc}. Averaging (105) over the training examples of the cluster matched to expert e gives

$$\delta_{\mathrm{disc}}(t) \;=\; a_1\,\frac{E}{E-1}\int_0^t R_-^v(t,\tau)\,\mathbb{E}_{c=e}\big[\mathrm{res}(\tau)\,\delta F_e(\tau)\big]\,d\tau, \tag{106}$$

where $\mathbb{E}_{c=e}$ is conditioned on the cluster matched to expert e. The prefactor $a_1 s_E^2 = a_1 E/(E-1)$ is that of (105), with no further counting factor, and by the S_E symmetry every cluster gives the same correction for its own matched expert.

We verify (106) against gradient flow. In gradient flow the difference in (103) equals the contrast direction projected on the empirical cluster noise mean, $s_E(v_e-\bar v)\cdot\bar\xi_e$, which isolates the router's memorization of its own training set. In short runs over 36 seeds ($m=300$, $d=1200$), (106) gives 0.96 and 0.86 of the measured δ_{disc} at $E=4$ and 8, within the sampling error of about 27%. The measured values, 4.8×10^{-2} and 5.3×10^{-2}, are nearly independent of E, as the matched-cluster conditioning of (106) predicts, whereas a normalization with prefactor a_1/E and a single conditioning cluster would decay as $1/E$ (ratios 0.52 and 0.20).

Step 3: Differentiate. Apply Leibniz's rule to (106), using the equal-time identity $R^v_-(t,t) = R_{dv}(t,t) - R_{vo}(t,t) = 1$ and the contrast response equation. Subtracting the R_{vo} equation stated below (81) from (81), the intermediate-expert multiplicities cancel, the diagonal memory $M_R^{vd}R_{dv} + (E-1)M_R^{vo}R_{vo}$ minus the off-diagonal $M_R^{vd}R_{vo}+M_R^{vo}\big[R_{dv}+(E-2)R_{vo}\big]$ equals $\big(M_R^{vd}-M_R^{vo}\big)R^v_-$ for every E, so

$$\partial_t R^v_-(t,\tau) = \delta(t-\tau) - \nu_v(t)\, R^v_-(t,\tau) - \int_\tau^t \big[M_R^{vd}(t,s) - M_R^{vo}(t,s)\big] R^v_-(s,\tau)\, ds, \tag{107}$$

and therefore

$$\begin{aligned}\frac{d}{dt}\delta_{\rm disc}(t) \;=\; &\mathrm{Source}(t) \;-\; \nu_v(t)\,\delta_{\rm disc}(t) \\ &-\; a_1\frac{E}{E-1}\int_0^t d\tau \int_\tau^t ds\,\big[M_R^{vd} - M_R^{vo}\big](t,s)\, R^v_-(s,\tau)\,\mathbb{E}_{c=e}[\mathrm{res}(\tau)\,\delta F_e(\tau)],\end{aligned}$$

with $\mathrm{Source}(t) := a_1\frac{E}{E-1}\,\mathbb{E}_{c=e}[\mathrm{res}(t)\,\delta F_e(t)]$. Exchanging the order of integration in the double integral and recognizing the inner integral as $\frac{E-1}{a_1 E}\,\delta_{\rm disc}(s)$ via (106) with $t \to s$, we obtain the closed Volterra integro-differential equation of Section 4.3,

$$\frac{d}{dt}\delta_{\rm disc}(t) \;=\; \mathrm{Source}(t) \;-\; \nu_v(t)\,\delta_{\rm disc}(t) \;-\; \int_0^t \big[M_R^{vd}(t,s) - M_R^{vo}(t,s)\big]\,\delta_{\rm disc}(s)\, ds. \tag{108}$$

Both kernels are the memory kernels (57)–(58), closures already computed for the router Volterra equations of Appendix D, so (108) adds no new two-time matrices. Only their difference enters, because $\delta_{\rm disc}$ propagates along the contrast direction. In our discretization these kernels carry R_A in its cell-integral form and contribute at order h (Appendix D.4), so the memory term vanishes as $h \to 0$. At the settings of Figure 5, $\delta_{\rm disc}(50)$ is 0.535 at $h = 0.1$ and 0.553 without the memory term, which approximates the limit $h \to 0$.

E.4 Source term

The source $\mathrm{Source}(t) = a_1\frac{E}{E-1}\,\mathbb{E}_{c=e}[\mathrm{res}(t)\,\delta F_e(t)]$ is a Gaussian expectation over the teacher fields of a single example and needs no approximation beyond the closure already used for the drive. For an example of cluster $c = e$ let $\lambda_j = U_j \cdot x$, so that $\lambda \sim \mathcal{N}(\kappa\,\mathbf{e}_e, I_E)$, and let $\lambda_{\rm tot} = \sum_j \lambda_j$. Averaged over the neuron noise at the current order parameters, the expert outputs are $F_j = \bar{a}\,S(u_j)$ with

$$u_j = (P_{\rm self} - P_{\rm cross})\,\lambda_j + P_{\rm cross}\,\lambda_{\rm tot}, \quad S(u) = \mathrm{erf}\Big(\frac{u}{\sqrt{2(2-P_2)}}\Big), \quad P_2 = P_{\rm self}^2 + (E-1)P_{\rm cross}^2, \tag{109}$$

and the gates are $g_j = \frac{1}{E} + a_1 s_E\,\Delta_v\,(\lambda_j - \lambda_{\rm tot}/E)$. With the teacher output $\sigma(\lambda_e)$ and the student output $f = \sum_j g_j F_j$,

$$\mathrm{Source}(t) \;=\; a_1\,\frac{E}{E-1}\,\mathbb{E}_\lambda\Big[\big(\sigma(\lambda_e) - f\big)\Big(F_e - \tfrac{1}{E}\textstyle\sum_j F_j\Big)\Big], \tag{110}$$

a teacher term minus a student term. Only products of at most two expert outputs appear, so the expectation depends on the matched coordinate, two mismatched coordinates and the sum of the remaining $E - 3$, and Gauss–Hermite quadrature over these four variables evaluates it without further approximation. The source carries the amplitude $\bar{a}$ through F_j and contains no fitted constant. Measured directly in gradient flow (36 seeds, $m = 300$, $d = 1200$, where $\bar{a}$ stays at 0.50), $a_1\frac{E}{E-1}\,\mathbb{E}_{c=e}[\mathrm{res}\,\delta F_e]$ is 1.001, 1.015 and 1.027 times (110) at $E = 2, 4, 8$, each within two standard errors.

E.5 How the correction enters the dynamics

The expert drive receives the per-example self-reaction (105), not its cluster mean. Each training example's gate and residual are shifted by its own retarded kick, and because the kick is largest for the examples with large $\mathrm{res}\,\delta F_e$, the shifts are correlated with the quantities the drive averages. Appendix D.3 gives the resulting corrections: the gate channel (63), the residual channel (65), and

the four-term expansion (66) with its cross term. The residual that sources the kick is itself shifted by the self-reaction, and we place this feedback inside the accumulating memory integral, so the correction is solved self-consistently at each step. All reported results include these terms, and the ablation of Appendix E.6 removes them.

The mean shift $\delta_{\rm disc}$ is the prediction for the observable routing bias. In the reported results it enters the dynamics only through the expert constraint multiplier ν_w, evaluated at the shifted discrimination $r_D + \delta_{\rm disc}/\kappa$, an approximation, since the drive itself uses the per-example self-reaction, so changing $\delta_{\rm disc}$ by 3% changes $P_{\rm cross}$ by less than 10^{-8}. Shifting every example's routing score by the cluster mean instead of applying the per-example self-reaction captures part of the effect, but not all of it (Appendix E.6). Two further rules complete the specification.

1. The amplitude kernels $\hat{\Phi}_{\rm self/cross}$ of the $\bar{a}$ ODE carry no correction. This is an approximation: the $\bar{a}$ gradient (16) is also a training-set sum. Applying the mean shift to these kernels inflates the $\bar{a}$ source by 22.8%, making $\bar{a}$ grow too fast and $P_{\rm self}$ too slow, so we leave them uncorrected.

2. The router drive uses the training-set residual. The router gradient is $\propto \mathrm{res}\,\delta F_e\, x$, and δF_e depends on the expert weights rather than on v, so only the residual carries a self-reaction, and unlike (66) the expansion has two terms:

$$\tilde{\Phi}_v^{\rm tr}(t) \;=\; \mathbb{E}_\lambda\big[\mathrm{res}\,\delta F_0\,\lambda\big] \;-\; \mathbb{E}_\lambda\big[\delta f\,\delta F_0\,\lambda\big], \tag{111}$$

with δf as in (64). The second term is 6.0% of the router drive at $t = 25$ for $\bar{\alpha} = 1$, $m = 1200$, and it grows from zero with router specialization. The source (110) uses the population residual, so $\delta_{\rm disc}$ is the cluster mean of the first iterate of the self-reaction, while the expert drive uses the self-consistent one.

E.6 Numerical validation: ablating the correction

Figure 5 ablates the router self-reaction. Removing it from the expert drive, together with $\delta_{\rm disc}$, makes the DMFT predict monotone growth of $P_{\rm cross}$ (red dashed), missing the GF peak and decay. Restoring it cuts the trajectory RMSE on $P_{\rm cross}$ by $4.3\times$ ($1.4 \times 10^{-3} \to 3.3 \times 10^{-4}$ at $d = 3000$, 16 seeds) and on Δ by $9.3\times$. Feeding the expert drive only the cluster-mean shift $\delta_{\rm disc}$, with no per-example correlation, gives an intermediate result. $P_{\rm cross}$ still rises and decays, but peaks later and higher than in gradient flow (2.97×10^{-3} at $t = 33$, against 2.39×10^{-3} at $t = 27$), the RMSE is 5.3×10^{-4}, and the effect on $P_{\rm cross}$ and Δ at $t = 50$ is about half that of the full self-reaction. The right panel shows the predicted mean shift $\delta_{\rm disc}(t)$. As $h \to 0$ it reaches about 0.55 at $t = 50$ (dotted), 3% above the 0.536 ± 0.008 measured in gradient flow over 36 seeds at the same settings ($d = 3000$). At our step $h = 0.1$ the order-h memory term lowers it to 0.535 (solid, Appendix E.7).

E.7 Discrete implementation

In the forward-Euler scheme of Appendix D with step size h, equation (108) is integrated as

$$\delta_{\rm disc}[n+1] = \delta_{\rm disc}[n] + h\Big(-\nu_v[n]\,\delta_{\rm disc}[n] + \mathrm{Source}[n] - h\sum_{j=0}^{n}\big[M_R^{vd}[n,j] - M_R^{vo}[n,j]\big]\,\delta_{\rm disc}[j]\Big), \tag{112}$$

with $\mathrm{Source}[n]$ obtained by evaluating (110) at the current order parameters. The kernels $M_R[n,j]$ contain R_A in the cell-integral form (85), so the memory sum is of order h and vanishes as $h \to 0$. The memory sum reuses $M_R^{vd}[n,j]$ and $M_R^{vo}[n,j]$ already computed for the R_{dv} and R_{vo} Volterra equations, so $\delta_{\rm disc}$ adds one scalar trajectory. The per-example self-reaction is accumulated in the same way: at every node of the drive's quadrature grid the solver stores the per-step source $\delta F\,\mathrm{res}$ and forms the memory sum with R_-^v, which adds one grid-valued row per step and no new two-time matrices.

Alongside (112), the router multiplier enters in its population-plus-Onsager form (96), and the equal-time off-diagonal router correlator is integrated from its own equation (102).

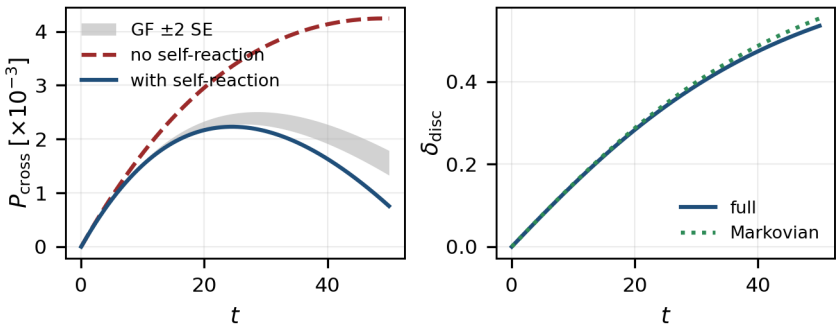


Figure 5: **The router self-reaction is required for quantitative DMFT–GF agreement.** *Left*: $P_{\text{cross}}(t)$ for GF (gray, ± 2 SE over 16 seeds) and the DMFT with and without the per-example router self-reaction, the arm without it also setting $\delta_{\text{disc}} = 0$ (RMSE 3.3×10^{-4} against 1.4×10^{-3}). *Right*: the predicted mean shift $\delta_{\text{disc}}(t)$ of (108) at step $h = 0.1$ (solid) and without its memory term (dotted). The memory term is of order h in our discretization, so the dotted curve approximates the limit $h \to 0$ (Appendix E.7). Settings: $E = 2$, $\bar{\alpha} = \kappa = 1$, $a_1 = 0.3$, $\sigma(z) = \text{erf}(z/\sqrt{2})$, $d = 3000$, $m = 300$, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, integration step 0.1.

F REM Correspondence: Proofs and Quenched Sampling

This appendix supplies the proofs of the softmax-to-REM identification of Section 5.1, the derivation of the matched-condensation temperature (11), the relation between the closure and the quenched and annealed averages, and the sampling that an untruncated treatment of the glassy phase would need. (The numerical validation of the sharpening with E is reported in Section 4.4 and Appendix I.)

F.1 Decomposition of router preactivations under symmetric initialization

Consider the router preactivation $z_e^v(t) = v_e(t)\cdot x$ for an input $x = \kappa U_c + \xi$ from cluster c, with signal $\kappa P_{e,c}^v(t)$ and noise part $v_e(t)\cdot\xi$. Under the diagonal-S_E ansatz of Appendix C.7, the noise parts have equal-time covariance

$$\text{Cov}\big(v_e(t)\cdot\xi,\ v_{e'}(t)\cdot\xi\big) \ = \ \begin{cases} 1 & e = e' \\ C_o^v(t,t) & e \neq e' \end{cases} \tag{113}$$

(the diagonal follows from $\|v_e\| = 1$ and the off-diagonal is the router correlator C_o^v of Appendix B.1). The covariance matrix $(1 - C_o^v)I_E + C_o^v \mathbf{1}_E \mathbf{1}_E^\top$ is positive semidefinite for $C_o^v \geq -1/(E-1)$, and for every such C_o^v the noise decomposes as

$$v_e(t)\cdot\xi \ = \ c(t) \ + \ \sqrt{1 - C_o^v(t,t)}\,\eta_e(t), \qquad c(t) \ = \ \sqrt{\tfrac{1+(E-1)C_o^v}{E}}\,Z(t) \ - \ \sqrt{1 - C_o^v}\,\bar{\eta}(t), \tag{114}$$

with $Z, \eta_e \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\bar{\eta} = \frac{1}{E}\sum_j \eta_j$. The right-hand side has the covariance (113), and the common mode $c(t)$ is the same for every expert, so it cancels in the translation-invariant softmax, which sees only the i.i.d. part $\sqrt{1 - C_o^v}\,\eta_e$. For $C_o^v \geq 0$ the simpler form $\sqrt{C_o^v}\,Z + \sqrt{1 - C_o^v}\,\eta_e$ is equivalent, but gradient flow drives C_o^v toward $-1/(E-1)$ (Appendix F.8), where only (114) applies.

F.2 Identification with the Random Energy Model

The softmax routing weights at temperature T are

$$g_e(t) \ = \ \frac{\exp(z_e^v(t)/T)}{\sum_{e'}\exp(z_{e'}^v(t)/T)} \ = \ \frac{\exp(\eta_e(t)/T_{\text{eff}})}{\sum_{e'}\exp(\eta_{e'}(t)/T_{\text{eff}})}, \qquad T_{\text{eff}}(t) \ = \ \frac{T}{\sqrt{1 - C_o^v(t,t)}}, \tag{115}$$

where the second equality follows from (114) and the cancellation of the common-mode $c(t)$ in the translation-invariant softmax.

REM identification. Equation (115) is exactly the Boltzmann distribution of Derrida (1981)'s Random Energy Model (the same identification was made for dense associative memories by Lucibello & Mézard, 2024 and for attention scores by Giorlandino & Goldt, 2026) with E microstates labeled by $e \in \{1, \dots, E\}$, quenched random energies $\mathcal{E}_e = -\eta_e \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and inverse temperature $\beta = 1/T_{\text{eff}}$. The REM partition function $Z(\beta) = \sum_e e^{-\beta \mathcal{E}_e}$ exhibits a freezing transition at the critical inverse temperature $\beta_c = \sqrt{2 \ln E}$, giving the bare freezing temperature

$$T_c^{(0)} \;=\; \frac{1}{\beta_c} \;=\; \frac{1}{\sqrt{2 \ln E}}. \tag{116}$$

The crossover is determined by setting the annealed free energy $-T_{\text{eff}} \ln \mathbb{E}[Z] = -T_{\text{eff}}(\ln E + \beta^2/2)$ equal to the frozen-phase free energy $-\sqrt{2 \ln E}$ (the ground-state energy, which is temperature-independent) at the touching point β_c (Derrida, 1981). Equivalently, the annealed entropy $\ln E - \beta^2/2$ vanishes there, giving $\beta_c = \sqrt{2 \ln E}$. For $T_{\text{eff}} < T_c^{(0)}$ the partition function condenses onto a few of the largest η_e, with $Y_2 := \sum_e g_e^2$ averaging $1 - T_{\text{eff}}/T_c^{(0)}$ as $E \to \infty$. For $T_{\text{eff}} > T_c^{(0)}$ the routing remains diffuse, and $Y_2 \to 0$ as E grows.

F.3 Signal-modified condensation onto the matched expert

Under the initialization (3), with $P^v_{\text{self}}(0) = \varepsilon^v$ and $P^v_{\text{cross}}(0) = 0$, the cluster-conditional router preactivation acquires a deterministic mean shift. For input $x = \kappa U_c + \xi$ from cluster c:

$$z_e^v(t) \;=\; v_e(t) \cdot x \;=\; \underbrace{\kappa\, P^v_{e,c}(t)}_{\text{signal (deterministic)}} \;+\; \underbrace{v_e(t) \cdot \xi}_{\text{noise}}. \tag{117}$$

Substituting the decomposition (114) of the noise term and dropping the common mode $c(t)$, which cancels under the translation-invariant softmax,

$$z_e^v(t) \;=\; \kappa\, P^v_{e,c}(t) \;+\; \sqrt{1 - C_o^v(t,t)}\, \eta_e(t). \tag{118}$$

The matched expert ($e = c$) sees the signal κP^v_{self}. The $E - 1$ mismatched experts see the smaller signal $\kappa P^v_{\text{cross}}$. The matched-state advantage is $\kappa\, \Delta_v(t)$ with $\Delta_v := P^v_{\text{self}} - P^v_{\text{cross}}$.

Matched-state condensation temperature. As T_{eff} decreases, matched-state condensation begins when the matched advantage exceeds the annealed free energy of the mismatched subset, which includes both its entropy and its Gaussian bulk energy. Writing the matched advantage in effective-temperature units as $s := \kappa\, \Delta_v(t)/\sqrt{1 - C_o^v(t,t)}$, the annealed comparison is

$$\beta_{\text{eff}}\, s \;>\; \ln(E-1) \;+\; \tfrac{1}{2}\beta_{\text{eff}}^2, \tag{119}$$

the last term being the annealed bulk free energy of the $E - 1$ mismatched states, which the bare-freezing derivation in Appendix F.2 already carries (the annealed free energy there is $-T_{\text{eff}}(\ln E + \beta^2/2)$). Solving the quadratic $\frac{1}{2}\beta_{\text{eff}}^2 - s\,\beta_{\text{eff}} + \ln(E-1) < 0$ gives condensation for β_{eff} between its two roots, that is, for T_{eff} between the two branches

$$T_c^{\text{eff},\pm}(t) \;=\; \frac{s \pm \sqrt{s^2 - 2\ln(E-1)}}{2\ln(E-1)}, \tag{120}$$

which requires a real root, $s \geq \sqrt{2\ln(E-1)}$: a *minimum-signal* condition for condensation. Two remarks. First, the upper branch $T_c^{\text{eff},+}$ is the physical condensation line. The lower branch lies at or below the bulk freezing temperature $1/\sqrt{2\ln(E-1)}$, just above $T_c^{(0)}$, because its inverse temperature is at least $\sqrt{2\ln(E-1)}$. There the annealed treatment of the mismatched bulk is invalid: the frozen bulk's largest noise is about $\sqrt{2\ln(E-1)} < s$, so the matched expert stays condensed at every temperature below $T_c^{\text{eff},+}$, and the lower branch is spurious. Second, at large signal $T_c^{\text{eff},+} \to \kappa\, \Delta_v/(\sqrt{1 - C_o^v}\, \ln(E-1))$, growing linearly in the signal, with $\ln(E-1)$ setting the slope. The positive feedback loop (13) follows: Δ_v grows during specialization, raising s, lifting $T_c^{\text{eff},+}$ and deepening the matched-state condensation.

Specialization vs frozen-collapse boundary. Matched-state condensation is possible only above the minimum-signal threshold implied by (120), which separates the specialization and frozen-collapse regimes:

$$s \;\geq\; \sqrt{2\ln(E-1)} \quad\Longleftrightarrow\quad \kappa\,\Delta_v(t) \;\geq\; \sqrt{1-C_o^v(t,t)}\,\sqrt{2\ln(E-1)} \;=:\; \kappa\,\Delta_v^*. \tag{121}$$

Above this threshold, cooling from high T first crosses $T_c^{\mathrm{eff},+} > T_c^{(0)}$ and condenses onto the matched expert (specialization phase). Below it, no matched-state condensation exists above the bare floor, bare-REM freezing happens first, and the router condenses onto a random microstate-like configuration (frozen-collapse phase). At $\kappa = 1$ and $C_o^v = 0$ this gives $\Delta_v^* = 1.48$, 1.97 and 2.33 for $E = 4$, 8 and 16 respectively. No unit-norm router reaches these values: by (124), $\Delta_v \leq \sqrt{1-C_o^v}$, so $s \leq \kappa$, and at leading order in $\ln E$ matched-state condensation is accessible only for $\kappa \geq \sqrt{2\ln(E-1)}$. Figure 6 shows the geometry at $\kappa = 3$, where $\Delta_v^* \approx 0.78$ lies inside the reachable range: the condensation branch $T_c^{\mathrm{eff},+}(\Delta_v)$ emerges at Δ_v^*, just above the bare-freezing floor, and rises to the right of it.

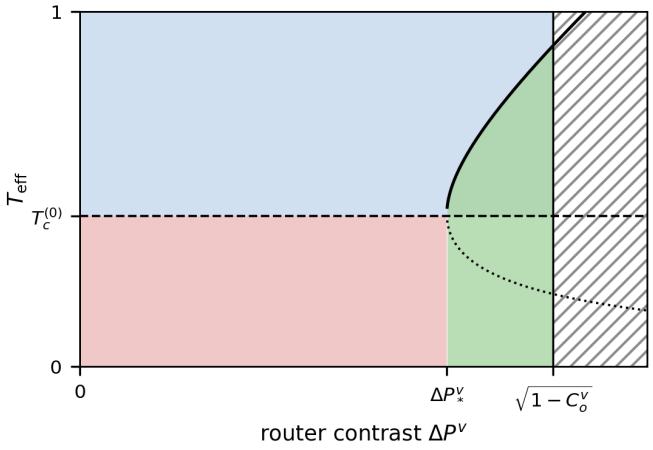


Figure 6: **Matched-condensation temperature versus the bare freezing temperature** at $E = 16$, $\kappa = 3$, $C_o^v = 0$. Dashed: $T_c^{(0)} = 1/\sqrt{2\ln E} \approx 0.42$. Solid: the matched-state condensation branch $T_c^{\mathrm{eff},+}(\Delta_v)$ of (120), which exists for $\Delta_v \geq \Delta_v^* = \sqrt{1-C_o^v}\sqrt{2\ln 15}/\kappa \approx 0.78$ and emerges at $T \approx 0.43$, just above the bare floor. Hatched: $\Delta_v > \sqrt{1-C_o^v}$, which no unit-norm router reaches (Appendix F.8). At $\kappa = 1$ the threshold would be 2.33, inside the hatched region. Left of Δ_v^* the matched advantage cannot lift condensation above the bare floor, and the router risks frozen collapse. Right of it the router condenses onto the matched expert for every $T < T_c^{\mathrm{eff},+}(\Delta_v)$: above the bulk freezing temperature by the annealed comparison, and below it because the matched head start s exceeds the largest bulk noise, $\sqrt{2\ln 15}$. The feedback loop (13) raises $T_c^{\mathrm{eff},+}$ as Δ_v grows. The dotted lower branch of (120) is spurious: it falls below $T_c^{(0)}$ just past threshold, where the annealed treatment of the mismatched bulk is invalid.

F.4 Quenched, annealed, and the closure

With the per-example action X_1 of (32), two averages over the routing randomness can be formed: the quenched average $\mathbb{E}_{z^v}[X_1] = \kappa_1(X_1)$ and the annealed one $\mathcal{F} = \ln \mathbb{E}_{z^v}[e^{X_1}] = \sum_{\ell\geq 1} \kappa_\ell(X_1)/\ell!$. The data average of the MSRDJ generating functional produces $\mathcal{F}$ (Appendix C.4), so $\mathcal{F}$ is the target. By Jensen's inequality $\mathcal{F} \geq \kappa_1$, with equality only when X_1 does not fluctuate, and $\mathcal{F} - \kappa_1 = \sum_{\ell\geq 2} \kappa_\ell/\ell!$. The second-cumulant closure is neither average: it evaluates the log-determinant at the mean covariance and so also drops the Jensen gap (34) inside κ_1. Its error is therefore not measured by $\mathcal{F} - \kappa_1$. Where routing is diffuse and E is large, the gate fluctuations that X_1 depends on shrink (Appendix C.4), and the three are expected to agree to leading order. In the specialization and frozen-collapse regimes each input is dominated by one expert, X_1 fluctuates at order one, and we measure the error of the DMFT against gradient flow instead (Section 4.4).

F.5 UNTRUNCATED EVALUATION BY SAMPLING

Evaluating $\mathcal{F}$ without the closure requires sampling each example's routing trajectory z^v from its Gaussian law at the tracked two-point functions, evaluating X_1 on each sample, and averaging e^{X_1}. No result reported in this paper uses such sampling. Its main difficulty is weight degeneracy: when X_1 fluctuates at order one, as in the glassy regime, e^{X_1} is dominated by rare trajectories and the effective sample size $(\sum_s e^{X^{(s)}})^2/\sum_s e^{2X^{(s)}}$ can collapse to order one. Annealed importance sampling along a temperature ladder, with resampling whenever the effective sample size falls, is the standard remedy.

F.6 BARE-SOFTMAX THERMODYNAMICS: THE PARTICIPATION RATIO

Section 4.4 tests the cumulant truncation on softmax *trained dynamics*, by comparing the linear-DMFT solver against softmax-GF. A purely thermodynamic check on the same temperature scale, with no gradient flow and no signal, is given by sampling bare softmax weights $g_e \propto \exp(\eta_e/T)$ on iid energies $\eta_e \sim \mathcal{N}(0,1)$ and measuring the participation ratio $Y_2(T) = \sum_e g_e^2$, Derrida's REM order parameter. As $E \to \infty$, $\langle Y_2\rangle = 1 - T/T_c^{(0)}$ for $T < T_c^{(0)}$ and $\langle Y_2\rangle \to 0$ for $T > T_c^{(0)}$ (Derrida & Toulouse, 1985; Mézard et al., 1987). Figure 2 shows Y_2 at $E \in \{4, 16, 64, 512\}$ (5×10^4 samples per (E,T)) against this limit. At $T_c^{(0)}$ the finite-E values are 0.52, 0.38, 0.31 and 0.25, so the approach to the limit is slow in E, and at the expert counts used in practice freezing is a crossover centered near $T_c^{(0)}$.

F.7 LOAD IMBALANCE FROM INPUT-SHARED LOGITS

In the symmetric model each input's frozen winner is picked by that input's own noise $\eta_e(x)$, so every expert carries load $1/E$ at every temperature. Load imbalance requires a logit component that is shared by all inputs but differs between experts, such as router biases or unequal overlaps with a common input direction. We model it by

$$z_e(x) \;=\; \sqrt{a}\, b_e + \sqrt{1-a}\, \eta_e(x), \qquad b_e,\ \eta_e(x) \text{ i.i.d. } \mathcal{N}(0,1), \tag{122}$$

with b_e fixed across inputs and $\eta_e(x)$ drawn afresh for each input, so $a \in [0,1]$ is the shared fraction of the unit logit variance, and $a = 0$ in the symmetric model. The load of expert e is $L_e = \mathbb{E}_x[g_e(x)]$, and $N_{\rm load} = 1/\sum_e L_e^2$ is the effective number of experts that carry it, the analog of $1/Y_2$ for the load.

For each input the logits (122) are E independent unit-variance Gaussians, so per-input routing freezes at $T_c^{(0)}$ for every a. For the load, write $\hat\beta = T_c^{(0)}/T$ and measure both components in units of $\sqrt{2\ln E}$, $b_e = p\sqrt{2\ln E}$ and $\eta_e = q\sqrt{2\ln E}$. About E^{1-p^2} experts have shared level $p \le 1$, and a given input gives one of them the level q with probability E^{-q^2}. Below $T_c^{(0)}$ an input's weight goes to the experts whose logit lies within $\mathcal{O}(1)$ of the maximum $\sqrt{2\ln E}$, so an expert at level p wins an input when $\sqrt{a}\,p + \sqrt{1-a}\,q \ge 1$, and its load from winning scales as $E^{-(1-\sqrt{a}\,p)^2/(1-a)}$, independent of T. Summing L_e^2 over levels, the exponent $1 - p^2 - 2(1-\sqrt{a}\,p)^2/(1-a)$ is maximal at $p = 2\sqrt{a}/(1+a)$, which gives

$$N_{\rm load} \;\asymp\; E^{\lambda}, \qquad \lambda \;=\; \frac{1-a}{1+a} \qquad (T \le T_c^{(0)}). \tag{123}$$

Experts that do not win an input still collect small, T-dependent weights, but for every $T \le T_c^{(0)}$ their contribution to $\sum_e L_e^2$ is of lower order in E. Above $T_c^{(0)}$ the same counting, with the unfrozen per-input partition function $\ln Z \simeq (1+\hat\beta^2)\ln E$, gives $\lambda = 1 - 2a\hat\beta^2$ for $T \ge \max(\sqrt{1+2a},\, 2\sqrt{a})\, T_c^{(0)}$, and for $a \le 1/2$ the exponent decreases continuously to (123) as T falls to $T_c^{(0)}$. Load imbalance therefore freezes at the same temperature as per-input routing: cooling below $T_c^{(0)}$ concentrates each input on fewer experts but does not concentrate the load further. The frozen imbalance is set by the shared fraction, from balanced load ($\lambda = 1$) at $a = 0$ to collapse onto $\mathcal{O}(1)$ experts ($\lambda \to 0$) as $a \to 1$.

Sampling (122) at E from 64 to 16384 (three draws of b, 10^4 inputs each) is consistent with the freezing. At $a = 1/2$, lowering T from $T_c^{(0)}$ to $0.5\,T_c^{(0)}$ lowers $N_{\rm load}$ by a factor between 1.4 and

1.7 that drifts only slowly with E, while lowering it from $1.5\,T_c^{(0)}$ to $T_c^{(0)}$ lowers $N_{\rm load}$ by a factor that grows from 1.5 to 3.2 over the same range. At and below $T_c^{(0)}$ the fitted exponents are 0.62 to 0.66, 0.41 to 0.44 and 0.25 to 0.30 at $a = 1/4, 1/2, 3/4$, against 0.60, 0.33 and 0.14, and at $1.5\,T_c^{(0)}$ they are 0.76 and 0.57 at $a = 1/4, 1/2$, against 0.78 and 0.56. The approach to the leading-order exponents is slow at large a, as for Y_2 in Figure 2.

Load-balancing losses penalize the deviation of L_e from $1/E$. At $a = 0$ the load is balanced at every temperature, so at leading order these losses exert no force against frozen collapse, and they act only on the shared component that sets a.

F.8 The signal bound and the condensation threshold in κ

Let $M = V^\top U \in \mathbb{R}^{E\times E}$ collect the router–cluster overlaps, $M_{ec} = v_e \cdot U_c$, which under the diagonal S_E symmetry is $M = \Delta_v\, I + P^v_{\rm cross}\mathbf{1}\mathbf{1}^\top$, and let $G = V^\top V = (1 - C^v_o)I + C^v_o\mathbf{1}\mathbf{1}^\top$ be the router Gram matrix. Because the columns of U are orthonormal, $UU^\top \preceq I$, so $MM^\top = V^\top UU^\top V \preceq G$. On vectors orthogonal to $\mathbf{1}$ the left side acts as $(\Delta_v)^2$ and the right side as $1 - C^v_o$, hence

$$(\Delta_v)^2 \;\le\; 1 - C^v_o, \qquad s \;=\; \frac{\kappa\,\Delta_v}{\sqrt{1 - C^v_o}} \;\le\; \kappa, \tag{124}$$

with equality when the router contrast lies in the span of the cluster directions. The threshold (12) requires $s \ge \sqrt{2\ln(E-1)}$, so at leading order in $\ln E$ condensation onto the matched expert is impossible for $\kappa < \sqrt{2\ln(E-1)}$, which is 1.48, 1.97 and 2.33 at $E = 4, 8, 16$. The gradient-flow runs of Figure 4 use $\kappa = 1$, and Appendix F.9 shows runs at $\kappa = 1$ and $\kappa = 3$.

We test this in softmax gradient flow at $E = 4$, $\bar\alpha = 4$, $d = 2000$, $m = 120$, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, $\kappa \in \{1, 2, 3\}$ and $T \in \{0.4, 0.8, 1.2\}$ (two seeds each, $t \le 400$). In every run C^v_o settles near $-1/(E-1)$ and s stays below κ, with s/κ ending between 0.56 and 0.68 at $\kappa = 1$ (after peaking at up to 0.79), 0.80 and 0.88 at $\kappa = 2$, and 0.88 and 0.92 at $\kappa = 3$. The mean gate weight on the matched expert is 0.43 to 0.53 at $\kappa = 1$ ($s < 1.48$), 0.62 to 0.79 at $\kappa = 2$ and 0.78 to 0.94 at $\kappa = 3$, falling with T at each κ. Evaluating the REM at each run's measured s and $T_{\rm eff}$ predicts these weights to within 0.05 to 0.14, with gradient flow, measured on training inputs, always the more confident. On 10^4 fresh inputs the same trained routers match the REM prediction to within 0.004 in every cell, for example 0.392 against 0.396 at $\kappa = 1$, $T = 0.4$, where the training inputs give 0.533. The excess confidence on training inputs is therefore the router's self-reaction of Section 4.3, measured directly. At $E = 4$ the dependence on s is smooth. The threshold sharpens with E: at $T_{\rm eff} = T_c^{(0)}/2$, the matched weight as a function of $s/\sqrt{2\ln(E-1)}$ passes through 0.66 at 1 for every E from 4 to 4096, while at 0.6 it falls from 0.50 to 0.12 and at 1.4 it rises from 0.81 to 0.97 over that range.

Figure 7 repeats the $E = 8$ gradient-flow grid of Figure 4 at $\kappa = 3$, where $s_* = \sqrt{2\ln 7} \approx 1.97$ lies below the bound $s \le 3$ (five values of $\bar\alpha \in [0.3, 4]$ by six of $T \in [0.1, 1.5]$, $d = 2500$, $m = 120$, training time 200, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$). All three regimes appear, with Y_2^∞ measured on the training inputs. At small $\bar\alpha$ and low T routing is frozen ($Y_2^\infty \approx 1$) and sends 17 to 26% of fresh inputs to their matched expert, 1.4 to 2 times the chance level $1/8$. At each cell's measured s and $T_{\rm eff}$ the REM places eight cells in the specialization regime ($s \ge s_*$ and $T_{\rm eff} < T_c^{\rm eff,+}$). In all eight the router gives the matched expert a mean weight of 0.54 to 0.80 on fresh inputs, more than all other experts together. The three cells with $s \ge s_*$ at $T = 1.5$ lie above the REM line. There the matched weight is 0.39 to 0.46, although 81 to 89% of fresh inputs still give the matched expert their largest weight, so the fraction routed to the matched expert does not by itself identify the regime. At $E = 8$ the threshold is a crossover: four cells with s from 1.50 to 1.97 also give the matched expert more than half the weight (0.52 to 0.64). At $T = 1.5$ routing is least concentrated ($Y_2^\infty = 0.38$ to 0.54). Among all cells with $Y_2^\infty \ge 0.9$ the matched fraction is 17 to 70% where $s < s_*$ and 74 to 86% where $s \ge s_*$. At each cell's measured s and $T_{\rm eff}$ the REM predicts the matched gate weight on fresh inputs to within 0.011 and the matched fraction to within 0.012, in all 30 cells.

F.9 Empirical phase regions: GF trajectories

Section 5 draws the regimes in the head start s and $T_{\rm eff}$ (Figure 4, left). Figure 8 overlays softmax-GF trajectories at $E = 4$, where the condensation threshold is $s_* = \sqrt{2\ln 3} \approx 1.48$. At $\kappa = 1$ no run

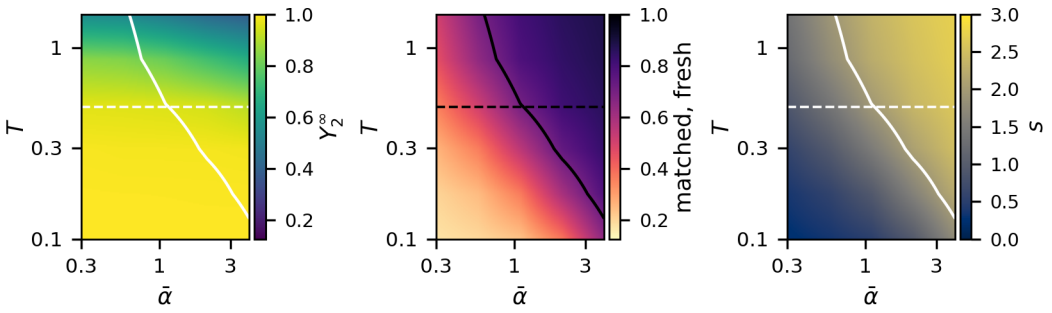


Figure 7: **Gradient flow at $E = 8$, $\kappa = 3$.** Final Y_2^∞ on the training inputs (left), fraction of fresh inputs routed to their matched expert (center) and head start s (right), against $\bar{\alpha}$ and T. Dashed: $T_c^{(0)} \approx 0.49$. Solid: the contour $s = s_* = \sqrt{2 \ln 7}$. Below $T_c^{(0)}$ the contour marks the crossover from frozen collapse (left) to condensation onto the matched expert (right). Other settings as in Figure 4.

can reach it, since $s \leq \kappa$ (Appendix F.8), and the three $\kappa = 1$ runs show one observed behavior each. The democratic run ($T = 1.20$, $\bar{\alpha} = 1$, small ε) stays in the high-T_{eff} region. The frozen-collapse run ($T = 0.30 < T_c^{(0)}$, small ε) stays at small s below the freezing line. The expert-specialization run ($T = 0.45$, $\bar{\alpha} = 4$, large ε) grows furthest, to $\Delta_v \approx 0.83$ ($s \approx 0.75$) before Δ_v settles at 0.74, while routing stays below $T_c^{(0)}$. Its experts specialize through the onset without the router condensing onto them. At $\kappa = 3$ the threshold is reachable. Both runs from the sweep of Appendix F.8, at $T = 0.4$ and 1.2, cross s_* by $t \approx 4$ and end in the matched-condensation region, with 95% and 94% of their training inputs, and 92 to 94% of fresh inputs, routed to their matched expert. The run at $T = 0.4$ starts below $T_c^{(0)}$ and leaves the frozen-collapse region once its head start passes s_*. The $\kappa = 1$ runs stop at training times 80, 200 and 300, before reaching equilibrium. This is a feasibility constraint, not a statement about the regions' attractor structure.

G Cross-Router Specialization Comparison

How fast the seed grows depends on the gate. Expanding the teacher activation and the gate response in Hermite polynomials, the natural basis for functions of a Gaussian, the heuristic of Appendix H ties the leading order of $\Lambda(t)$ to the joint information exponent $k^* := \min\{k \geq 1 : \hat{\sigma}_k \hat{\rho}_k \neq 0\}$, with $\hat{\sigma}_k$ and $\hat{\rho}_k$ the k-th Hermite coefficients of the teacher activation and the gate response. For most gates the linear terms already overlap ($k^* = 1$), while a gate without a linear term must rely on weaker higher-order terms.

We run gradient flow at $E = 2$ with four routers that differ only in the gate function: linear-discriminant, quadratic-probit and softmax-T, all with $k^* = 1$, and a cubic-Hermite gate $g_e = \frac{1}{2} + a_3 \, \mathrm{He}_3(\mathrm{disc}_e)$ with $a_3 = a_1/\sqrt{6}$, clipped to $[10^{-4}, 1 - 10^{-4}]$. Without the clipping this gate has $\hat{\rho}_1 = 0$ and $k^* = 3$. The clipping gives it a negative linear coefficient, $\hat{\rho}_1 \approx -0.057$ against $+0.30$ for the linear gate, and roughly halves $\hat{\rho}_3$, so the simulated gate has $k^* = 1$ with the opposite sign. Its router contrast turns negative, the direction a negative $\hat{\rho}_1$ favors, its routing stays nearly uniform, and its experts still specialize slightly (Figure 9). This run therefore does not test the case $k^* = 3$. Hard top-k is not differentiable and is treated separately in Appendix C.6, where its gate sector is exactly solvable.

H Hermite-Spectrum Router Observations and Future Directions

This appendix collects observations on how the joint Hermite spectrum $\hat{\sigma}_k \hat{\rho}_k$ of the teacher activation and gate response could organize routers by the speed of their onset. We define the joint information exponent k^*, give a heuristic for its effect on the onset, catalog the four canonical routers and their Hermite spectra, and outline future directions. The framing extends the single-index information-exponent argument of Ben Arous et al. (2021) to the coupled router–experts setting. It is a heuristic:

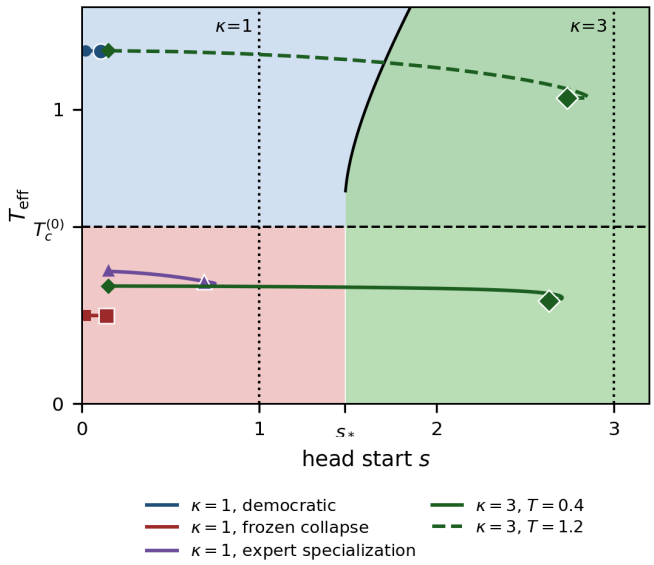


Figure 8: **GF trajectories at $E = 4$ in the head start s and T_{eff}.** Boundaries: $T_c^{(0)}$ (dashed), the condensation branch $T_c^{\text{eff},+}$ (11) (solid) from $s_* = \sqrt{2 \ln 3} \approx 1.48$, and the bound $s \leq \kappa$ (dotted) at $\kappa = 1$ and $\kappa = 3$. Small and large markers give the start and end of each run. At $\kappa = 1$ ($d = 2000$, $m = 120$) all three runs stay left of $s = 1$. For these runs C_o^v was not recorded, and s and T_{eff} use an approximation of C_o^v from the router–teacher overlaps, which lies above the value gradient flow reaches. The measured value would move the curves slightly toward smaller s and T_{eff} without changing the region each run occupies. At $\kappa = 3$ ($\bar{\alpha} = 4$, $d = 2000$, $m = 120$, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$) both runs cross s_* and end in the matched-condensation region.

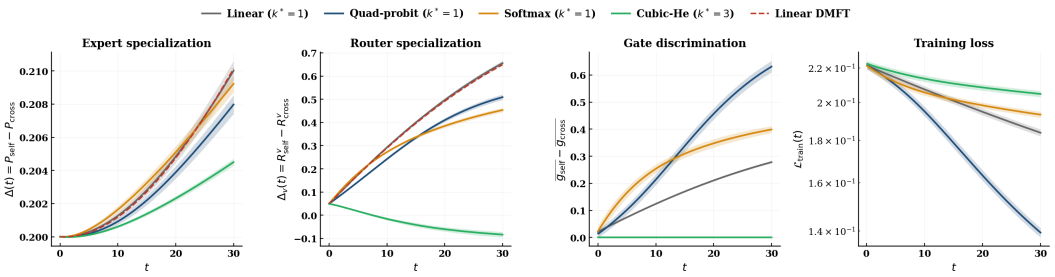


Figure 9: **Per-router specialization at $E = 2$** ($d = 3000$, $m = 200$, $T_{\max} = 30$, 4 seeds, bands ± 2 SE, $a_1 = 0.3$, $a_2 = 0.15$, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, softmax at $T = 0.5$, linear-gate DMFT dashed). Panels: Δ, Δ_v, the mean gate discrimination $\bar{g}_{\text{self}} - \bar{g}_{\text{cross}}$ and the training loss. The three $k^* = 1$ routers align on similar timescales. The gate labeled $k^* = 3$ is the clipped cubic gate, whose clipping gives it a negative linear coefficient: its router contrast drifts from 0.05 to -0.08, as a negative linear coefficient drives it, while its mean gate discrimination stays near zero. The separation is in the *router* order parameter (Δ_v rises from 0.05 to between 0.45 and 0.66 for $k^* = 1$) rather than the expert overlap (Δ rises from 0.200 to between 0.208 and 0.210 for $k^* = 1$, and to 0.2045 for the cubic gate).

our simulations do not isolate k^*, because the only simulated gate without a linear term is clipped, which restores one (Appendix G).

H.1 INFORMATION EXPONENT AND DESIGN PRINCIPLE

Both the teacher activation σ and the routing function r are square-integrable under standard Gaussian inputs and admit (probabilist's) Hermite expansions

$$\sigma(z) \;=\; \sum_{k\ge 0} \frac{\hat\sigma_k}{k!}\,\mathrm{He}_k(z), \qquad r_e(x)\,\big|_{V^\top x=z} \;=\; \bar r_e \;+\; \sum_{k\ge 1} \frac{\hat\rho_k}{k!}\,\mathrm{He}_k(z_e) \;+\; \text{cross-coordinate terms}, \tag{125}$$

where He_k is the k-th probabilist's Hermite polynomial. The coefficient $\hat\sigma_k$ measures how strongly σ depends on the k-th-order Hermite mode of its argument. For $\sigma(z) = \mathrm{erf}(z/\sqrt{2})$ all odd-k coefficients are nonzero. Likewise $\hat\rho_k$ measures how strongly the gate r_e responds to the k-th-order Hermite mode of its own preactivation z_e^v. We define the *joint information exponent* of the MoE pair (σ, r) by

$$k^* \;:=\; \min\{\, k \ge 1 \,:\, \hat\sigma_k\,\hat\rho_k \neq 0 \,\}. \tag{126}$$

When every Hermite coefficient of the gate is nonzero, (126) reduces to the classical teacher-only exponent of Ben Arous et al. (2021), the smallest k with $\hat\sigma_k \neq 0$. Our definition extends it to the joint object that enters the linearized MoE dynamics.

Hermite structure of the router-mediated coupling. Substituting (125) into the linearized equations of motion (Appendix H.2) and applying Mehler's formula (the closed-form evaluation of $\mathbb{E}[\mathrm{He}_j(G_1)\mathrm{He}_k(G_2)] = j!\,\delta_{jk}\,c^j$ for centered jointly Gaussian G_1, G_2 with cross-correlation c), the router-mediated coupling of the linearized system, the geometric mean $\sqrt{BC}$ of its off-diagonal entries, scales at small initialization scale ε of (3) as

$$\sqrt{B(t)\,C(t)} \;\simeq\; \gamma(t)\,\big|\hat\sigma_{k^*}\,\hat\rho_{k^*}\big|\,\big(\Delta_v(t)\big)^{k^*-1}, \tag{127}$$

where $\gamma(t) > 0$ depends on $\bar\alpha$, the second-layer amplitude $\bar a(t)$ and the integrated residual response $\chi_R(t) = \int_0^t \mathcal{R}_A(t,s)\,ds$, and carries a factor $m^{-1/2}$ because the expert row of the linearized system carries the explicit $1/m$ (Appendix H.2). Only the k^*-th joint Hermite product survives at leading order, because lower orders vanish by the definition of k^* and higher orders are suppressed by additional powers of the small overlap. For $k^* = 1$ the diagonal entries are small and $\sqrt{BC}$ is the growth rate $\Lambda(t)$ at leading order. For $k^* \ge 2$ it is not, and since $\Delta_v(0) = \varepsilon^v$ the two cases behave qualitatively differently.

Case $k^* = 1$. When $\hat\sigma_1\hat\rho_1 \neq 0$, the right side of (127) is independent of Δ_v, so $\Lambda(0) \simeq \sqrt{BC}$ does not vanish as the seed shrinks. The linearized dynamics satisfy $\delta P(t) \sim \varepsilon\,\exp(\int_0^t \Lambda(s)\,ds)$. If Λ stayed constant the seed would leave the linearized regime after a time of order $\ln(1/\varepsilon)/\Lambda$. In gradient flow $\Lambda(t)$ decays as the loss falls. That alone does not make $\int \Lambda\,dt$ finite, but at $\bar\alpha = 8$ the growth stops at a finite gain, well after a constant rate fitted to larger biases predicts escape, so no $\ln(1/\varepsilon)$ law holds there (Section 5.2). The growth rate is also throttled while $\bar a$ develops, on a time scale of order $m/\bar\alpha$, so absolute timescales depend on m.

Case $k^* \ge 2$. When $\hat\sigma_1\hat\rho_1 = 0$, the router-mediated coupling vanishes at the symmetric saddle (Appendix H.2), so the router cannot drive the onset at linear order. If the expert diagonal entry A of the linearized system is negative, the router-mediated rate along the trajectory is $BC/|A| \propto (\Delta_v)^{2(k^*-1)}$, so a seed δ grows as $d\delta/dt \propto \delta^{2k^*-1}$ and a seed of size ε leaves the symmetric region after a time

$$t_{\mathrm{spec}} \;=\; \Theta\big(\varepsilon^{2(1-k^*)}\big) \tag{128}$$

in the small-overlap limit $\varepsilon \to 0$ at fixed $\bar\alpha,\ m,\ T$. If $A > 0$, as in our softmax Jacobians at and below $1.5\,T_c^{(0)}$ (Appendix H.2), the diagonal entry alone gives growth at rate A. Both cases are heuristics: they treat the coefficients as constant, while for $k^* = 1$ the measured rate decays as the loss falls. Neither is the single-index result of Ben Arous et al. (2021), whose population escape time scales as ε^{2-k} for $k \ge 3$. Below $T_c^{(0)}$ the router is frozen from initialization. Freezing means $T_{\mathrm{eff}} < T_c^{(0)}$, which with $T_{\mathrm{eff}} = T/\sqrt{1 - C_o^v}$ reads

$$1 - C_o^v(t,t) \;>\; T^2/T_c^{(0)\,2}, \tag{129}$$

and holds at $t = 0$, where $C_o^v \approx 0$, whenever $T < T_c^{(0)}$. In gradient flow C_o^v then settles near $-1/(E-1)$, which lowers T_{eff} by the factor $\sqrt{(E-1)/E}$ and deepens the freezing. A slower onset leaves the router frozen for longer before the experts' signal can lift it, which favors $k^* = 1$ for routing design.

Design observation. Equation (126) suggests reading the router-side Hermite spectrum $\{\hat{\rho}_k\}$ as a design parameter: routers with nonzero linear Hermite coefficient $\hat{\rho}_1$ belong to the $k^* = 1$ class. Among these, the quadratic-probit router has no normalized measure over experts and so no REM freezing (Appendix C.4). The soft-discriminant router has $k^* = 1$ and a router gradient $\partial_{v_e} r_e \propto x$ that does not depend on the router state. Softmax has $k^* = 1$ at every T but freezes below $T_c^{(0)}$. Hard top-k has an exactly solvable gate sector but is not differentiable: a top-1 gate routes every input to one expert from the start, which is frozen routing at zero temperature, and gives the router no per-example gradient, although the population gradient keeps the boundary contribution of Appendix C.6. Table 3 and Figure 10 catalog the four routers. A quantitative theory of Hermite-spectrum router design, including its interaction with auxiliary load-balancing losses, is left to future work.

Catalog of routers and their information-exponent properties. Table 3 summarizes the Hermite spectrum, joint information exponent k^*, frozen-collapse risk and gate-moment evaluation of the four routing functions, all under a teacher activation with $\hat{\sigma}_1 \neq 0$ (e.g. $\sigma(z) = \operatorname{erf}(z/\sqrt{2})$).

Routing function r	Hermite spectrum $\{\hat{\rho}_k\}$	k^*	FC risk	Gate moments
Soft-disc.: $a_0 + a_1 \operatorname{disc}_e(x)$	$\hat{\rho}_1 = a_1$, $\hat{\rho}_{k\geq 2} = 0$	1	none	closed-form Gaussian
Quadratic-probit: $a_0 + a_1 z + a_2(z^2-1)$	$\hat{\rho}_1, \hat{\rho}_2 \neq 0$, $\hat{\rho}_{k\geq 3} = 0$	1	none	closed-form Gaussian
Softmax-T: $\operatorname{smax}_T(V^\top x)_e$	all odd $\hat{\rho}_k \neq 0$, even $\hat{\rho}_k = 0$ at $E = 2$	1	below $T_c^{(0)}$	exact reduction (Appendix C.5)
Hard top-k: $\mathbf{1}\{e \in \operatorname{top}_k\}/k$	all odd $\hat{\rho}_k \neq 0$, even $\hat{\rho}_k = 0$ at $E = 2$	1	maximal	exact finite mixture

Table 3: Routing functions, Hermite-coefficient profiles, joint information exponent k^*, frozen-collapse (FC) risk, and gate-moment evaluation. All four achieve $k^* = 1$ when paired with a teacher whose first Hermite coefficient is nonzero. Soft-discriminant and quadratic-probit have terminating gate Hermite series, so their gate moments are closed-form Gaussian moments. Each hard top-k gate takes only the values 0 and $1/k$, so its gate average is an exact finite mixture, and its routing is frozen by construction, at zero temperature. For softmax the DMFT error grows as T falls (Section 4.4). All four use the second-cumulant closure of the effective action.

Figure 10 shows the corresponding normalized spectra $|\hat{\rho}_k|/k!$ for $k \in \{1, \dots, 5\}$, computed by Gauss–Hermite quadrature against the router preactivation. At $E = 2$ softmax, top-k and the soft-discriminant gate depend on the logits only through their difference and are odd in it, so their even-order coefficients vanish. Quadratic-probit acts on each raw logit and carries $\hat{\rho}_2 = 2a_2$. At $E > 2$ softmax and top-k also carry even-order coefficients. Soft-discriminant terminates exactly at $k = 1$ and quadratic-probit at $k = 2$, while softmax and hard top-k carry full odd-order tails.

H.2 Linearized dynamics around the symmetric saddle

Appendix J defines $\Lambda(t)$ as the dominant eigenvalue of a linearized matrix $M(t)$ acting on antisymmetric perturbations of the symmetric saddle, eq. (133). We give the construction here. Define the symmetric and antisymmetric expert-overlap modes

$$\bar{P}(t) \;=\; \tfrac{1}{E}\big(P_{\text{self}}(t) + (E-1)\,P_{\text{cross}}(t)\big), \qquad \delta P(t) \;=\; P_{\text{self}}(t) - P_{\text{cross}}(t) \;=\; \Delta(t),$$

and analogously $\bar{P}^v$ and $\delta P^v = \Delta_v$ for the router. Under the initialization (3), $\bar{P}(0) = \mathcal{O}(\varepsilon)$, $\bar{P}^v(0) = \mathcal{O}(\varepsilon^v)$, $\delta P(0) = \varepsilon$ and $\delta P^v(0) = \varepsilon^v$. The symmetric modes obey their own ODEs, which respect the permutation symmetry. The antisymmetric modes $(\delta P, \delta P^v)$ measure the breaking of that

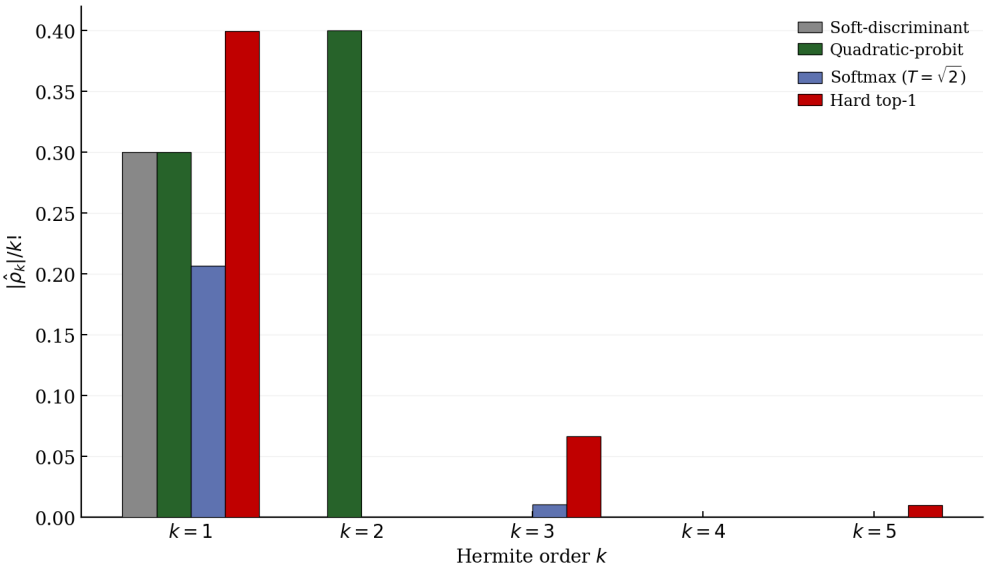


Figure 10: **Normalized Hermite spectra $|\hat{\rho}_k|/k!$ of the four canonical routers**, $k \in \{1, \dots, 5\}$. All four have $\hat{\rho}_1 \neq 0$ and therefore all achieve $k^* = 1$ when paired with an erf teacher (which has $\hat{\sigma}_1 \neq 0$). At $E = 2$, even-order coefficients vanish for the three routers that are odd about the discriminant origin. Quadratic-probit is the exception by construction, carrying $\hat{\rho}_2 = 2a_2$. Soft-discriminant and quadratic-probit have a finite spectrum (terminating at $k = 1$ and $k = 2$ respectively). Softmax and hard top-k carry an infinite odd-order tail that decays at different rates. Coefficients computed by 256-node Gauss–Hermite quadrature. Settings are $E = 2$, $a_1 = 0.30$, $a_2 = 0.40$ (quadratic-probit), $T = \sqrt{2}$ (softmax).

symmetry and obey

$$\frac{d}{dt}\begin{pmatrix}\delta P \\ \delta P^v\end{pmatrix} = \underbrace{\begin{pmatrix}A(t) & B(t) \\ C(t) & D(t)\end{pmatrix}}_{M(t)}\begin{pmatrix}\delta P \\ \delta P^v\end{pmatrix} + \mathcal{O}\big(\delta P^2,\, \delta P\, \delta P^v,\, (\delta P^v)^2\big). \tag{130}$$

The entries of $M(t)$ are obtained by differentiating the expert and router signal kernels of Appendix D with respect to the antisymmetric overlaps and evaluating at the symmetric saddle $\delta P = \delta P^v = 0$. Each entry is a Gaussian expectation involving σ', r_e and their derivatives, integrated against a covariance determined by the symmetric-mode order parameters $\bar{P}(t), \bar{P}^v(t)$ and the equal-time correlators C_d, C_{dv} of Appendix B.1. Substituting the Hermite expansions (125) and applying Mehler's formula gives, schematically,

$$A(t) = \nu_w(t) + \sum_{k\geq 1} a_k(t)\, \hat{\sigma}_k^2 + \mathcal{O}(\delta P, \delta P^v), \quad D(t) = -\nu_v(t) + \sum_{k\geq 1} d_k(t)\, \hat{\rho}_k^2 + \mathcal{O}(\delta P, \delta P^v), \tag{131}$$

$$B(t) = \sum_{k\geq 1} b_k(t)\, \hat{\sigma}_k \hat{\rho}_k\, (\delta P^v)^{k-1} + \cdots, \qquad C(t) = \sum_{k\geq 1} c_k(t)\, \hat{\sigma}_k \hat{\rho}_k\, (\delta P^v)^{k-1} + \cdots, \tag{132}$$

where a_k, b_k, c_k, d_k are coefficient functions of $\bar{\alpha}, \bar{a}(t), \chi_R(t)$ and the equal-time correlators, and the diagonal entries A and D contain the spherical-constraint drifts. The expert row (A, B) carries the explicit factor $1/m$. At the symmetric saddle, expert equivalence and the shift invariance of the gate remove the leading contribution to the router diagonal D. In the DMFT Jacobians we evaluate for the softmax router at T from 0.7 to $2\,T_c^{(0)}$, the product $|AD|$ lies 6.0 to 8.0 orders of magnitude below BC at $E = 2$, and 6.2 to 9.5 orders at $E = 4$ and 8. Hence $\det M = AD - BC < 0$ and M has a positive eigenvalue: the symmetric solution is a *saddle*, not a stable node. The dominant eigenvalue $\lambda_{\max}(M(t))$ is the linearized growth rate $\Lambda(t)$ of (133).

Leading order in k^*. The dominant eigenvalue is $\lambda_{\max} = \frac{A+D}{2} + \sqrt{\left(\frac{A-D}{2}\right)^2 + BC}$. For $k^* = 1$ the router diagonal D nearly vanishes at the symmetric saddle, and A^2 is at most 4% of BC in the Jacobians we evaluate, so $\Lambda \simeq A/2 + \sqrt{BC}$ with $\sqrt{BC} \propto |\hat{\sigma}_1 \hat{\rho}_1|$ dominant. Because B carries the $1/m$ of the expert row and C does not, $\sqrt{BC} \propto m^{-1/2}$, the factor in (127). The sign of A depends on the temperature: in these Jacobians $A < 0$ at $2\,T_c^{(0)}$ and $A > 0$ at and below $1.5\,T_c^{(0)}$. For $k^* \geq 2$ the terms of order $k \geq 2$ in (132) vanish at the saddle, so $B = C = 0$ there and $\lambda_{\max}$ reduces to the larger diagonal entry: the router cannot drive the onset at linear order. Along the trajectory B and C scale as $(\delta P^v)^{k^*-1}$, which for $A < 0$ gives the rate $BC/|A|$ behind (128). Once δP and δP^v have grown appreciably the diagonal entries are no longer negligible, so the k^* ordering governs the *rate of departure from symmetry*, not the growth rate at every point of the trajectory.

H.3 Frozen collapse and the onset

Frozen collapse is set by the bath temperature. When $T < T_c^{(0)}$ the freezing condition (129) holds already at initialization, where $C_o^v(t,t) \approx 0$, so the router is frozen from the start: in gradient flow at $E = 8$ and $\bar{\alpha} = 4$, Y_2 is 0.75 to 0.86 at $t = 0$ for T between 0.2 and $0.37\,T_c^{(0)}$ (Appendix J). The drift of C_o^v toward $-1/(E-1)$ lowers T_{eff} and deepens the freezing. Within the linearized regime, $\delta P^v(t) = \mathcal{O}(\varepsilon^v)$ lies far below the minimum signal Δ_v^* of (12), so the condensation branch (11) does not yet exist.

A frozen gate makes each input's router gradient small. For one input the gradient on v_e is proportional to $g_e(F_e - f)$ with $f = \sum_j g_j F_j$ (40). As the gate concentrates on a winner e^*, $f \to F_{e^*}$, so the winner loses its edge over f while every loser loses its weight. The population gradient does not vanish, however: inputs near a decision boundary between two experts contribute at order one, because the softmax derivative carries the factor $1/T$ there. In gradient flow at $E = 8$ below $T_c^{(0)}$, the router alignment Δ_v grows over the first few hundred time units and then declines slightly, and Y_2 on the training inputs peaks early and then falls slowly, without the router leaving the frozen state by $t = 3200$ (Appendix J). The per-input suppression is specific to normalized gates and is absent for the soft-discriminant gate, consistent with frozen collapse being a softmax phenomenon. Frozen collapse therefore occurs when $T < T_c^{(0)}$ and the experts' signal stays too small to lift the router out of the frozen state. For $k^* = 1$ the onset is limited by the amplitude ramp, of order $m/\bar{\alpha}$, and by the decay of Λ. For $k^* \geq 2$ the heuristic escape time (128) grows as the seed shrinks.

H.4 Future directions

The above observations suggest a richer program of Hermite-spectrum router design that we leave to future work. These include: (i) an explicit computation of the prefactor $\gamma(t)$ and of the diagonal entries of $M(t)$, which would turn (128) into a quantitative prediction, (ii) a phase-conditional theory of auxiliary load-balancing-loss interventions, predicting where in the $(T, \bar{\alpha}, \varepsilon)$ control space the penalty stabilizes versus opposes specialization, and (iii) a test of the heuristic scaling (128) for $k^* \geq 2$ at $\varepsilon \ll 10^{-2}$ and $T_{\max} \gg 10^4$, beyond our compute budget, with a gate whose higher Hermite modes stay orthogonal to the linear one during training. The clipped cubic gate of Appendix G does not provide such a test.

I ADDITIONAL VALIDATION DETAIL

This appendix expands the validation of Section 4.4. It records the full E-dependence of the error of the linearized-gate DMFT and separates most of the gate linearization from the second-cumulant closure, two approximations that Figure 11 conflates.

Soft-discriminant router at $E = 4$. With the implementation of Appendix D.8 and the settings of Figure 3 at $m = 100$, except $d = 3000$ ($a_1 = 0.3$, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, $T_{\max} = 50$, 8 gradient-flow seeds), the DMFT tracks Δ to within 5.9% of its gradient-flow change and P_{self} to within 3.0%, against 6.7% and 2.4% at $E = 2$ and the same width. P_{cross} is again the least accurate, at 25% of its peak (22% at $E = 2$): the DMFT peak is 0.84 times the gradient-flow peak and comes 7 time units early. At $t = 50$, $m(\Delta - \varepsilon)$ is 2.3 in the DMFT and 2.0 in gradient flow.

Where does the cumulant truncation fail? Sections 4.1–4.3 develop the DMFT for a generic router with the second-cumulant closure. Its gate-level input terminates for the soft-discriminant router and is evaluated by the reduction of Appendix C.5 for softmax. Here we relate the DMFT's accuracy to the freezing scale of Section 5.

Empirical breakdown across E. We run the general-E Volterra solver against multi-seed softmax gradient flow at $E \in \{2, 4, 8\}$ across nine temperatures, with the implementation that predates the corrections of Appendix D.8 at $E = 4$ and 8. Time-RMSE on Δ_v tracks the residual. Figure 11 shows two views. *Absolute axes (left)*: RMSE shrinks with E at fixed T. Both the matched $a_1 \propto \sqrt{E-1}/E^{3/2}$ and the initial biases ($\varepsilon^v = 0.05$ at $E = 2$, 0.0125 at $E = 4, 8$) are smaller at larger E, and both reduce the scale of Δ_v. *Rescaled axes (right)*: plotted against $T_c^{(0)}(E)/T$, the residual rises smoothly through $T_c^{(0)}$ at every E, with no feature at the freezing scale, because the linearization error dominates it (next paragraph).

Relative to how much each curve changes during training, the errors on the expert order parameter Δ are comparable to those on Δ_v. At the four temperatures whose trajectories we store, the RMSE divided by the gradient-flow change is 0.21 to 0.79 for Δ and 0.14 to 1.03 for Δ_v at $E = 2$, and 0.05 to 0.23 and 0.04 to 0.76 at $E = 8$. At $E = 2$ and $T \leq 0.85$ the DMFT's rise in Δ is 2.2 to 2.6 times the gradient-flow rise. The matched coefficient is not an ansatz: expanding softmax at small z_e/T gives $g_e \simeq 1/E + (z_e - \bar{z})/(ET)$, so the linear-discriminant router is the leading term of softmax's high-temperature expansion at exactly this a_1, and the gate moments admit an exact reduction at *any* temperature (Appendix C.5). With the gate means and drives treated exactly, the remaining error at $0.7\,T_c^{(0)}$ is 4.9% of the gradient-flow peak at $E = 2$ (next paragraph). A thermodynamic check on bare-softmax statistics is in Appendix F.6.

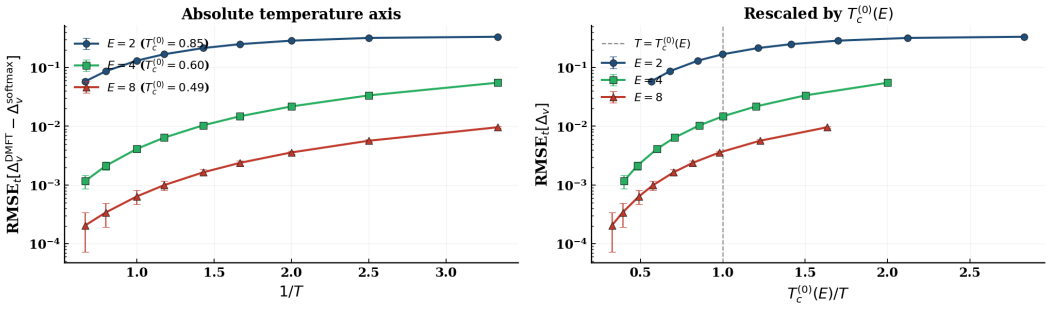


Figure 11: **Linearized-gate DMFT against softmax gradient flow, $E \in \{2, 4, 8\}$.** Linear-DMFT at matched a_1 vs. softmax-GF (4 seeds, $T_{\max} = 30$, $\bar{\alpha} = \kappa = 1$, with $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, $d = 3000$ and $m = 200$ at $E = 2$, and $\varepsilon = 0.05$, $\varepsilon^v = 0.0125$, $d = 2000$ and $m = 120$ at $E = 4, 8$). *Left*: time-RMSE on Δ_v vs. $1/T$, curves vertically offset because the matched a_1 and the initial biases are smaller at larger E. The residual rises smoothly as T falls, with no feature at $T_c^{(0)}$. *Right*: rescaled by $T_c^{(0)}(E)$. The $E = 4$ and 8 curves predate the corrections of Appendix D.8.

Separating the gate approximation from the closure. Figure 11 measures the *linearized* gate against softmax gradient flow, so its residual mixes two distinct approximations: the κ_2 closure and the high-temperature linearization of the gate itself. The exact reduction of Appendix C.5, exact for $E \leq 3$, separates most of the linearization from the closure. It lets the same closure be integrated with the softmax gate means and second moments, and with the expert and router drives evaluated for the softmax gate. Figure 12 does this. The other router-side terms listed in Appendix D.8, among them the population part of ν_v, Σ_R, Σ_C, the router memory kernels and the self-reaction, keep their linear-gate forms at every E. At $E = 2$ the softmax router force on an input is $4g_0g_1 \leq 1$ times the linear one, so these terms overstate it increasingly as T falls, and the remaining residual combines the closure with this router-side linearization. With these gate moments and drives, the residual on Δ_v at $2.5\,T_c^{(0)}$ is 0.8% of the gradient-flow peak at $E = 2$ and 0.1% at $E = 4$ and 8, and it grows as T falls, to 4.9%, 8.6% and 11.8% at $0.70\,T_c^{(0)}$. There the residual with the fully linearized gate is larger by $10.9\times$, $7.0\times$ and $5.1\times$ at $E = 2, 4, 8$, so the linearization dominates any residual on Δ_v measured with it. On the expert order parameter Δ the exact gate and drives remove only about half of the error, the linearized residual being 1.9 to 2.4 times the exact one at $E = 2$ across these temperatures. These runs start from $\varepsilon = 0.2$, $\varepsilon^v = 0.05$ and $\bar{a}(0) = 0.5$ and end at $t = 20$ (30 seeds, $m = 120$, $d = 2000$), so they are not the runs of Figure 11. The values at $E = 4$ and 8 carry the further approximations of Appendix D.8: the expert-drive quadrature evaluates the unresolved experts at one resolved field (Appendix C.6), and several kernels weight the mismatched clusters by one instead of $E - 1$. We have not quantified their effect, so we do not compare these residuals across E.

Equation (10) is a thermodynamic-limit statement, and at finite E the transition is a crossover of relative width $\mathcal{O}(1/\sqrt{\ln E})$, already $\mathcal{O}(1)$ at $E = 2$, so no sharp feature is expected in these residuals at these E, and none is seen.

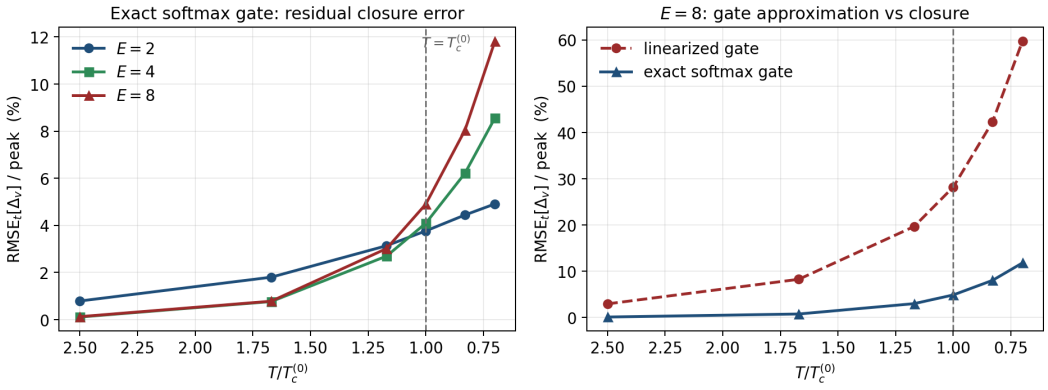


Figure 12: **With the softmax gate moments and drives of Appendix C.5, the remaining error grows as T falls.** Time-RMSE on Δ_v against softmax gradient flow, normalized by the GF peak so the vertical axis is a fractional error ($\bar{\alpha} = \kappa = 1$, $m = 120$, $d = 2000$, 30 seeds, $\varepsilon = 0.2$, $\varepsilon^v = 0.05$, $\bar{a}(0) = 0.5$, $T_{\max} = 20$, with T decreasing to the right). *Left*: the same closure integrated with the softmax gate moments and drives, the other router-side terms keeping their linear-gate form (Appendix D.8). The residual is below 1% far above $T_c^{(0)}$ and grows as T falls, to 4.9% at $E = 2$ at $0.70\,T_c^{(0)}$. The $E = 4$ and 8 curves predate the corrections of Appendix D.8. *Right*: at $E = 8$, the same quantity with the linearized gate (dashed) and the exact gate (solid). The linearized residual is 5.1 times the exact one at $0.70\,T_c^{(0)}$.

A thermodynamic cross-check. Independently of the trained dynamics, bare-softmax statistics on i.i.d. Gaussian logits, with no learning at all, show the same freezing scale (Figure 2, Appendix F.6).

J PHASE-BOUNDARY ANALYSIS: DETERMINANT DETAIL

This appendix derives the marginality condition (135), evaluates it on the symmetric solution, measures the gain of the onset, and reports how the two measured contours depend on training budget and $\bar{\alpha}$.

Linearization and the characteristic equation. Linearizing the two-scalar DMFT around $\Delta = 0$ with perturbation $\delta P = P_{\text{self}} - P_{\text{cross}} = \varepsilon$, and projecting onto the dominant eigenvector of the linearized system, gives

$$\partial_t(\delta P) \;=\; \Lambda(t)\,\delta P \;+\; \mathcal{O}(\delta P^2), \tag{133}$$

where $\Lambda(t)$ is the dominant eigenvalue of the linearized 2×2 matrix $M(t)$ acting on $(\delta P, \delta P^v)$ (Appendix H). It depends on the routing differential $\partial_{P^v}(\bar{g}_{\text{self}} - \bar{g}_{\text{cross}})$, the current signal $\Delta_v(t)$ and the cluster separability $\Omega^{UU}_{cc'}$. Appendix H relates its router-mediated part to the joint Hermite spectrum of teacher and gate, as a heuristic. Linear stability of a DMFT fixed point goes back to Sompolinsky et al. (1988). The marginality condition is obtained from the linearized dynamics, which are *not* instantaneous: the memory kernels make the antisymmetric sector a linear Volterra system,

$$\dot{u}(t) \;=\; M_0\,u(t) \;-\; \int_0^t K(t-s)\,u(s)\,ds, \qquad u := (\delta P,\ \delta P^v)^\top, \tag{134}$$

on a background treated as time-translation invariant near threshold, with integrable kernels. This is an approximation, since $\Lambda(t)$ decays as the loss falls. Substituting $u \propto e^{\lambda t}$ turns the convolution into multiplication by the Laplace transform $\widehat{K}(\lambda) = \int_0^\infty e^{-\lambda s} K(s)\,ds$, giving the characteristic equation $\det[\lambda I - M_0 + \widehat{K}(\lambda)] = 0$. The marginality condition is the surface on which the marginal, non-oscillatory root $\lambda = 0$ is attained:

$$\det\bigl[-M_0 + \widehat{K}(0)\bigr] \;=\; 0, \qquad \widehat{K}(0) = \int_0^\infty K(s)\,ds. \tag{135}$$

A 2×2 determinant therefore survives the memory, but its entries are renormalized by $\widehat{K}(0)$ rather than being those of the instantaneous matrix $M(t)$. The two rows carry different prefactors (the expert row an explicit $1/m$, the router row none), so the memory shifts the router-row entries at order one and the expert-row entries only weakly. Oscillatory instabilities would require a root with $\operatorname{Im}\lambda \neq 0$ and cannot be excluded without analyzing the full frequency-dependent characteristic equation. Evaluated on the symmetric solution, however, the marginal case is not attained at accessible parameters. Expert equivalence together with softmax shift invariance removes the leading part of the router-row diagonal entry there, so, in the notation of Appendix H.2, $|AD|$ lies at least six orders of magnitude below BC at every (T, E) we evaluate ($E = 2, 4, 8$, T from 0.7 to $2\,T_c^{(0)}$), and $\det M_0 < 0$. Because the whole expert row carries the explicit $1/m$, the determinant factorizes into m^{-1} times an m-independent quantity, so m rescales its magnitude without touching its sign. The sign is also the same at every T and E we evaluate, all at $\bar{\alpha} = 1$. The full linearized DMFT, memory terms included, gives a positive growth rate at every (T, E) we evaluate, so the memory does not reverse the sign either. The symmetric solution is therefore a saddle throughout the accessible range, and (135) serves as the correct marginality *condition* rather than as a boundary the dynamics crosses. Instead δP grows at every temperature (Section 5.2). The frozen-collapse regime is bounded separately by $T_{\text{eff}} < T_c^{(0)}$: below $T_c^{(0)}$ the router is frozen from initialization, before the amplification has lifted the seed. To our knowledge the frozen-collapse regime is new to MoE training dynamics. For single-pathway two-layer networks, Veiga et al. (2022) give a phase diagram of SGD.

Gain of the onset. At $E = 4$, $T = 1$ ($T_c^{(0)} = 0.60$), $d = 2000$ and $m = 120$ we plant $\varepsilon \in \{0.1, 0.03, 0.01, 0.003, 0.001\}$ (down to 0.003 at $\bar{\alpha} = 8$) with router bias $\varepsilon/4$, all other overlaps exactly zero, and record when Δ first reaches 0.2 (three seeds, $t \leq 800$). At $\bar{\alpha} = 8$ the escape times are 57 and 172 for $\varepsilon = 0.1$ and 0.03, and no smaller bias escapes. At $\bar{\alpha} = 4$ they are 110 and 551, and at $\bar{\alpha} = 1$ only $\varepsilon = 0.1$ escapes, at $t \approx 470$. For $\varepsilon = 0.01$ and 0.03, $\Delta(800)/\varepsilon$ is about 3 at $\bar{\alpha} = 1$, still growing, and about 8 at $\bar{\alpha} = 4$. At $\bar{\alpha} = 8$ it is 14 for $\varepsilon = 0.01$ and 16 for $\varepsilon = 0.003$ (per seed, 12.5 to 15.7 and 8.5 to 21.8). At $\bar{\alpha} = 1$ and $\varepsilon \leq 0.003$ individual seeds change sign, because the bias is then comparable to the finite-d fluctuations of the initial overlaps. The growth rate falls

steadily. At $\bar{\alpha} = 8$ a constant rate fitted to the escape times of $\varepsilon = 0.1$ and 0.03 predicts that $\varepsilon = 0.01$ escapes near $t = 280$. Continued to $t = 2400$ (one seed), that run instead peaks at $\Delta = 0.169$ near $t = 1600$ and then declines slightly, a gain of 17. At $\bar{\alpha} = 1$ and 4 the runs end at $t = 800$ while still growing (their growth rates over the last 100 time units are 5.0×10^{-4} and 2.7 to 3.0×10^{-4}), so they do not show whether the gain is finite. At $t = 800$ the gain is larger at larger $\bar{\alpha}$, but at equal $\bar{\alpha}t$ it grows much less: at $\bar{\alpha}t = 800$ it is 3.2 at $\bar{\alpha} = 1$, 4.0 at $\bar{\alpha} = 4$ and 4.3 at $\bar{\alpha} = 8$. Continued to $t = 2400$ at $\bar{\alpha} = 8$ (one seed each), the runs at $\varepsilon = 0.03$ and 0.1 peak at $\Delta = 0.386$ and 0.613 near $t \approx 1170$ and then decline slightly, like the run at $\varepsilon = 0.01$. They approach no common specialized state: the plateau grows continuously with ε, with gains of 13 and 6, so larger biases are amplified less than proportionally, and the escape level 0.2 is a convention rather than a sharp threshold.

Frozen collapse at long times. Below $T_c^{(0)}$ the router freezes first, and the experts specialize less the lower the temperature. At $E = 8$ and $\bar{\alpha} = 4$ we follow three cells of the $\kappa = 1$ grid, at $T = 0.10$, 0.135 and 0.183 with $\varepsilon = 0.2$ and $\varepsilon^v = 0.05$, to $t = 3200$ (one seed each, Y_2 on the training inputs). The growth of Δ saturates at 0.150, 0.191 and 0.230 at $T = 0.20$, 0.28 and $0.37\,T_c^{(0)}$. At the coldest cell it peaks near $t \approx 1450$ and then loses 2.4%, at $0.28\,T_c^{(0)}$ it loses 0.4%, and at $0.37\,T_c^{(0)}$ it is flat after $t \approx 2400$. On the training inputs the router is concentrated from the start ($Y_2 = 0.86$, 0.81 and 0.75 at $t = 0$), concentrates further until $t \approx 160$ to 270 ($Y_2 = 0.98$, 0.97 and 0.95), and then relaxes slightly, to $Y_2 = 0.92$, 0.89 and 0.86 at $t = 3200$. We did not record routing on fresh inputs in these runs. Its alignment Δ_v rises from 0.05 to 0.25, 0.31 and 0.39 by $t \approx 280$ to 525 and then declines slightly, to 0.24, 0.30 and 0.36.

Budget and signal dependence of the two contours. We sweep the $E = 8$ grid over a sixteenfold range of training budget at the two ends of the thirteenfold $\bar{\alpha}$ range, taking the onset contour as the temperature of largest growth and the concentration contour as the $Y_2 = 0.5$ crossing on the training inputs. At $\bar{\alpha} = 0.30$ the onset contour moves from 0.68 to $1.68\,T_c^{(0)}$ and the concentration contour from 1.31 to $2.82\,T_c^{(0)}$ by $t = 400$, leaving the grid by $t = 800$, so both are still moving. At $\bar{\alpha} = 4$ the onset contour stops at $0.92\,T_c^{(0)}$ after $t = 200$ and the concentration contour converges to $1.53\,T_c^{(0)}$, to about 3% over the final doubling. The concentration contour moves strongly with $\bar{\alpha}$, and the two contours stay distinct at both ends of the range. The concentration contour is not the REM line (11), which has no real root at $\kappa = 1$. Y_2 is measured on the training inputs here, and the contour sits highest at the smallest $\bar{\alpha}$, consistent with the training-set self-reaction of Section 4.3, which is largest there.

K Closure and Onsager Reduction: Supporting Detail

This appendix gives the two arguments that Section 4 states in summary form: why the gate-level and effective-action expansions must be distinguished, and why the mean routing shift δ_{disc} collapses to a single scalar.

The two expansions, in full. The distinction is between two different objects. The *gate-level* Hermite expansion of r terminates at first order for the soft-discriminant gate and at second order for quadratic-probit, so the soft-discriminant gate is Gaussian and the quadratic-probit gate adds only a finite polynomial, evaluated exactly. Each hard top-k gate takes only the values 0 and $1/k$, which reduces the gate average to an exact finite mixture. The *effective-action* cumulant series in (33), by contrast, does *not* terminate for any of these gates: the map $\Sigma \mapsto \operatorname{tr}\log(I + \Sigma)$ is non-linear in the router field, so $\kappa_{\ell \geq 3} \neq 0$ generically even when the gate is affine. We therefore treat the routing-induced part of the DMFT as a *second-cumulant closure* throughout, not as an exact expansion. For softmax the variance of the gate fluctuations shrinks as $1/T^2$, so the closure is expected to improve at high temperature. In the glassy regime the neglected structure is not small, and untruncated sampling (Appendix F.5) would be needed. The results reported here do not use it. Validity conditions, the per-router catalog (Table 2), and the quenched-vs-annealed distinction are in Appendix C.4 and Appendix F.

The single-scalar reduction, in full. At any E, the E-dimensional vector of per-expert logit corrections collapses to a single number, for two reasons. (This concerns the mean shift δ_{disc}. The

correction to the expert drive is the per-example self-reaction of Appendix E.5, a joint average over each example's history that does not reduce to a scalar.) First, within a cluster, permutation symmetry among the $E-1$ mismatched experts forces the vector of per-expert logit corrections into the form $(\delta_{\mathrm{match}}, \delta_{\mathrm{mis}}, \ldots, \delta_{\mathrm{mis}})$, and because the soft-discriminant gate, like softmax and top-k, depends on the logits only through their *differences* (shift invariance), the common-logit mode is unobservable, so $\sum_e \delta_e = 0$ and hence $\delta_{\mathrm{match}} = -(E-1)\delta_{\mathrm{mis}}$, one free number. Second, by the diagonal-S_E symmetry of the symmetric initialization (3), the per-cluster corrections $\delta^{(c)}_{\mathrm{disc}}(t)$ for $c = 1, \ldots, E$ are all related by permutation (we track $\delta_{\mathrm{disc}}(t) := \delta^{(c=1)}_{\mathrm{disc}}(t)$ and recover the others by symmetry). In the matched-expert convention of (103) they are all equal.