TWO ROUTES TO CLASS ORGANIZATION IN A SOLVABLE MODEL OF CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning can yield class-informative representations from unlabeled data, but how this class organization emerges as data scale grows remains unclear. We study this question in a high-dimensional InfoNCE model with data drawn from a latent Gaussian mixture and a single-layer encoder. In the proportional limit, a replica calculation under a reduced replica-symmetric (RS) and class-permutation-symmetric ansatz yields a free-energy functional and saddle-point equations for a small set of macroscopic order parameters. We quantify class organization by the fraction of representation variance explained by between-class mean differences. As sample complexity increases, we find two routes to this organization: a specialized branch either emerges continuously from the unspecialized branch or appears at finite amplitude and coexists with a locally stable unspecialized branch before becoming preferred through an RS free-energy crossing. The same nonlinear activation can exhibit either route as the number of classes changes, while the number of negatives shifts the transition location. A local Landau-type expansion of the profiled free energy identifies a cubic term associated with finite-amplitude competition. Its leading high-InfoNCE-temperature form reveals contributions from activation curvature, class structure, and contrastive weighting. Finite-dimensional simulations quantitatively reproduce the predicted representation statistics and qualitatively recover the branch structures.

1 INTRODUCTION

Self-supervised contrastive learning has been shown to yield class-informative representations even when class labels are absent from the training objective (Chen et al., 2020; Parulekar et al., 2023). However, the absence of representation collapse does not by itself guarantee class organization (Jing et al., 2022). Representations may retain substantial variance yet exhibit little separation between class means (Xue et al., 2023). Existing theory has characterized the conditions under which contrastive objectives recover class-informative representations, cluster structure, or discriminative subspaces (Saunshi et al., 2019; Zhang et al., 2025; Bansal et al., 2025). What remains unclear is whether class organization emerges continuously from an unspecialized state or through competition between distinct states as data scale increases.

Empirical studies have shown that contrastive representations vary with data scale, augmentation, negative sampling, temperature, and architecture (Cole et al., 2022; Chen et al., 2020; Wang & Liu, 2021; Chuang et al., 2020; Xue et al., 2024). These factors substantially affect class separability and downstream performance even within the same family of contrastive objectives. Much less is known about whether they also shape how class organization emerges, rather than how strongly it develops. Answering this question requires tracing branches of candidate macroscopic states as sample complexity increases and determining whether distinct states coexist at fixed sample complexity.

Training trajectories identify the states reached by a given optimizer and initialization, but not the full set of stationary macroscopic states or their competition. We therefore take an equilibrium perspective that distinguishes state existence, local stability, and thermodynamic preference as sample complexity varies. In this work, we draw inspiration from statistical physics, where high-dimensional learning problems are often described in terms of a small set of macroscopic order parameters (Seung et al., 1992; Watkin et al., 1993; Engel & Van den Broeck, 2001). Statistical-physics approaches have been used to study a broad range of learning transitions, from early analyses of generalization

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.

transitions in neural networks (Sompolinsky et al., 1990; Györgyi, 1990; Schwarze, 1993) to high-dimensional phase transitions and computational-statistical gaps (Aubin et al., 2018; Barbier et al., 2019; Maillard et al., 2020), and more recently, competing positional and semantic learning mechanisms in attention (Cui et al., 2024).

Building on tractable models of contrastive learning (Meng & Wang, 2024; Bansal et al., 2025), we study a high-dimensional InfoNCE model with data drawn from a latent Gaussian mixture and a single-layer encoder. We work in the proportional limit, where the input dimension and the number of training groups grow at a fixed ratio. A replica calculation under a reduced ansatz with replica symmetry (RS) and class-permutation symmetry yields saddle-point equations for a small set of macroscopic order parameters that describe candidate states. We quantify class organization by the fraction of representation variance explained by between-class mean differences. Within this ansatz, class specialization denotes the preferential alignment of individual encoder directions with particular latent classes. The corresponding symmetry-breaking order parameter tracks this specialization and, in turn, the emergence of class organization.

Analysis of the saddle-point branches reveals two routes to class organization through encoder specialization. For the linear encoder, the specialized branch emerges continuously from the unspecialized branch as sample complexity increases. With all other model parameters fixed, the smooth-ReLU model instead exhibits a finite-amplitude specialized branch that coexists with a locally stable unspecialized branch. This branch appears before it becomes thermodynamically preferred, so branch emergence precedes the free-energy crossing. In other configurations, however, the same smooth-ReLU activation exhibits continuous emergence, showing that the activation alone does not determine the route.

To understand how the emergence route varies with model configuration, we expand the profiled RS free energy in the specialization amplitude about the unspecialized branch. The cubic term captures the local asymmetry associated with finite-amplitude competition in the configurations studied. At high InfoNCE temperature, the leading-order expansion gives an explicit expression for the cubic coefficient, with contributions from activation curvature, class structure, and contrastive weighting. Branch continuation establishes coexistence, while RS free-energy comparisons determine thermodynamic preference. Finite-dimensional simulations quantitatively reproduce the predicted representation statistics and qualitatively recover the branch structures.

In particular our contributions are:

- We develop a replica characterization of a high-dimensional InfoNCE model that links encoder specialization to class organization through the class-explained variance fraction.
- We identify continuous emergence and finite-amplitude competition as two distinct routes to class organization, and show that the route varies with the activation and other model parameters.
- We characterize the profiled RS free-energy landscape around the unspecialized branch and derive the leading high-InfoNCE-temperature form of the cubic coefficient, revealing contributions from activation curvature, class structure, and contrastive weighting.

2 SETUP

Input data model. We consider a Gaussian mixture model in $\mathbb{R}^N$ with $C \geq 2$ equally likely latent classes. The class directions $\boldsymbol{\xi}_c$, $c \in \{1, \dots, C\}$, are mutually orthogonal and form the rows of $\boldsymbol{\Xi} \in \mathbb{R}^{C \times N}$, with normalization $N^{-1}\boldsymbol{\Xi}\boldsymbol{\Xi}^\top = \boldsymbol{I}_C$. Conditional on class c, an input is generated as $\boldsymbol{x} = \sqrt{\gamma}\,\boldsymbol{\xi}_c + \sqrt{1-\gamma}\,\boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N)$ is class-independent Gaussian noise, $\gamma \in [0, 1)$ controls the signal strength, and $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. We use the weak-signal scaling $\kappa = N\gamma = \Theta_N(1)$, under which the signal and noise contributions to the encoder preactivations remain comparable. This construction builds on latent-class abstractions of contrastive learning (Saunshi et al., 2019) and Gaussian-mixture analyses of InfoNCE (Bansal et al., 2025).

The training set $\mathcal{D} = \{(\boldsymbol{x}^p, \boldsymbol{x}^{+p}, \boldsymbol{x}_1^{-p}, \dots, \boldsymbol{x}_K^{-p})\}_{p=1}^P$ comprises P independent groups, each containing an anchor, a positive, and K negatives. Within each group, the anchor class is sampled uniformly from the C classes, the positive shares the anchor class, and the K negative classes are sampled independently and uniformly with replacement from the other $C-1$ classes. Conditional

on these class labels, all inputs are drawn independently. For clarity, we restrict the main discussion to this simple instance, which already exhibits the emergence routes studied in Section 4. Unconditioned negative sampling, which permits same-class negatives, is treated in Appendix A.8.

Encoder and training objective. We use a single-layer encoder with preactivations $\boldsymbol{h} = \boldsymbol{W}\boldsymbol{x}/\sqrt{N}$ and representations $\boldsymbol{v} = \phi(\boldsymbol{h})$, where $\boldsymbol{W} \in \mathbb{R}^{M\times N}$, M is the representation dimension, and $\phi : \mathbb{R} \to \mathbb{R}$ acts coordinatewise. We specialize to the symmetry-matched setting $M = C$, which isolates the specialization of representation directions toward latent classes. The activation is either linear, $\phi(t) = t$, or smooth ReLU, $\phi(t) = \frac{1}{2}(t + \sqrt{t^2 + \varepsilon_{\rm sm}})$, with $\varepsilon_{\rm sm} = 0.01$. The encoder is trained by minimizing the regularized InfoNCE objective (van den Oord et al., 2018), summed over the P groups and given by

$$\mathcal{L}(\boldsymbol{W}) = -\sum_{p=1}^{P} \log \frac{\exp\big((\boldsymbol{v}^p)^\top \boldsymbol{v}^{+p}/\tau\big)}{\exp((\boldsymbol{v}^p)^\top \boldsymbol{v}^{+p}/\tau) + \sum_{k=1}^{K} \exp\big((\boldsymbol{v}^p)^\top \boldsymbol{v}_k^{-p}/\tau\big)} + \frac{\lambda_w}{2}\|\boldsymbol{W}\|_F^2, \qquad (1)$$

where $\boldsymbol{v}^p$, $\boldsymbol{v}^{+p}$, and $\boldsymbol{v}_k^{-p}$ are the anchor, positive, and negative representations produced by the same encoder. Here, τ is the InfoNCE temperature, λ_w is the weight-regularization strength, and $\|\cdot\|_F$ denotes the Frobenius norm.

Class organization. We quantify class organization by the fraction of representation variance explained by between-class mean differences. For a fixed encoder, let $\boldsymbol{m}_c = \mathbb{E}[\boldsymbol{v} \mid c]$ denote the class-conditional mean, $\bar{\boldsymbol{m}} = C^{-1}\sum_{c=1}^{C} \boldsymbol{m}_c$ the overall mean, and $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{v}]$ the population covariance, with all moments taken over the input distribution. The covariance then decomposes as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$, where $\boldsymbol{\Sigma}_B = C^{-1}\sum_{c=1}^{C}(\boldsymbol{m}_c - \bar{\boldsymbol{m}})(\boldsymbol{m}_c - \bar{\boldsymbol{m}})^\top$ is the between-class covariance and $\boldsymbol{\Sigma}_W$ is the average within-class covariance. For $\mathrm{Tr}(\boldsymbol{\Sigma}) > 0$, the class-explained variance fraction is $\eta_{\rm cls} = \mathrm{Tr}(\boldsymbol{\Sigma}_B)/\mathrm{Tr}(\boldsymbol{\Sigma}) \in [0, 1]$, where Tr denotes the trace.

Class specialization. Let $\boldsymbol{w}_i^\top$ denote the ith row of $\boldsymbol{W}$. The overlap matrix $\boldsymbol{R} = N^{-1}\boldsymbol{W}\boldsymbol{\Xi}^\top$ measures the alignment between encoder coordinates and class directions, with entries $R_{ic} = N^{-1}\boldsymbol{w}_i^\top \boldsymbol{\xi}_c$. Following analyses of specialization in symmetry-matched teacher–student models (Biehl et al., 1998; Oostwal et al., 2021), the overlap matrix is parametrized as $R_{ic} = r_o + (r_d - r_o)\delta_{ic}$ under permutation symmetry, where r_d and r_o are the matched and unmatched overlaps. Their difference, $\Delta r = r_d - r_o$, measures coordinate-dependent class specialization, whereas $\eta_{\rm cls}$ is invariant under orthogonal changes of representation coordinates. For small specialization, $\eta_{\rm cls} = O((\Delta r)^2)$ in our model, directly relating Δr to class organization. For comparison, the centered effective-rank deficit appears only at $O((\Delta r)^4)$; see Appendix B.

3 Replica Theory of InfoNCE

We analyze the model of Section 2 in the proportional limit $N, P \to \infty$ at fixed sample complexity $\alpha = P/N$, while C, M, K, κ, τ, and λ_w remain $\Theta_N(1)$ and $\gamma = \kappa/N$. Rather than tracking individual weights, we characterize encoder states by a set of macroscopic summary statistics, or order parameters, that describe alignment with the class directions and overlaps among encoder directions.

To characterize these states without tying the analysis to a particular optimizer, we introduce the Gibbs measure

$$p_\beta(\boldsymbol{W} \mid \mathcal{D}) = \frac{1}{Z_\beta(\mathcal{D})} \exp[-\beta\mathcal{L}(\boldsymbol{W})]\,, \qquad (2)$$

where $\mathcal{L}$ is the regularized empirical objective in Eq. (1), and $\beta > 0$ is an auxiliary inverse temperature distinct from the InfoNCE temperature τ. The partition function $Z_\beta(\mathcal{D}) = \int \exp[-\beta\mathcal{L}(\boldsymbol{W})]\, d\boldsymbol{W}$ normalizes the measure and is finite for $\lambda_w > 0$. We study the quenched free-energy density $f_\beta = -(\beta N)^{-1}\mathbb{E}_{\mathcal{D}}[\log Z_\beta(\mathcal{D})]$, where the expectation is over training sets. As $\beta \to \infty$, the Gibbs measure concentrates on global minimizers of $\mathcal{L}$, and f_β approaches the average minimum regularized training objective per input dimension. We evaluate the high-dimensional limit of f_β using the replica method (Engel & Van den Broeck, 2001).

For a fixed training set, the replica method introduces independent encoder copies drawn from the Gibbs measure and summarizes them by their overlaps. For replicas a and b, the encoder–class overlaps are $R^a_{ic} = N^{-1}(\boldsymbol{w}^a_i)^\top \boldsymbol{\xi}_c$, and the weight overlaps are $Q^{ab}_{ij} = N^{-1}(\boldsymbol{w}^a_i)^\top \boldsymbol{w}^b_j$, where $(\boldsymbol{w}^a_i)^\top$ is the ith row of $\boldsymbol{W}^a$. We then impose the following reduced overlap structure.

Assumption 3.1 (Replica-symmetric and class-permutation-symmetric reduced ansatz)**.** *The replicated overlaps satisfy*

$$R^a_{ic} = r_o + (r_d - r_o)\delta_{ic}, \qquad Q^{aa}_{ij} = q\delta_{ij}, \qquad Q^{ab}_{ij} = Q\delta_{ij} \quad (a \neq b). \tag{3}$$

Realizability of the within-replica overlaps requires $q \geq \max\{(r_d - r_o)^2, [r_d + (C-1)r_o]^2\}$. Under Assumption 3.1, the encoder–class overlaps are replica independent, while the within- and cross-replica weight-overlap matrices are proportional to the identity. The scalar q is the squared norm of each encoder direction divided by N, and Q is the normalized overlap between corresponding directions in distinct replicas. All subsequent predictions are derived within this ansatz. Finite-dimensional simulations support its directly observable components. The encoder–class overlap matrix is well described by common matched and unmatched values, while the within-replica weight-overlap matrix is predominantly diagonal; see Appendix C.2. The cross-replica overlap Q is not directly observable from a single trained encoder. Conditional on the replicated weights and class labels, the preactivations are jointly Gaussian, with means and covariances determined by (r_d, r_o, q, Q).

Within this ansatz, we take the zero-temperature limit $\beta \to \infty$ with $q - Q = O(\beta^{-1})$. The resulting finite susceptibility χ is defined by

$$Q = q - \frac{\chi}{\beta} + o(\beta^{-1}), \qquad \chi = \lim_{\beta\to\infty} \beta(q - Q) \in (0, \infty). \tag{4}$$

The following result gives the zero-temperature reduction to four scalar order parameters and an effective single-group problem.

Result 3.2 (Zero-temperature characterization of InfoNCE)**.** *Under Assumption 3.1 and the scaling in Eq. (4), the relevant order parameters are (r_d, r_o, q, χ), where r_d and r_o are the matched and unmatched encoder–class overlaps, q is the encoder self-overlap, and χ is the finite zero-temperature susceptibility.*

By class-permutation symmetry, it suffices to fix an arbitrary anchor class c. Let $\mathbf{1}_C \in \mathbb{R}^C$ be the all-ones vector and $\boldsymbol{e}^{(c)} \in \mathbb{R}^C$ the cth standard basis vector, and define

$$\boldsymbol{\mu}_c = \sqrt{\kappa}\left[r_o \mathbf{1}_C + (r_d - r_o)\boldsymbol{e}^{(c)}\right]. \tag{5}$$

Sample

$$c^-_k \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left(\{1, \dots, C\} \setminus \{c\}\right), \qquad k = 1, \dots, K, \tag{6}$$

and independently draw $\boldsymbol{g}_0, \boldsymbol{g}^+_0, \{\boldsymbol{g}^-_{0,k}\}_{k=1}^K \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_C)$, collected in $\boldsymbol{G}_0 \in \mathbb{R}^{C(K+2)}$. The corresponding field centers are

$$\boldsymbol{u}_c = \boldsymbol{\mu}_c + \sqrt{(1-\gamma)q}\,\boldsymbol{g}_0, \qquad \boldsymbol{u}^+_c = \boldsymbol{\mu}_c + \sqrt{(1-\gamma)q}\,\boldsymbol{g}^+_0, \qquad \boldsymbol{u}^-_{c^-_k} = \boldsymbol{\mu}_{c^-_k} + \sqrt{(1-\gamma)q}\,\boldsymbol{g}^-_{0,k}. \tag{7}$$

For each realization of $\{c^-_k\}$ and $\boldsymbol{G}_0$, define the effective single-group potential

$$V^*_\chi = \min_{\boldsymbol{h}, \boldsymbol{h}^+, \{\boldsymbol{h}^-_k\}_{k=1}^K} \left\{ \log\left[1 + \sum_{k=1}^K \exp\left(\frac{\phi(\boldsymbol{h})^\top\phi(\boldsymbol{h}^-_k) - \phi(\boldsymbol{h})^\top\phi(\boldsymbol{h}^+)}{\tau}\right)\right] + \frac{\|\boldsymbol{h} - \boldsymbol{u}_c\|_2^2 + \|\boldsymbol{h}^+ - \boldsymbol{u}^+_c\|_2^2 + \sum_{k=1}^K \|\boldsymbol{h}^-_k - \boldsymbol{u}^-_{c^-_k}\|_2^2}{2(1-\gamma)\chi} \right\}. \tag{8}$$

Here $\gamma = \kappa/N \to 0$ in the proportional limit, while the optimization dimension $C(K+2)$ remains independent of N. Let $\boldsymbol{h}^*$, $\boldsymbol{h}^{+,*}$, and $\{\boldsymbol{h}^{-,*}_k\}_{k=1}^K$ denote minimizers, and define

$$\delta\boldsymbol{h} = \boldsymbol{h}^* - \boldsymbol{u}_c, \qquad \delta\boldsymbol{h}^+ = \boldsymbol{h}^{+,*} - \boldsymbol{u}^+_c, \qquad \delta\boldsymbol{h}^-_k = \boldsymbol{h}^{-,*}_k - \boldsymbol{u}^-_{c^-_k}. \tag{9}$$

The energetic contribution and reduced zero-temperature free-energy density are

$$\mathcal{G}_\alpha(r_d, r_o, q, \chi) = \alpha\, \mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[V_\chi^*\right], \tag{10}$$

and

$$f_0(r_d, r_o, q, \chi) = \frac{C}{2\chi}\left[r_d^2 + (C-1)r_o^2 - q\right] + \frac{C\lambda_w}{2}q + \mathcal{G}_\alpha(r_d, r_o, q, \chi). \tag{11}$$

At differentiable interior stationary points, Eq. (11) yields

$$\begin{cases} r_d = \dfrac{\alpha\sqrt{\kappa}}{C(1-\gamma)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\delta h_c + \delta h_c^+ + \displaystyle\sum_{k=1}^{K}\delta h^-_{k,c_k^-}\right], \\ r_o = \dfrac{\alpha\sqrt{\kappa}}{C(C-1)(1-\gamma)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\displaystyle\sum_{i\neq c}(\delta h_i + \delta h_i^+) + \sum_{k=1}^{K}\sum_{i\neq c_k^-}\delta h^-_{k,i}\right], \\ q = r_d^2 + (C-1)r_o^2 + \dfrac{\alpha}{C(1-\gamma)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\|\delta\boldsymbol{h}\|_2^2 + \|\delta\boldsymbol{h}^+\|_2^2 + \displaystyle\sum_{k=1}^{K}\|\delta\boldsymbol{h}_k^-\|_2^2\right], \\ 1 = \lambda_w\chi - \dfrac{\alpha}{C\sqrt{(1-\gamma)q}}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\boldsymbol{g}_0^\top\delta\boldsymbol{h} + (\boldsymbol{g}_0^+)^\top\delta\boldsymbol{h}^+ + \displaystyle\sum_{k=1}^{K}(\boldsymbol{g}_{0,k}^-)^\top\delta\boldsymbol{h}_k^-\right]. \end{cases} \tag{12}$$

Here δh_i, δh_i^+, and $\delta h_{k,i}^-$ are the ith coordinates of the corresponding displacement vectors.

The reduced theory determines stationary branches through self-consistency between the four order parameters and the effective single-group problem. Multiple branches may coexist and differ in local stability and free energy. Once a branch is determined, its order parameters specify the specialization amplitude, representation statistics including η_{cls}, and the population linear-probe error. We numerically continue these branches as α varies, assess local stability from relaxed free-energy curvatures, and determine thermodynamic preference by comparing their free energies in Eq. (11). The replica derivation is given in Appendix A, while Appendix D details the stability and branch-selection analyses, including the scalar response diagnostic. The resulting branch structure is compared with finite-dimensional training in Section 4.

To assess local stability against specialization perturbations, we introduce the class-averaged overlap $\bar{r} = [r_d + (C-1)r_o]/C$. Together with $\Delta r = r_d - r_o$, this gives $r_d = \bar{r} + (C-1)\Delta r/C$ and $r_o = \bar{r} - \Delta r/C$. We write the reduced free energy as $f_0(\Delta r, \boldsymbol{s}; \alpha)$, with $\boldsymbol{s} = (\bar{r}, q, \chi)$. At fixed Δr, the class-symmetric variables $\boldsymbol{s}$ are determined by the stationarity conditions $\partial_{\boldsymbol{s}} f_0 = 0$.

Result 3.3 (Relaxed stiffness of the unspecialized state)**.** *Assume that a smooth stationary continuation $\boldsymbol{s}^\star(\Delta r; \alpha)$ satisfying $\partial_{\boldsymbol{s}} f_0 = 0$ exists in a neighborhood of the unspecialized solution and that the Hessian block $\partial_{\boldsymbol{s}}^2 f_0$ is invertible at $\Delta r = 0$. The profiled free energy is*

$$F(\Delta r; \alpha) = f_0(\Delta r, \boldsymbol{s}^\star(\Delta r; \alpha); \alpha).$$

The relaxed specialization stiffness is its curvature at $\Delta r = 0$,

$$\Lambda_{\mathrm{sp}} = \partial_{\Delta r}^2 F(\Delta r; \alpha)\big|_{\Delta r = 0}. \tag{13}$$

Differentiation along the stationary continuation gives

$$\Lambda_{\mathrm{sp}} = \left[\partial_{\Delta r}^2 f_0 - \partial_{\Delta r}\partial_{\boldsymbol{s}} f_0 \left(\partial_{\boldsymbol{s}}^2 f_0\right)^{-1} \partial_{\boldsymbol{s}}\partial_{\Delta r} f_0\right]\Big|_{\Delta r = 0,\ \boldsymbol{s} = \boldsymbol{s}^\star(0;\alpha)}. \tag{14}$$

On the unspecialized branch, class-permutation symmetry implies $\partial_{\Delta r}\partial_{\boldsymbol{s}} f_0 = 0$, so the specialization mode decouples from the class-symmetric variables to quadratic order. The stiffness therefore reduces to

$$\Lambda_{\mathrm{sp}} = \frac{C-1}{\chi} + \partial_{\Delta r}^2 \mathcal{G}_\alpha\big|_{\Delta r = 0}, \tag{15}$$

where the derivative is evaluated at fixed $\boldsymbol{s}$ and all quantities are taken on the unspecialized branch.

The sign of Λ_{sp} determines the local stability of the unspecialized branch against specialization perturbations, with positive, zero, and negative values corresponding to stability, marginality, and instability, respectively. This quadratic criterion identifies loss of local stability but does not determine the full emergence route. Distinguishing continuous emergence from finite-amplitude competition requires higher-order local analysis, while coexistence and thermodynamic preference require branch continuation and free-energy comparison.

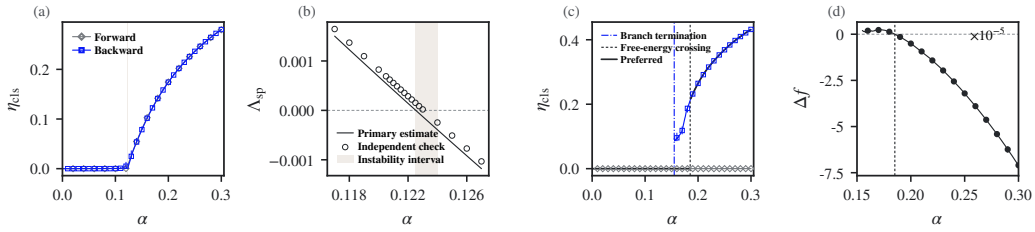

Figure 1: Replica theory predicts two routes to class organization. Gray diamonds and blue squares denote branches obtained by continuation from low and high α, respectively. In **(a,b)**, the two continuations coincide for the linear encoder, giving continuous emergence. In **(b)**, the specialization stiffness $\Lambda_{\rm sp}$ changes sign within the shaded interval. In **(c,d)**, smooth ReLU exhibits coexistence of unspecialized and specialized branches. Black segments mark the branch preferred by the RS free energy. The blue and black vertical lines mark the numerical endpoint of backward continuation and the RS free-energy crossing, respectively. In **(d)**, $\Delta f = f_0^{\rm backward} - f_0^{\rm forward}$, so $\Delta f < 0$ favors the specialized branch.

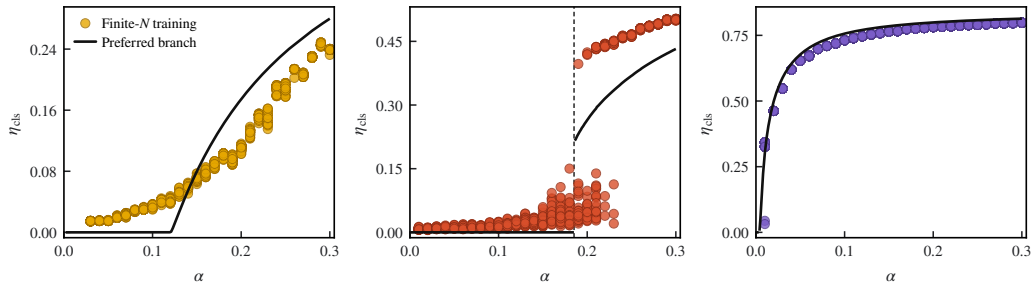

Figure 2: Finite-N training recovers the predicted qualitative emergence routes. Colored markers denote converged finite-N outcomes, and black curves show the branches preferred by the RS free energy. In both finite networks and theory, $\eta_{\rm cls}$ is computed from population moments, analytically for the linear encoder and by Gauss–Hermite quadrature for smooth ReLU, using trained weights and saddle-point order parameters, respectively. The left panel shows continuous emergence for the linear encoder with $C = 10$. The middle panel shows weakly and strongly organized outcomes across initializations for smooth ReLU with $C = 10$, with the dashed line marking the RS free-energy crossing. The right panel shows continuous emergence again for smooth ReLU with $C = 2$.

4 Two Routes to Class Organization

We solve the zero-temperature saddle-point equations in Eq. (12) across sample complexity α and compare the resulting stationary branches with finite-dimensional training under Eq. (1). Under Assumption 3.1, thermodynamic preference among locally stable branches is determined by the RS free energy. Multiple locally stable branches may coexist across the free-energy crossing (Biehl et al., 1998; Ahr et al., 1999). The saddle-point solver may therefore converge to either branch from different initializations.

Unless otherwise stated, theory and simulations use $C = M = 10$, $K = 5$, $\kappa = 15$, $\lambda_w = 0.01$, and $\tau = 0.1$. Forward and backward continuations start from low-α unspecialized and high-α specialized states, respectively. At each α, networks with $N = 10000$ are trained from 50 independent initializations using full-batch Adam followed by L-BFGS. Theoretical branches are distinguished by the normalized specialization amplitude $\Delta r/\sqrt{q}$, which is near zero for unspecialized branches and clearly nonzero for specialized ones. In simulations, weakly and strongly organized denote low- and high-$\eta_{\rm cls}$ outcomes, respectively; weak organization is not equated with exact unspecialization. Because the transitions occur at small α, the finite-N experiments contain relatively few training groups in the transition regime. We therefore emphasize the qualitative emergence routes. Appendix C.1 details optimization and convergence diagnostics, while Appendix C.2 reports quantitative comparisons of (r_d, r_o, q) and $\eta_{\rm cls}$ over $\alpha \in [0.1, 5]$.

For the linear encoder, forward and backward continuations trace the same branch, with both Δr and η_{cls} growing continuously from near zero as α increases (Fig. 1(a)). Numerical evaluation places the sign change of the relaxed specialization stiffness at $\alpha_{\mathrm{soft}} \in [0.121, 0.125]$, consistent with continuous specialization near the loss of local stability of the unspecialized branch (Fig. 1(b)). Finite-N training shows the same qualitative behavior, with class organization growing smoothly across initializations and no distinct weakly and strongly organized groups (Fig. 2, left).

By contrast, with all other parameters fixed, smooth ReLU exhibits coexisting unspecialized and specialized branches from $\alpha \approx 0.16$ (Fig. 1(c,d)). At the RS free-energy crossing, $\alpha_{\mathrm{th}} \in [0.173, 0.176]$, thermodynamic preference switches from the forward to the backward branch, and η_{cls} jumps from approximately 2.4×10^{-3} to 1.84×10^{-1}. The unspecialized branch nevertheless remains locally stable against specialization until $\alpha \approx 0.33$ (see Appendix D.3). Finite-N training similarly yields weakly and strongly organized outcomes across initializations (Fig. 2, middle). Across 12 smooth-ReLU configurations with $C \in \{2, 3, 8, 10\}$ and $K \in \{1, 5, 10\}$, $C = 2$ yields a single continuously specializing branch recovered by both continuations, consistent with finite-N training (Fig. 2, right). For $C = 3$, a specialized branch emerges continuously while a distinct nearly unspecialized branch persists. For $C \in \{8, 10\}$, specialized branches instead appear at finite amplitude and coexist with the unspecialized branch before free-energy selection (Figs. 7–8 in Appendix C.3). This qualitative dependence on C persists across the sampled values of K, while increasing K mainly shifts the characteristic transition scales.

We map the smooth-ReLU stationary states for $C = 10$ over $\alpha \in [0.01, 2]$ and $\tau \in [10^{-4}, 10^{4}]$. At each grid point, we solve the saddle-point system from multiple initializations, assess local stability, and compare RS free energies across stable solutions. The unspecialized branch is recovered throughout the scan, whereas specialized branches appear only over an intermediate temperature range. Within this range, varying α recovers the same coexistence and free-energy-selection mechanism described above. Crossing either temperature boundary produces a discontinuous change in η_{cls} on the preferred branch. Thus, temperature primarily determines where this mechanism is realized and the resulting degree of class organization (Fig. 3).

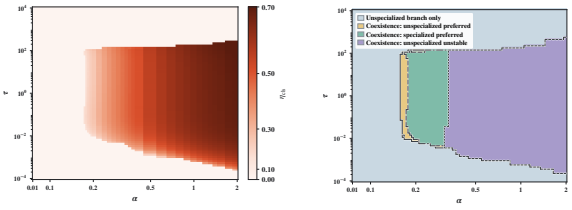

Figure 3: Smooth-ReLU (α, τ) phase structure. Left, η_{cls} on the preferred branch; right, branch existence, thermodynamic preference, and local stability.

At fixed C, the population-optimal multinomial linear-probe error is determined by $|\Delta_{\mathrm{LP}}|$, where $\Delta_{\mathrm{LP}} = \sqrt{\kappa/[(1-\gamma)q]}\,(r_d - r_o)$. For the linear encoder, η_{cls} increases monotonically with $|\Delta_{\mathrm{LP}}|$, linking greater class organization directly to lower probe error. The same probe-error relation holds for smooth ReLU, although η_{cls} additionally depends on the nonlinear operating point (see Appendix A.9). Probe performance therefore improves continuously along the continuous-specialization route and can improve abruptly when a finite-amplitude specialized branch becomes preferred. Finite-N simulations across six smooth-ReLU configurations recover the corresponding low-error regime under calibrated localist classification (see Appendix C.4). Beyond the solvable model, CIFAR-10 experiments show that η_{cls} follows the qualitative trend in projector linear-probe accuracy as the pretraining-set size increases, whereas participation-ratio and centered RankMe effective ranks can follow different trends (Garrido et al., 2023) (see Appendix C.6).

5 Local Nonlinear Signatures of Class Organization

We extend the local analysis of Result 3.3 beyond quadratic order to characterize nonlinear signatures of the emergence routes in Section 4. Assuming a sufficiently smooth stationary profile, the profiled RS free energy admits the local Landau-type expansion (Liu et al., 2023)

$$F(\Delta r; \alpha) = F_0 + \frac{a_2}{2}(\Delta r)^2 + \frac{a_3}{3}(\Delta r)^3 + \frac{a_4}{4}(\Delta r)^4 + O((\Delta r)^5), \tag{16}$$

where $F_0 = F(0; \alpha)$ and the coefficients depend on α and the model parameters. Because $\Delta r = 0$ is stationary, the linear term vanishes. The quadratic coefficient is $a_2 = \Lambda_{\mathrm{sp}}$, while $a_3 = \frac{1}{2}\partial^3_{\Delta r} F(0; \alpha)$ and $a_4 = \frac{1}{6}\partial^4_{\Delta r} F(0; \alpha)$ are derived in Appendix D.2.

Table 1: Local free-energy coefficients at representative sample complexities. For the linear encoder with $C = 10$ and sReLU with $C \in \{2, 3\}$, α is chosen near the zero crossing of $a_2 = \Lambda_{\mathrm{sp}}$. For sReLU with $C = 10$, $\alpha = 0.174$ lies near the free-energy crossing, where the unspecialized state remains locally stable with $a_2 > 0$. Here, sReLU denotes smooth ReLU. Values are reported as mean $\pm$ standard error over eight fully profiled disorder realizations.

Configuration	α	a_2	a_3	a_4
Linear, $C = 10$	0.1225	$(-2.66 \pm 0.36) \times 10^{-4}$	$(-0.72 \pm 0.59) \times 10^{-7}$	9.01 ± 0.03
sReLU, $C = 2$	0.0041	$(8.47 \pm 0.29) \times 10^{-4}$	$(3.2 \pm 6.2) \times 10^{-3}$	7.9 ± 2.2
sReLU, $C = 3$	0.010	$(4.095 \pm 0.037) \times 10^{-3}$	-0.152 ± 0.009	9.7 ± 1.1
sReLU, $C = 10$	0.174	$(1.872 \pm 0.012) \times 10^{-2}$	-0.807 ± 0.016	10.9 ± 7.2

Within the quartic approximation, $a_3 = 0$ and $a_4 > 0$ yield continuous onset as a_2 changes sign. For $a_3 < 0$ and $a_4 > 0$, positive finite-amplitude stationary points exist when $a_3^2 - 4a_2a_4 \geq 0$. In particular, $0 < a_2 < a_3^2/(4a_4)$ permits a finite-amplitude local minimum to coexist with the locally stable state $\Delta r = 0$. These conditions distinguish continuous onset from finite-amplitude competition within the local quartic landscape.

The cubic coefficient a_3 diagnoses local asymmetry rather than the global emergence route (Table 1). For the linear encoder with $C = 10$ and smooth ReLU with $C = 2$, both of which exhibit continuous onset, a_3 is consistent with zero and $a_4 > 0$. For smooth ReLU with $C = 10$, the negative a_3 indicates local asymmetry, while $a_2 > 0$ shows that the unspecialized state remains locally stable near the independently determined free-energy crossing. Smooth ReLU with $C = 3$ also has $a_3 < 0$, although its specialized branch emerges continuously while a nearly unspecialized branch persists (see Appendix C.3). A negative cubic response is therefore associated with local asymmetry but does not by itself distinguish the two emergence routes. For the linear encoder, an exact reflection symmetry makes the profiled free energy even in Δr, eliminating all odd-order terms (see Appendix D.2). For nonlinear activations, curvature couples the second-order shift of the class means to the first-order specialization response, producing a cubic contribution to the negative-minus-positive similarity difference.

At high τ, the leading cubic coefficient is

$$a_3 = -3\alpha \frac{K}{K+1} \frac{C-2}{C} \frac{\kappa^{3/2}}{\tau} \mathbb{E}[\phi'(X)]\mathbb{E}[\phi''(X)] + O(\tau^{-2}), \tag{17}$$

where $X = \sqrt{\kappa}\,\bar{r}_0 + \sqrt{(1-\gamma)q_0}\,Z$, $Z \sim \mathcal{N}(0,1)$, and $(\bar{r}_0, q_0)$ denote the unspecialized saddle values. Activation curvature and class structure enter through $\mathbb{E}[\phi'(X)]\mathbb{E}[\phi''(X)]$ and $(C-2)/C$, while $K/(K+1)$ sets the InfoNCE weighting. The leading term vanishes for linear activations or $C = 2$ and is negative for smooth ReLU with $C > 2$. At finite τ, higher-order couplings among negatives can modify the cubic coefficient; Table 1 reports its numerical values. Here the cubic contribution originates from positive–negative similarity comparisons, whereas supervised soft-committee specialization arises from teacher–student output matching (Oostwal et al., 2021; Afanah & Rosenow, 2026).

The cubic response captures local asymmetry, while branch continuation establishes coexistence and free-energy comparison determines thermodynamic selection. Whether this local response, evaluated along the unspecialized branch, can distinguish continuous onset from finite-amplitude branch formation without reconstructing the specialized branch remains open. The high-τ expression further motivates comparison across contrastive objectives. CLIP-type softmax and SigLIP-type pairwise sigmoid losses (Radford et al., 2021; Zhai et al., 2023) both compare similarities but differ in their normalization and coupling among negatives. Comparing their local nonlinear responses could separate effects of the shared similarity structure from loss-specific effects. Extending the analysis beyond signed similarity-difference objectives could further identify which properties of the objective govern the emergence route.

6 Related Work

Representation Geometry and Identifiability in Contrastive Learning. Prior theory has largely characterized what structures contrastive learning can recover, including class-relevant structure

and latent factors under latent-class, nonlinear-identifiability, and augmentation-based assumptions (Saunshi et al., 2019; Hyvärinen et al., 2019; Zimmermann et al., 2021; von Kügelgen et al., 2021; Lyu & Fu, 2022). Related work has studied the geometry induced by contrastive objectives through alignment, augmentation structure, cluster preservation, and minimizer geometry (Wang & Isola, 2020; HaoChen et al., 2021; Zhang et al., 2025; Parulekar et al., 2023; Koromilas et al., 2024). More recent analyses extend this picture to anisotropic variation, approximate sufficiency, and discriminative subspace recovery in Gaussian mixtures (Rusak et al., 2025; Lin & Mei, 2025; Bansal et al., 2025). Taken together, existing theory largely characterizes what representation structure is recovered and under what conditions, leaving open how class organization emerges with sample complexity and whether specialization develops continuously or through competing states.

Representation Variability and Equivalence in Contrastive Learning. Contrastive learning can select different predictive features or local optima under different function classes and optimization biases (Robinson et al., 2021; Saunshi et al., 2022; Tian, 2023; Xue et al., 2023). Related work has also examined cross-model variability arising from representational indeterminacies, including linear non-identifiability (Roeder et al., 2021) and hidden-unit permutation symmetries (Entezari et al., 2022; Ainsworth et al., 2023), as well as functional compatibility through model stitching (Bansal et al., 2021). These lines of work explain variability through optimization bias, representational indeterminacy, or cross-model compatibility, but do not address the coexistence of distinct macroscopic representation states within a common equilibrium landscape.

Phase Transitions in Neural Networks. Statistical-mechanical analyses of teacher–student networks have established specialization transitions and metastable regimes as sample complexity increases (Seung et al., 1992; Watkin et al., 1993; Biehl et al., 1998; Ahr et al., 1999), with later work showing that the transition can depend qualitatively on the activation function (Oostwal et al., 2021; Afanah & Rosenow, 2026). Related transition phenomena have been identified in self-supervised and representation-learning settings, including phase transitions in contrastive representations on physics-inspired datasets (Cy et al., 2023), collapse-related symmetry breaking (Liu et al., 2023), stepwise feature formation (Simon et al., 2023), spectral detectability transitions in masked self-supervised learning (Zurich et al., 2026), abrupt syntactic acquisition in masked language modeling (Chen et al., 2024), and positional–semantic transitions in attention (Cui et al., 2024). High-dimensional work has also examined nonlinear contrastive-learning dynamics (Meng & Wang, 2024) and label-based local contrastive learning (Xie et al., 2025). To our knowledge, phase-transition behavior underlying distinct routes to class organization in InfoNCE learning from unlabeled latent-class data has not been characterized theoretically.

7 Conclusion and discussion

We study how class organization emerges as sample complexity increases in a high-dimensional InfoNCE model with a single-layer encoder. Tracking the class-explained variance fraction along stationary branches reveals two routes. Specialization either emerges continuously from the unspecialized state or appears as a finite-amplitude specialized branch that becomes thermodynamically preferred at a higher sample complexity through an RS free-energy crossing. A local Landau-type expansion of the profiled free energy identifies a cubic response associated with finite-amplitude competition. Its leading high-InfoNCE-temperature form shows how activation curvature, class structure, and contrastive weighting contribute to this response.

As a minimal solvable model of the broader representation-learning problem, our setting retains only discrete latent class centers, so α primarily controls class-discriminative information. Recent self-supervised approaches increasingly target not only class discrimination but also the joint organization and prediction of semantic, spatial, and temporal structure (Mur-Labadia et al., 2026). Effective sample complexity in real data is therefore likely to depend not only on data volume but also on which learnable latent structures additional data provide. A broader theoretical challenge is to determine whether different structures emerge through distinct mechanisms as sample complexity increases and when optimization can access them. Class specialization is therefore a minimal instance of a broader question: which latent structures does self-supervised learning organize as effective data scale grows?

AI USE STATEMENT

In this work, generative AI tools assisted with checking and organizing the authors' theoretical derivations, discussing numerical methods and experimental diagnostics, implementing and debugging code, language editing, manuscript organization, LaTeX preparation, and selected literature searches and summaries. Generative AI was not used to generate experimental data or replace numerical computations; other mandatory disclosure categories are not applicable. AI-assisted derivations were independently verified against the original derivations, code was executed and tested, numerical results were produced through author-run computations, and references and citations were verified by the authors. The authors reviewed and approved the final text, mathematical conclusions, experimental results, and scientific claims and take full responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce the theoretical and numerical results. Appendix A develops the model, high-dimensional limit, overlap parametrization, reduced symmetry ansatz, free energy, and saddle-point equations, with the zero-temperature reduction given in Appendix A.7. Appendix B relates the specialization amplitude to class-explained variance and centered spectral effective ranks. Appendices C.1 and C.2 describe the finite-N protocol and theory–simulation comparisons. Appendices C.3, C.4, C.5, and C.6 cover the (C, K) scans, downstream generalization, unconditioned negative sampling, and CIFAR-10 experiments, with the corresponding theoretical derivations in Appendices A.8 and A.9. Appendix D analyzes specialization onset, finite-amplitude continuation, free-energy selection, local stability within the reduced ansatz, and the scalar response diagnostic, with further details in Appendices D.1, D.2, D.3, D.4, and D.5. Code for reproducing the reported results is included in the supplementary material.

REFERENCES

Assem Afanah and Bernd Rosenow. Continuous specialization transition in the soft committee machine with ReLU activation. arXiv preprint arXiv:2603.20010, 2026. URL https://arxiv.org/abs/2603.20010.

Martin Ahr, Michael Biehl, and Robert Urbanczik. Statistical physics and practical training of soft-committee machines. *The European Physical Journal B*, 10(3):583–588, 1999. doi: 10.1007/s100510050889.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.04836.

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966. doi: 10.2140/pjm.1966.16.1.

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

Parikshit Bansal, Ali Kavis, and Sujay Sanghavi. Understanding contrastive learning via Gaussian mixture models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 110853–110884, 2025. doi: 10.52202/085713-3703.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116.

Michael Biehl, Enno Schlösser, and Martin Ahr. Phase transitions in soft-committee machines. *Europhysics Letters*, 44(2):261–267, 1998. doi: 10.1209/epl/i1998-00466-6.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations*, pp. 14479–14510, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14755–14764, 2022.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36342–36389, 2024. doi: 10.52202/079017-1146.

Ali Cy, Anugrah Chemparathy, Michael Han, Rumen Dangovski, Peter Y Lu, and Marin Soljacic. Studying phase transitions in contrastive learning with physics-inspired datasets. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023.

J. R. L. de Almeida and D. J. Thouless. Stability of the Sherrington-Kirkpatrick solution of a spin glass model. *Journal of Physics A: Mathematical and General*, 11(5):983–990, 1978. doi: 10.1088/0305-4470/11/5/028.

Andreas Engel and Christian Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001. ISBN 9780521773072.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023.

Géza Györgyi. First-order transition to perfect generalization in a neural network with binary synapses. *Physical Review A*, 41(12):7097–7100, 1990. doi: 10.1103/PhysRevA.41.7097.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5000–5011, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Panagiotis Koromilas, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Bridging mini-batch and asymptotic analysis in contrastive learning: From InfoNCE to kernel-based losses. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25276–25301. PMLR, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Licong Lin and Song Mei. A statistical theory of contrastive learning via approximate sufficient statistics. In *Advances in Neural Information Processing Systems*, volume 38, pp. 127952–127996, 2025. doi: 10.52202/085713-4262.

Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1–3):503–528, 1989. doi: 10.1007/BF01589116.

Ziyin Liu, Ekdeep Singh Lubana, Masahito Ueda, and Hidenori Tanaka. What shapes the loss landscape of self supervised learning? In *International Conference on Learning Representations*, 2023.

Qi Lyu and Xiao Fu. On finite-sample identifiability of contrastive learning-based nonlinear independent component analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14582–14600. PMLR, 2022.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11071–11082, 2020.

Linghuan Meng and Chuang Wang. Training dynamics of nonlinear contrastive learning model in the high dimensional limit. *IEEE Signal Processing Letters*, 31:2535–2539, 2024. doi: 10.1109/LSP.2024.3434392.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. In *Computer Vision – ECCV 2026*, pp. 671–689. Springer, 2026.

Elisa Oostwal, Michiel Straat, and Michael Biehl. Hidden unit specialization in layered neural networks: ReLU vs. sigmoidal activation. *Physica A: Statistical Mechanics and its Applications*, 564:125517, 2021. doi: 10.1016/j.physa.2020.125517.

Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. InfoNCE loss provably learns cluster-preserving representations. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1914–1961. PMLR, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Joshua W. Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *Advances in Neural Information Processing Systems*, volume 34, pp. 4974–4986, 2021.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9030–9039. PMLR, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pp. 606–610, 2007.

Evgenia Rusak, Patrik Reizinger, Attila Juhos, Oliver Bringmann, Roland S. Zimmermann, and Wieland Brendel. InfoNCE: Identifying the gap between theory and practice. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4159–4167. PMLR, 2025.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 2019.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19250–19286. PMLR, 2022.

Henry Schwarze. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21):5781–5794, 1993. doi: 10.1088/0305-4470/26/21/017.

H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992. doi: 10.1103/PhysRevA.45.6056.

James B. Simon, Maksis Knutins, Ziyin Liu, Daniel Geisz, Abraham J. Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31852–31876. PMLR, 2023.

Haim Sompolinsky, Naftali Tishby, and H. Sebastian Seung. Learning from examples in large neural networks. *Physical Review Letters*, 65(13):1683–1686, 1990. doi: 10.1103/PhysRevLett.65.1683.

Yuandong Tian. Understanding the role of nonlinearity in training dynamics of contrastive learning. In *International Conference on Learning Representations*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. URL `https://arxiv.org/abs/1807.03748`.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467, 2021.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504. IEEE, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499–556, 1993. doi: 10.1103/RevModPhys.65.499.

Mingshan Xie, Yuchen Wang, and Haiping Huang. Local-contrastive-learning machine with both generalization and adversarial robustness: A statistical physics analysis. *Science China Physics, Mechanics & Astronomy*, 68:210511, 2025. doi: 10.1007/s11433-024-2504-8.

Yihao Xue, Siddharth Joshi, Eric Gan, Pin-Yu Chen, and Baharan Mirzasoleiman. Which features are learnt by contrastive learning? On the role of simplicity bias in class collapse and feature suppression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38938–38970. PMLR, 2023.

Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. In *International Conference on Learning Representations*, pp. 38101–38123, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Qi Zhang, Yifei Wang, and Yisen Wang. An augmentation overlap theory of contrastive learning. *Journal of Machine Learning Research*, 26(228):1–42, 2025.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 2021.

Arie Wortsman Zurich, Federica Gerace, Bruno Loureiro, and Yue M. Lu. A random matrix theory of masked self-supervised learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

A Replica Theory

This appendix derives the scalar replica equations used in the main text. We first recast the replicated problem in terms of encoder–class and replica overlaps, which determine a Gaussian effective description of the preactivations (Secs. A.1–A.2). Imposing replica symmetry (RS) and class-permutation symmetry reduces the theory to four physical and four conjugate scalars, yielding the finite-temperature free energy, saddle-point equations, and energetic response terms (Secs. A.3–A.6). The zero-temperature limit then reduces the description to the four physical variables (r_d, r_o, q, χ) used in the main text (Sec. A.7). The modifications for unconditioned negative sampling and the mapping from the zero-temperature order parameters to population linear-probe error are given in Secs. A.8 and A.9, respectively.

A.1 Replica Formulation for InfoNCE

We consider the Gaussian-mixture InfoNCE model (van den Oord et al., 2018) introduced in Section 2. We work in the proportional limit $N, P \to \infty$ at fixed sample complexity $\alpha = P/N = \Theta_N(1)$, with $C, M, K, \kappa, \tau, \lambda_w = \Theta_N(1)$ and $\gamma = \kappa/N$. The class directions are collected in $\boldsymbol{\Xi} \in \mathbb{R}^{C\times N}$ and satisfy $N^{-1}\boldsymbol{\Xi}\boldsymbol{\Xi}^\top = I_C$. Throughout the replica calculation, we condition on $\boldsymbol{\Xi}$.

The training set $\mathcal{D} = \{\mathcal{D}^p\}_{p=1}^P$ contains P independent contrastive groups. For each group p, we sample the anchor class c^p uniformly from $\{1, \dots, C\}$ and generate

$$\boldsymbol{x}^p = \sqrt{\gamma}\,\boldsymbol{\xi}_{c^p} + \sqrt{1-\gamma}\,\boldsymbol{z}^p, \tag{18}$$

$$\boldsymbol{x}^{+p} = \sqrt{\gamma}\,\boldsymbol{\xi}_{c^p} + \sqrt{1-\gamma}\,\boldsymbol{z}^{+p}, \tag{19}$$

$$\boldsymbol{x}_k^{-p} = \sqrt{\gamma}\,\boldsymbol{\xi}_{c_k^{-p}} + \sqrt{1-\gamma}\,\boldsymbol{z}_k^{-p}, \qquad k = 1, \dots, K. \tag{20}$$

Here, $\boldsymbol{z}^p, \boldsymbol{z}^{+p}, \boldsymbol{z}_1^{-p}, \dots, \boldsymbol{z}_K^{-p} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, I_N)$ and $c_k^{-p} \mid c^p = c \sim \mathrm{Unif}(\{1, \dots, C\} \setminus \{c\})$. The resulting conditional negative distribution is

$$P_{\mathrm{neg}}(\boldsymbol{x}^- \mid c) = \frac{1}{C-1} \sum_{c' \neq c}^{C} \mathcal{N}\big(\boldsymbol{x}^-; \sqrt{\gamma}\,\boldsymbol{\xi}_{c'}, (1-\gamma) I_N\big)\,. \tag{21}$$

For $\boldsymbol{W} \in \mathbb{R}^{M\times N}$, define the preactivation and representation maps

$$\boldsymbol{h}(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^M, \qquad \boldsymbol{v}(\boldsymbol{x}) = \phi(\boldsymbol{h}(\boldsymbol{x})) \in \mathbb{R}^M, \tag{22}$$

where ϕ acts coordinatewise. The regularized empirical objective is

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\boldsymbol{W}) &= \sum_{p=1}^{P} \mathcal{L}^p(\boldsymbol{W}) + \frac{\lambda_w}{2}\|\boldsymbol{W}\|_F^2 \\ &= -\sum_{p=1}^{P} \log \frac{\exp\big((\boldsymbol{v}^p)^\top \boldsymbol{v}^{+p}/\tau\big)}{\exp\big((\boldsymbol{v}^p)^\top \boldsymbol{v}^{+p}/\tau\big) + \sum_{k=1}^{K} \exp\big((\boldsymbol{v}^p)^\top \boldsymbol{v}_k^{-p}/\tau\big)} + \frac{\lambda_w}{2}\|\boldsymbol{W}\|_F^2. \end{aligned} \tag{23}$$

Here, $\mathcal{L}^p$ denotes the InfoNCE loss for group p, in which the positive pair $(\boldsymbol{v}^p, \boldsymbol{v}^{+p})$ is contrasted with the K negatives $\{\boldsymbol{v}_k^{-p}\}_{k=1}^K$.

We introduce the Gibbs measure

$$p_\beta(\boldsymbol{W} \mid \mathcal{D}) = \frac{1}{Z_\beta(\mathcal{D})} \exp\left[-\beta \mathcal{L}_\mathcal{D}(\boldsymbol{W})\right], \qquad Z_\beta(\mathcal{D}) = \int d\boldsymbol{W}\, e^{-\beta \mathcal{L}_\mathcal{D}(\boldsymbol{W})}, \tag{24}$$

where β is an auxiliary inverse temperature distinct from the InfoNCE temperature τ. The quenched free-energy density is

$$f_{\beta,N} = -\frac{1}{\beta N} \mathbb{E}_{\mathcal{D}|\boldsymbol{\Xi}} \log Z_\beta(\mathcal{D}). \tag{25}$$

For any $Z > 0$,

$$\log Z = \lim_{n \to 0} \frac{Z^n - 1}{n}. \tag{26}$$

The replica method evaluates the replicated moment for positive integer n and formally continues the resulting expression to $n \to 0$ (Engel & Van den Broeck, 2001). Taking the high-dimensional limit before the replica limit gives

$$-\beta f_\beta = \lim_{n \to 0} \frac{1}{n} \lim_{N \to \infty} \frac{1}{N} \log \mathbb{E}_{\mathcal{D}|\boldsymbol{\Xi}} \left[Z_\beta(\mathcal{D})^n\right]. \tag{27}$$

For integer $n \geq 1$, introduce replicated weights $\{\boldsymbol{W}^a\}_{a=1}^n$, with $\boldsymbol{W}^a \in \mathbb{R}^{M \times N}$. Then

$$\begin{aligned} \mathbb{E}_{\mathcal{D}|\boldsymbol{\Xi}} \left[Z_\beta(\mathcal{D})^n\right] &= \mathbb{E}_{\mathcal{D}|\boldsymbol{\Xi}} \int \prod_{a=1}^n d\boldsymbol{W}^a \, \exp\left[-\beta \sum_{a=1}^n \mathcal{L}_\mathcal{D}(\boldsymbol{W}^a)\right] \\ &= \int \prod_{a=1}^n d\boldsymbol{W}^a \, \exp\left[-\frac{\beta \lambda_w}{2} \sum_{a=1}^n \|\boldsymbol{W}^a\|_F^2\right] \prod_{p=1}^P \mathbb{E}_{\mathcal{D}^p|\boldsymbol{\Xi}} \exp\left[-\beta \sum_{a=1}^n \mathcal{L}^p(\boldsymbol{W}^a)\right]. \end{aligned} \tag{28}$$

Because the P training groups are conditionally i.i.d. given $\boldsymbol{\Xi}$, the replicated data contribution factorizes into a single-group factor raised to the power $P = \alpha N$.

A.2 Gaussian Preactivations and Overlap Constraints

The replicated data term can be expressed through the preactivations entering the contrastive loss. For fixed replicated weights $\{\boldsymbol{W}^a\}_{a=1}^n$, let $(\boldsymbol{w}_i^a)^\top$ denote row i of $\boldsymbol{W}^a$ and define $h_i^a = (\boldsymbol{w}_i^a)^\top \boldsymbol{x}/\sqrt{N}$, with $i = 1, \ldots, M$. For an input from class c, $\boldsymbol{x} = \sqrt{\gamma}\, \boldsymbol{\xi}_c + \sqrt{1-\gamma}\, \boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, I_N)$, giving

$$h_i^a = \sqrt{\kappa}\, \frac{(\boldsymbol{w}_i^a)^\top \boldsymbol{\xi}_c}{N} + \sqrt{\frac{1-\gamma}{N}}\, (\boldsymbol{w}_i^a)^\top \boldsymbol{z}. \tag{29}$$

Conditional on c, the replicated preactivations are jointly Gaussian, with means and covariances determined by the encoder–class and weight overlaps

$$\begin{aligned} R_{ic}^a &= \frac{1}{N} (\boldsymbol{w}_i^a)^\top \boldsymbol{\xi}_c, & Q_{ij}^{ab} &= \frac{1}{N} (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^b, \\ \mathbb{E}[h_i^a \mid c] &= \sqrt{\kappa}\, R_{ic}^a, & \mathrm{Cov}(h_i^a, h_j^b \mid c) &= (1-\gamma) Q_{ij}^{ab}, \end{aligned} \tag{30}$$

where $i, j = 1, \ldots, M$ and $c = 1, \ldots, C$. Once R and Q are fixed, the microscopic weights affect the data average only through these Gaussian moments. Let $\boldsymbol{H} = (\boldsymbol{h}^1, \ldots, \boldsymbol{h}^n) \in \mathbb{R}^{nM}$ denote the stacked replica preactivations. For class c,

$$\boldsymbol{H} \mid c \sim \mathcal{N}_{nM}\big(\sqrt{\kappa}\, \boldsymbol{R}_c,\, (1-\gamma)\boldsymbol{Q}\big), \tag{31}$$

where $(\boldsymbol{R}_c)_{(a,i)} = R_{ic}^a$ and $\boldsymbol{Q}_{(a,i),(b,j)} = Q_{ij}^{ab}$. We denote this Gaussian measure by $\mathcal{P}_c(d\boldsymbol{H})$.

We next treat the overlaps as independent integration variables by enforcing their definitions with delta functions and introducing conjugate variables through Fourier representations. We omit normalization factors that are independent of the saddle variables. For the encoder–class overlaps,

$$\begin{aligned} 1 &= \int \mathcal{D}R \prod_{a=1}^{n} \prod_{i=1}^{M} \prod_{c=1}^{C} \delta\big(NR_{ic}^{a} - (\boldsymbol{w}_i^a)^\top \boldsymbol{\xi}_c\big) \\ &\propto \int \mathcal{D}R\, \mathcal{D}\hat{R} \exp\left[\sum_{a=1}^{n} \sum_{i=1}^{M} \sum_{c=1}^{C} \hat{R}_{ic}^{a} \left(NR_{ic}^{a} - (\boldsymbol{w}_i^a)^\top \boldsymbol{\xi}_c\right)\right]. \end{aligned} \tag{32}$$

For the weight overlaps, write $q_{ij}^a \equiv Q_{ij}^{aa}$ for the self-replica blocks and retain Q_{ij}^{ab} for $a < b$. Since q^a is symmetric in the representation indices, its independent entries are enforced by $\hat{q}_{ij}^a$ with $1 \le i \le j \le M$, whereas each cross-replica block is enforced by $\hat{Q}_{ij}^{ab}$:

$$\begin{aligned} 1 &= \int \mathcal{D}q\, \mathcal{D}Q \prod_{a=1}^{n} \prod_{1\le i\le j\le M} \delta\big(Nq_{ij}^{a} - (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^a\big) \prod_{a<b}^{M} \prod_{i,j=1} \delta\big(NQ_{ij}^{ab} - (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^b\big) \\ &\propto \int \mathcal{D}q\, \mathcal{D}\hat{q}\, \mathcal{D}Q\, \mathcal{D}\hat{Q} \exp\left[\sum_{a=1}^{n} \sum_{1\le i\le j\le M} \hat{q}_{ij}^{a} \left(Nq_{ij}^{a} - (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^a\right) + \sum_{a<b} \sum_{i,j=1}^{M} \hat{Q}_{ij}^{ab} \left(NQ_{ij}^{ab} - (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^b\right)\right]. \end{aligned} \tag{33}$$

The conjugate variables are integrated along the corresponding Fourier contours, with saddle-independent normalization factors omitted.

Substituting these constraints into Eq. (28) and using the Gaussian preactivation law gives

$$\begin{aligned} \mathbb{E}_{\mathcal{D}|\boldsymbol{\Xi}}\left[Z_\beta(\mathcal{D})^n\right] \propto \int \mathcal{D}R\, \mathcal{D}\hat{R}\, \mathcal{D}q\, \mathcal{D}\hat{q}\, \mathcal{D}Q\, \mathcal{D}\hat{Q} \int \prod_{a=1}^{n} d\boldsymbol{W}^a \exp\left[-\frac{\beta\lambda_w}{2} \sum_{a=1}^{n} \|\boldsymbol{W}^a\|_F^2\right] \\ \times \exp\left[\sum_{a=1}^{n} \sum_{i=1}^{M} \sum_{c=1}^{C} \hat{R}_{ic}^{a} \left(NR_{ic}^{a} - (\boldsymbol{w}_i^a)^\top \boldsymbol{\xi}_c\right)\right] \\ \times \exp\left[\sum_{a=1}^{n} \sum_{1\le i\le j\le M} \hat{q}_{ij}^{a} \left(Nq_{ij}^{a} - (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^a\right)\right] \\ \times \exp\left[\sum_{a<b} \sum_{i,j=1}^{M} \hat{Q}_{ij}^{ab} \left(NQ_{ij}^{ab} - (\boldsymbol{w}_i^a)^\top \boldsymbol{w}_j^b\right)\right] \\ \times \left\{ \frac{1}{C(C-1)^K} \sum_{c=1}^{C} \sum_{\{c_k^- \neq c\}_{k=1}^{K}} \int \left[\mathcal{P}_c(d\boldsymbol{H})\, \mathcal{P}_c(d\boldsymbol{H}^+) \prod_{k=1}^{K} \mathcal{P}_{c_k^-}(d\boldsymbol{H}_k^-)\right] \right. \\ \left. \times \left[\prod_{a=1}^{n} \frac{\exp\big(\phi(\boldsymbol{h}^a)^\top \phi(\boldsymbol{h}^{a+})/\tau\big)}{\exp(\phi(\boldsymbol{h}^a)^\top \phi(\boldsymbol{h}^{a+})/\tau) + \sum_{k=1}^{K} \exp\big(\phi(\boldsymbol{h}^a)^\top \phi(\boldsymbol{h}_k^{a-})/\tau\big)}\right]^{\beta} \right\}^{P}. \end{aligned} \tag{34}$$

Collecting the extensive terms, the replicated moment can be written as

$$\mathbb{E}_{\mathcal{D}|\boldsymbol{\Xi}}\left[Z_\beta(\mathcal{D})^n\right] \propto \int \mathcal{D}R\, \mathcal{D}\hat{R}\, \mathcal{D}q\, \mathcal{D}\hat{q}\, \mathcal{D}Q\, \mathcal{D}\hat{Q} \exp[-N\beta f_n], \tag{35}$$

where the replicated action decomposes as

$$-\beta f_n = \Psi_0 + \Psi_S + \Psi_E. \tag{36}$$

The overlap–conjugate contribution is

$$\Psi_0 = \sum_{a=1}^{n} \sum_{i=1}^{M} \sum_{c=1}^{C} \hat{R}_{ic}^{a} R_{ic}^{a} + \sum_{a=1}^{n} \sum_{1\le i\le j\le M} \hat{q}_{ij}^{a} q_{ij}^{a} + \sum_{a<b} \sum_{i,j=1}^{M} \hat{Q}_{ij}^{ab} Q_{ij}^{ab}. \tag{37}$$

The weight contribution is

$$\Psi_S = \frac{1}{N}\log\int\prod_{a=1}^{n} d\boldsymbol{W}^a\,\exp\left[-\frac{\beta\lambda_w}{2}\sum_{a=1}^{n}\|\boldsymbol{W}^a\|_F^2\right] \times\exp\left[-\sum_{a=1}^{n}\sum_{i=1}^{M}\sum_{c=1}^{C}\hat{R}_{ic}^{a}(\boldsymbol{w}_i^a)^\top\boldsymbol{\xi}_c-\sum_{a=1}^{n}\sum_{1\le i\le j\le M}\hat{q}_{ij}^{a}(\boldsymbol{w}_i^a)^\top\boldsymbol{w}_j^a-\sum_{a<b}\sum_{i,j=1}^{M}\hat{Q}_{ij}^{ab}(\boldsymbol{w}_i^a)^\top\boldsymbol{w}_j^b\right]. \tag{38}$$

Finally, since $P=\alpha N$, the energetic contribution is

$$\Psi_E=\alpha\log\left\{\frac{1}{C(C-1)^K}\sum_{c=1}^{C}\sum_{\{c_k^-\neq c\}_{k=1}^K}\int\mathcal{P}_c(d\boldsymbol{H})\,\mathcal{P}_c(d\boldsymbol{H}^+)\prod_{k=1}^{K}\mathcal{P}_{c_k^-}(d\boldsymbol{H}_k^-) \times\left[\prod_{a=1}^{n}\frac{\exp\left[\phi(\boldsymbol{h}^a)^\top\phi(\boldsymbol{h}^{a+})/\tau\right]}{\exp[\phi(\boldsymbol{h}^a)^\top\phi(\boldsymbol{h}^{a+})/\tau]+\sum_{k=1}^{K}\exp\left[\phi(\boldsymbol{h}^a)^\top\phi(\boldsymbol{h}_k^{a-})/\tau\right]}\right]^\beta\right\}. \tag{39}$$

A.3 Replica-Symmetric Reduced Ansatz

We set $M=C$ and impose replica symmetry (RS), class-permutation symmetry, and coordinate-diagonal weight overlaps. Sec. C.2 assesses the observable overlap structure in finite-N simulations. Let $\boldsymbol{J}_C=\mathbf{1}_C\mathbf{1}_C^\top$, where $\mathbf{1}_C$ is the all-ones vector.

The encoder–class overlaps are replica-independent and have the matched–unmatched form

$$\begin{aligned}\boldsymbol{R}^a&=\boldsymbol{R}=(r_d-r_o)\boldsymbol{I}_C+r_o\boldsymbol{J}_C,\\ \hat{\boldsymbol{R}}^a&=\hat{\boldsymbol{R}}=(\hat{r}_d-\hat{r}_o)\boldsymbol{I}_C+\hat{r}_o\boldsymbol{J}_C,\end{aligned} \tag{40}$$

where the hatted parameters are auxiliary conjugate variables in the reduced functional. Writing $\boldsymbol{q}^a\equiv\boldsymbol{Q}^{aa}$ for the self-replica blocks, we restrict the physical weight overlaps to

$$\boldsymbol{q}^a=q\boldsymbol{I}_C,\qquad \boldsymbol{Q}^{ab}=Q\boldsymbol{I}_C\quad(a\neq b). \tag{41}$$

The reduced calculation therefore involves four physical variables,

$$(r_d,r_o,q,Q), \tag{42}$$

and four auxiliary conjugates,

$$(\hat{r}_d,\hat{r}_o,\hat{q},\hat{Q}). \tag{43}$$

To obtain the overlap contribution, we first eliminate the conjugate variables at the full matrix level. Under the reduced ansatz, the stacked physical overlaps are

$$\begin{aligned}\mathsf{R}&=\mathbf{1}_n\otimes\boldsymbol{R}\in\mathbb{R}^{nC\times C},\\ \mathsf{Q}&=(q-Q)\boldsymbol{I}_n\otimes\boldsymbol{I}_C+Q\,\mathbf{1}_n\mathbf{1}_n^\top\otimes\boldsymbol{I}_C\in\mathbb{R}^{nC\times nC},\end{aligned} \tag{44}$$

where $\mathbf{1}_n$ is the n-dimensional all-ones vector. Projecting the weights onto the subspace orthogonal to the class directions yields the residual Gram matrix

$$\mathsf{S}=\mathsf{Q}-\mathsf{R}\mathsf{R}^\top. \tag{45}$$

The overlap-dependent entropic contribution is determined by $\log\det\mathsf{S}$. Defining $d=q-Q>0$, the reduced ansatz gives

$$\mathsf{S}=d\,\boldsymbol{I}_n\otimes\boldsymbol{I}_C+\mathbf{1}_n\mathbf{1}_n^\top\otimes\left(Q\boldsymbol{I}_C-\boldsymbol{R}\boldsymbol{R}^\top\right). \tag{46}$$

Within the positive-definite domain, decomposing replica space into the uniform direction and its orthogonal complement gives

$$\begin{aligned}\log\det\mathsf{S}&=(n-1)C\log d+\log\det\left[d\boldsymbol{I}_C+n\left(Q\boldsymbol{I}_C-\boldsymbol{R}\boldsymbol{R}^\top\right)\right]\\ &=nC\log d+\log\det\left[\boldsymbol{I}_C+\frac{n}{d}\left(Q\boldsymbol{I}_C-\boldsymbol{R}\boldsymbol{R}^\top\right)\right].\end{aligned} \tag{47}$$

The formal replica limit then yields

$$\begin{aligned}\lim_{n\to 0}\frac{1}{2n}\log\det \mathsf{S} &= \frac{C}{2}\log d + \frac{CQ - \mathrm{Tr}(\boldsymbol{R}\boldsymbol{R}^\top)}{2d} \\ &= \frac{C}{2}\log d + \frac{C\left[Q - r_d^2 - (C-1)r_o^2\right]}{2d}.\end{aligned} \tag{48}$$

This is the reduced overlap contribution used below. The scalar conjugates provide an auxiliary parametrization of this contribution. Their stationarity equations enforce stationarity within the chosen physical sector and do not require the full matrix saddle to close under diagonal conjugates.

A.4 Finite-Temperature Free Energy

We now evaluate the three contributions to the replicated action under the replica-symmetric scalar ansatz of Sec. A.3 before taking the replica limit $n \to 0$.

Substituting the scalar overlap structures into Eq. (37) gives

$$\begin{aligned}\Psi_0|_{\mathrm{RS}} &= \sum_{a=1}^{n}\sum_{i,c=1}^{C}\hat{R}_{ic}^a R_{ic}^a + \sum_{a=1}^{n}\sum_{i=1}^{C}\hat{q}_{ii}^a q_{ii}^a + \sum_{a<b}\sum_{i=1}^{C}\hat{Q}_{ii}^{ab}Q_{ii}^{ab} \\ &= n\sum_{i,c=1}^{C}\left[\hat{r}_o + (\hat{r}_d - \hat{r}_o)\delta_{ic}\right]\left[r_o + (r_d - r_o)\delta_{ic}\right] \\ &\quad + nC\hat{q}q + \frac{n(n-1)}{2}C\hat{Q}Q \\ &= nC\left[\hat{r}_d r_d + (C-1)\hat{r}_o r_o + \hat{q}q\right] + \frac{n(n-1)C}{2}\hat{Q}Q.\end{aligned} \tag{49}$$

The second equality inserts the matched–unmatched forms of R_{ic} and $\hat{R}_{ic}$. The final equality counts C matched pairs with $i = c$ and $C(C-1)$ unmatched pairs with $i \neq c$.

Dividing Eq. (49) by n and taking $n \to 0$ gives

$$\begin{aligned}\psi_0 &\equiv \lim_{n\to 0}\frac{1}{n}\,\Psi_0|_{\mathrm{RS}} \\ &= C\left[\hat{r}_d r_d + (C-1)\hat{r}_o r_o + \hat{q}q\right] + \lim_{n\to 0}\frac{(n-1)C}{2}\hat{Q}Q \\ &= C\left[\hat{r}_d r_d + (C-1)\hat{r}_o r_o + \hat{q}q - \frac{1}{2}\hat{Q}Q\right].\end{aligned} \tag{50}$$

Applying the RS ansatz to the weight contribution Ψ_S gives

$$
\begin{aligned}
\Psi_S|_{\mathrm{RS}} &= \frac{1}{N}\log\int\prod_{a,i,\nu}dw_{i\nu}^a\,\exp\left[-\frac{\beta\lambda_w}{2}\sum_{a,i,\nu}(w_{i\nu}^a)^2-\sum_{a,i,c,\nu}\hat{R}_{ic}\xi_{c\nu}w_{i\nu}^a-\hat{q}\sum_{a,i,\nu}(w_{i\nu}^a)^2-\hat{Q}\sum_{a<b,i,\nu}w_{i\nu}^a w_{i\nu}^b\right]\\
&= \frac{1}{N}\log\int\prod_{a,i,\nu}dw_{i\nu}^a\,\exp\left[-\sum_{a,i,c,\nu}\hat{R}_{ic}\xi_{c\nu}w_{i\nu}^a-\frac{\beta\lambda_w+2\hat{q}-\hat{Q}}{2}\sum_{a,i,\nu}(w_{i\nu}^a)^2-\frac{\hat{Q}}{2}\sum_{i,\nu}\left(\sum_a w_{i\nu}^a\right)^2\right]\\
&= \frac{1}{N}\log\int\prod_{i,\nu}Dz_{i\nu}\int\prod_{a,i,\nu}dw_{i\nu}^a\,\exp\left[-\frac{\beta\lambda_w+2\hat{q}-\hat{Q}}{2}\sum_{a,i,\nu}(w_{i\nu}^a)^2+\sum_{a,i,\nu}\left(z_{i\nu}\sqrt{-\hat{Q}}-\sum_c\hat{R}_{ic}\xi_{c\nu}\right)w_{i\nu}^a\right]\\
&= \frac{1}{N}\sum_{i,\nu}\log\int Dz_{i\nu}\left[\int dw\,\exp\left(-\frac{\beta\lambda_w+2\hat{q}-\hat{Q}}{2}w^2+\left[z_{i\nu}\sqrt{-\hat{Q}}-\sum_c\hat{R}_{ic}\xi_{c\nu}\right]w\right)\right]^n\\
&= \frac{1}{N}\sum_{i,\nu}\log\int Dz_{i\nu}\left(\frac{2\pi}{\beta\lambda_w+2\hat{q}-\hat{Q}}\right)^{n/2}\exp\left[\frac{n\left(z_{i\nu}\sqrt{-\hat{Q}}-\sum_c\hat{R}_{ic}\xi_{c\nu}\right)^2}{2(\beta\lambda_w+2\hat{q}-\hat{Q})}\right]\\
&= \frac{nC}{2}\log\frac{2\pi}{\beta\lambda_w+2\hat{q}-\hat{Q}}-\frac{C}{2}\log\left(1+\frac{n\hat{Q}}{\beta\lambda_w+2\hat{q}-\hat{Q}}\right)+\frac{n\sum_{i,\nu}\left(\sum_c\hat{R}_{ic}\xi_{c\nu}\right)^2}{2N[\beta\lambda_w+2\hat{q}+(n-1)\hat{Q}]}\\
&= \frac{nC}{2}\log\frac{2\pi}{\beta\lambda_w+2\hat{q}-\hat{Q}}-\frac{C}{2}\log\left(1+\frac{n\hat{Q}}{\beta\lambda_w+2\hat{q}-\hat{Q}}\right)+\frac{nC[\hat{r}_d^2+(C-1)\hat{r}_o^2]}{2[\beta\lambda_w+2\hat{q}+(n-1)\hat{Q}]}.
\end{aligned}
\tag{51}
$$

The second equality follows from

$$
\sum_{a<b}w^a w^b=\frac{1}{2}\left[\left(\sum_a w^a\right)^2-\sum_a(w^a)^2\right]. \tag{52}
$$

The third equality applies the Gaussian decoupling

$$
\exp\left[-\frac{\hat{Q}}{2}\left(\sum_a w^a\right)^2\right]=\int Dz\,\exp\left[z\sqrt{-\hat{Q}}\sum_a w^a\right], \tag{53}
$$

with

$$
Dz=\frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz. \tag{54}
$$

The fourth equality factorizes the integral over (i,ν) and replicas. The fifth evaluates the scalar Gaussian integral over w, and the sixth performs the remaining Gaussian integral over z. The final equality follows from the orthogonality of the class directions,

$$
\frac{1}{N}\sum_{\nu=1}^{N}\xi_{c\nu}\xi_{c'\nu}=\delta_{cc'}, \tag{55}
$$

which implies

$$
\frac{1}{N}\sum_{i,\nu}\left(\sum_c\hat{R}_{ic}\xi_{c\nu}\right)^2=C\left[\hat{r}_d^2+(C-1)\hat{r}_o^2\right]. \tag{56}
$$

Dividing Eq. (51) by n and taking $n \to 0$ gives

$$\begin{aligned}
\psi_S &\equiv \lim_{n\to 0} \frac{1}{n}\, \Psi_S|_{\mathrm{RS}} \\
&= \frac{C}{2} \log \frac{2\pi}{\beta\lambda_w + 2\hat{q} - \hat{Q}} - \frac{C}{2} \lim_{n\to 0} \frac{1}{n} \log\left(1 + \frac{n\hat{Q}}{\beta\lambda_w + 2\hat{q} - \hat{Q}}\right) + \frac{C}{2} \lim_{n\to 0} \frac{\hat{r}_d^2 + (C-1)\hat{r}_o^2}{\beta\lambda_w + 2\hat{q} + (n-1)\hat{Q}} \\
&= \frac{C}{2} \log \frac{2\pi}{\beta\lambda_w + 2\hat{q} - \hat{Q}} - \frac{C\hat{Q}}{2(\beta\lambda_w + 2\hat{q} - \hat{Q})} + \frac{C[\hat{r}_d^2 + (C-1)\hat{r}_o^2]}{2(\beta\lambda_w + 2\hat{q} - \hat{Q})} \\
&= \frac{C}{2}\left[\log \frac{2\pi}{\beta\lambda_w + 2\hat{q} - \hat{Q}} + \frac{\hat{r}_d^2 + (C-1)\hat{r}_o^2 - \hat{Q}}{\beta\lambda_w + 2\hat{q} - \hat{Q}}\right].
\end{aligned} \tag{57}$$

Here we use $\lim_{n\to 0} n^{-1}\log(1+nx) = x$ and $\beta\lambda_w + 2\hat{q} + (n-1)\hat{Q} \to \beta\lambda_w + 2\hat{q} - \hat{Q}$.

Under the RS ansatz, the replicated preactivations of the anchor, positive view, and kth negative can be represented as

$$\begin{aligned}
h_i^a(c) &= \sqrt{\kappa}\,[r_o + (r_d - r_o)\delta_{ic}] + \sqrt{(1-\gamma)Q}\, g_{0,i} + \sqrt{(1-\gamma)(q-Q)}\, g_{a,i}, \\
h_i^{a+}(c) &= \sqrt{\kappa}\,[r_o + (r_d - r_o)\delta_{ic}] + \sqrt{(1-\gamma)Q}\, g_{0,i}^{+} + \sqrt{(1-\gamma)(q-Q)}\, g_{a,i}^{+}, \\
h_{i,k}^{a-}(c_k^-) &= \sqrt{\kappa}\left[r_o + (r_d - r_o)\delta_{ic_k^-}\right] + \sqrt{(1-\gamma)Q}\, g_{0,k,i}^{-} + \sqrt{(1-\gamma)(q-Q)}\, g_{a,k,i}^{-},
\end{aligned} \tag{58}$$

where the displayed Gaussian vectors are mutually independent and have i.i.d. standard-normal components. The subscript 0 denotes fields shared across replicas, whereas a denotes replica-specific fields. Let $\boldsymbol{v}^a$, $\boldsymbol{v}^{a+}$, and $\boldsymbol{v}_k^{a-}$ denote the corresponding activated representations.

Class-permutation symmetry makes all anchor classes equivalent. We therefore fix an arbitrary reference class c and sample

$$c_k^- \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\left(\{1,\dots,C\} \setminus \{c\}\right), \qquad k = 1,\dots,K. \tag{59}$$

For each replica, define

$$\Delta_k^a = \frac{(\boldsymbol{v}^a)^\top \boldsymbol{v}_k^{a-} - (\boldsymbol{v}^a)^\top \boldsymbol{v}^{a+}}{\tau}, \tag{60}$$

and let

$$\Delta_k = \frac{\boldsymbol{v}^\top \boldsymbol{v}_k^- - \boldsymbol{v}^\top \boldsymbol{v}^+}{\tau} \tag{61}$$

denote the corresponding single-replica quantity. Substituting Eq. (58) into Eq. (39) then gives

$$\begin{aligned}
\Psi_E|_{\mathrm{RS}} &= \alpha \log\left\{\mathbb{E}_{\{c_k^-\},\boldsymbol{g}_0,\boldsymbol{g}_0^+,\{\boldsymbol{g}_{0,k}^-\}} \prod_{a=1}^{n} \mathbb{E}_{\boldsymbol{g}_a,\boldsymbol{g}_a^+,\{\boldsymbol{g}_{a,k}^-\}} \exp\left[-\beta \log\left(1 + \sum_{k=1}^{K} e^{\Delta_k^a}\right)\right]\right\} \\
&= \alpha \log\left\{\mathbb{E}_{\{c_k^-\},\boldsymbol{g}_0,\boldsymbol{g}_0^+,\{\boldsymbol{g}_{0,k}^-\}} \prod_{a=1}^{n} \mathbb{E}_{\boldsymbol{g}_a,\boldsymbol{g}_a^+,\{\boldsymbol{g}_{a,k}^-\}} \left(1 + \sum_{k=1}^{K} e^{\Delta_k^a}\right)^{-\beta}\right\} \\
&= \alpha \log\left\{\mathbb{E}_{\{c_k^-\},\boldsymbol{g}_0,\boldsymbol{g}_0^+,\{\boldsymbol{g}_{0,k}^-\}} \prod_{a=1}^{n} \left[\frac{1}{\Gamma(\beta)} \int_0^\infty dt_a\, t_a^{\beta-1} e^{-t_a} \mathbb{E}_{\boldsymbol{g}_a,\boldsymbol{g}_a^+,\{\boldsymbol{g}_{a,k}^-\}} \exp\left(-t_a \sum_{k=1}^{K} e^{\Delta_k^a}\right)\right]\right\} \\
&= \alpha \log\left\{\mathbb{E}_{\{c_k^-\},\boldsymbol{g}_0,\boldsymbol{g}_0^+,\{\boldsymbol{g}_{0,k}^-\}} \left[\frac{1}{\Gamma(\beta)} \int_0^\infty dt\, t^{\beta-1} e^{-t} \mathbb{E}_{\boldsymbol{g},\boldsymbol{g}^+,\{\boldsymbol{g}_k^-\}} \exp\left(-t \sum_{k=1}^{K} e^{\Delta_k}\right)\right]^n\right\}.
\end{aligned} \tag{62}$$

The expectation over $\{c_k^-\}$ includes the factor $(C-1)^{-K}$ in Eq. (39); the factor C^{-1} is removed because all anchor classes are equivalent. The second equality uses

$$e^{-\beta \log(1+x)} = (1+x)^{-\beta}, \tag{63}$$

and the third applies

$$(1+x)^{-\beta} = \frac{1}{\Gamma(\beta)} \int_0^\infty dt\, t^{\beta-1} e^{-t} e^{-tx}. \tag{64}$$

Replica symmetry and the i.i.d. private Gaussian fields make the inner factor identical across replicas, yielding the final equality.

Dividing Eq. (62) by n and taking $n \to 0$ gives

$$\begin{aligned}\psi_E &\equiv \lim_{n\to 0} \frac{1}{n}\, \Psi_E|_{\mathrm{RS}} \\ &= \alpha\, \mathbb{E}_{\{c_k^-\},\, \boldsymbol{g}_0, \boldsymbol{g}_0^+, \{\boldsymbol{g}_{0,k}^-\}} \log \left[\frac{1}{\Gamma(\beta)} \int_0^\infty dt\, t^{\beta-1} e^{-t} \mathbb{E}_{\boldsymbol{g}, \boldsymbol{g}^+, \{\boldsymbol{g}_k^-\}} \exp\left(-t \sum_{k=1}^K e^{\Delta_k} \right) \right].\end{aligned} \tag{65}$$

The final step uses $\lim_{n\to 0} n^{-1} \log \mathbb{E}[X^n] = \mathbb{E}[\log X]$.

A.5 Saddle-Point Equations

Combining the replica-limit contributions from the overlap–conjugate, weight, and energetic terms yields

$$\begin{aligned}-\beta f_\beta^{\mathrm{RS}} = \underset{r_d, r_o, q, Q, \hat{r}_d, \hat{r}_o, \hat{q}, \hat{Q}}{\mathrm{Extr}} \Bigg\{ C \left[\hat{r}_d r_d + (C-1)\hat{r}_o r_o + \hat{q} q - \frac{\hat{Q} Q}{2} \right] \\ + \psi_S(\hat{r}_d, \hat{r}_o, \hat{q}, \hat{Q}) + \psi_E(r_d, r_o, q, Q) \Bigg\}.\end{aligned} \tag{66}$$

Stationarity of the variational functional in Eq. (66) with respect to each physical variable and its conjugate yields

$$\begin{cases} r_d = -\dfrac{\hat{r}_d}{\beta\lambda_w + 2\hat{q} - \hat{Q}}, & \hat{r}_d = -\dfrac{1}{C}\, \partial_{r_d} \psi_E, \\ r_o = -\dfrac{\hat{r}_o}{\beta\lambda_w + 2\hat{q} - \hat{Q}}, & \hat{r}_o = -\dfrac{1}{C(C-1)}\, \partial_{r_o} \psi_E, \\ q = \dfrac{1}{\beta\lambda_w + 2\hat{q} - \hat{Q}} + \dfrac{\hat{r}_d^2 + (C-1)\hat{r}_o^2 - \hat{Q}}{(\beta\lambda_w + 2\hat{q} - \hat{Q})^2}, & \hat{q} = -\dfrac{1}{C}\, \partial_q \psi_E, \\ Q = \dfrac{\hat{r}_d^2 + (C-1)\hat{r}_o^2 - \hat{Q}}{(\beta\lambda_w + 2\hat{q} - \hat{Q})^2}, & \hat{Q} = \dfrac{2}{C}\, \partial_Q \psi_E. \end{cases} \tag{67}$$

Here, each derivative of ψ_E is taken with respect to the indicated physical order parameter, with all other variables held fixed.

A.6 Energetic Response Terms

The four conjugate variables are obtained from derivatives of the energetic contribution in Eq. (67). For the fixed reference class c, conditional on the negative classes and shared Gaussian fields, define

$$\mathcal{U} = -\beta \log \left(1 + \sum_{k=1}^K e^{\Delta_k} \right), \tag{68}$$

and the normalized private-field average

$$\langle X \rangle = \mathbb{E}_{\boldsymbol{g}, \boldsymbol{g}^+, \{\boldsymbol{g}_k^-\}} \left[e^{\mathcal{U}} X \right] \left(\mathbb{E}_{\boldsymbol{g}, \boldsymbol{g}^+, \{\boldsymbol{g}_k^-\}} \left[e^{\mathcal{U}} \right] \right)^{-1}. \tag{69}$$

For compactness, collect the shared Gaussian fields in

$$\boldsymbol{G}_0 = \left(\boldsymbol{g}_0,\, \boldsymbol{g}_0^+,\, \{\boldsymbol{g}_{0,k}^-\}_{k=1}^K \right). \tag{70}$$

Differentiating with respect to r_d yields

$$\begin{aligned}\hat{r}_d &= -\frac{1}{C}\partial_{r_d}\psi_E \\ &= -\frac{\alpha\sqrt{\kappa}}{C}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left\langle \partial_{h_c}\mathcal{U} + \partial_{h_c^+}\mathcal{U} + \sum_{k=1}^{K}\partial_{h^-_{k,c_k^-}}\mathcal{U}\right\rangle \\ &= \frac{\alpha\beta\sqrt{\kappa}}{C}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left\langle \left(1+\sum_{\ell=1}^{K}e^{\Delta_\ell}\right)^{-1}\sum_{\ell=1}^{K}e^{\Delta_\ell}\left(\partial_{h_c}\Delta_\ell + \partial_{h_c^+}\Delta_\ell + \partial_{h^-_{\ell,c_\ell^-}}\Delta_\ell\right)\right\rangle.\end{aligned} \tag{71}$$

Likewise,

$$\begin{aligned}\hat{r}_o &= -\frac{1}{C(C-1)}\partial_{r_o}\psi_E \\ &= -\frac{\alpha\sqrt{\kappa}}{C(C-1)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left\langle \sum_{i\neq c}\left(\partial_{h_i}\mathcal{U} + \partial_{h_i^+}\mathcal{U}\right) + \sum_{k=1}^{K}\sum_{i\neq c_k^-}\partial_{h^-_{k,i}}\mathcal{U}\right\rangle \\ &= \frac{\alpha\beta\sqrt{\kappa}}{C(C-1)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left\langle \left(1+\sum_{\ell=1}^{K}e^{\Delta_\ell}\right)^{-1}\sum_{\ell=1}^{K}e^{\Delta_\ell}\left[\sum_{i\neq c}\left(\partial_{h_i}\Delta_\ell + \partial_{h_i^+}\Delta_\ell\right) + \sum_{i\neq c_\ell^-}\partial_{h^-_{\ell,i}}\Delta_\ell\right]\right\rangle.\end{aligned} \tag{72}$$

Differentiating with respect to q yields

$$\begin{aligned}\hat{q} &= -\frac{1}{C}\partial_q\psi_E \\ &= -\frac{\alpha\sqrt{1-\gamma}}{2C\sqrt{q-Q}}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left\langle \sum_{i=1}^{C}\left[g_i\,\partial_{h_i}\mathcal{U} + g_i^+\,\partial_{h_i^+}\mathcal{U} + \sum_{k=1}^{K}g^-_{k,i}\,\partial_{h^-_{k,i}}\mathcal{U}\right]\right\rangle \\ &= -\frac{\alpha(1-\gamma)}{2C}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left\langle \sum_{i=1}^{C}\left[\partial^2_{h_i}\mathcal{U} + (\partial_{h_i}\mathcal{U})^2 + \partial^2_{h_i^+}\mathcal{U} + (\partial_{h_i^+}\mathcal{U})^2 + \sum_{k=1}^{K}\left(\partial^2_{h^-_{k,i}}\mathcal{U} + (\partial_{h^-_{k,i}}\mathcal{U})^2\right)\right]\right\rangle.\end{aligned} \tag{73}$$

At fixed q, varying Q changes the shared and private Gaussian scales in opposite directions, giving

$$\begin{aligned}\hat{Q} &= \frac{2}{C}\partial_Q\psi_E \\ &= \frac{\alpha\sqrt{1-\gamma}}{C}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\frac{1}{\sqrt{Q}}\sum_{i=1}^{C}\left(g_{0,i}\langle\partial_{h_i}\mathcal{U}\rangle + g^+_{0,i}\langle\partial_{h_i^+}\mathcal{U}\rangle + \sum_{k=1}^{K}g^-_{0,k,i}\langle\partial_{h^-_{k,i}}\mathcal{U}\rangle\right)\right. \\ &\qquad\left. - \frac{1}{\sqrt{q-Q}}\left\langle\sum_{i=1}^{C}\left[g_i\,\partial_{h_i}\mathcal{U} + g_i^+\,\partial_{h_i^+}\mathcal{U} + \sum_{k=1}^{K}g^-_{k,i}\,\partial_{h^-_{k,i}}\mathcal{U}\right]\right\rangle\right] \\ &= -\frac{\alpha(1-\gamma)}{C}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\sum_{i=1}^{C}\left[\langle\partial_{h_i}\mathcal{U}\rangle^2 + \langle\partial_{h_i^+}\mathcal{U}\rangle^2 + \sum_{k=1}^{K}\langle\partial_{h^-_{k,i}}\mathcal{U}\rangle^2\right].\end{aligned} \tag{74}$$

The final equalities in Eq. (73) and Eq. (74) follow from Gaussian integration by parts, $\mathbb{E}[gF(g)] = \mathbb{E}[F'(g)]$. For any preactivation variable x, the required derivatives of $\mathcal{U}$ are

$$\begin{aligned}\partial_x\mathcal{U} &= -\beta\left(1+\sum_{k=1}^{K}e^{\Delta_k}\right)^{-1}\sum_{k=1}^{K}e^{\Delta_k}\partial_x\Delta_k, \\ \partial_x^2\mathcal{U} &= -\beta\left[\left(1+\sum_{k=1}^{K}e^{\Delta_k}\right)^{-1}\sum_{k=1}^{K}e^{\Delta_k}\left(\partial_x^2\Delta_k + (\partial_x\Delta_k)^2\right) - \left(\left(1+\sum_{k=1}^{K}e^{\Delta_k}\right)^{-1}\sum_{k=1}^{K}e^{\Delta_k}\partial_x\Delta_k\right)^2\right].\end{aligned} \tag{75}$$

The corresponding derivatives of the similarity differences are

$$\begin{aligned}
\partial_{h_i}\Delta_k &= \tau^{-1}\phi'(h_i)\left(v^-_{k,i} - v^+_i\right), & \partial^2_{h_i}\Delta_k &= \tau^{-1}\phi''(h_i)\left(v^-_{k,i} - v^+_i\right), \\
\partial_{h^+_i}\Delta_k &= -\tau^{-1}v_i\phi'(h^+_i), & \partial^2_{h^+_i}\Delta_k &= -\tau^{-1}v_i\phi''(h^+_i), \\
\partial_{h^-_{k,i}}\Delta_\ell &= \tau^{-1}v_i\phi'(h^-_{k,i})\delta_{k\ell}, & \partial^2_{h^-_{k,i}}\Delta_\ell &= \tau^{-1}v_i\phi''(h^-_{k,i})\delta_{k\ell}.
\end{aligned} \tag{76}$$

A.7 Zero-Temperature Reduction

We take the zero-temperature limit of the finite-temperature saddle by setting

$$q - Q = \frac{\chi}{\beta} + o(\beta^{-1}), \qquad \chi \in (0, \infty). \tag{77}$$

We assume that r_d, r_o, and q converge to finite limits, denoted by the same symbols. To eliminate the auxiliary conjugates, define $d = q - Q$ and $B = r_d^2 + (C-1)r_o^2$. The finite-temperature saddle equations then give

$$\begin{aligned}
\beta\lambda_w + 2\hat{q} - \hat{Q} = d^{-1}, \qquad \hat{r}_d = -\frac{r_d}{d}, \qquad \hat{r}_o = -\frac{r_o}{d}, \\
\hat{Q} = \frac{B - Q}{d^2}, \quad 2\hat{q} = d^{-1} - \beta\lambda_w + \hat{Q}.
\end{aligned} \tag{78}$$

Substituting these relations into the overlap and weight contributions yields

$$\psi_0 + \psi_S = \frac{C}{2}\left[\log(2\pi d) + 1 + \frac{Q - B}{d} - \beta\lambda_w q\right]. \tag{79}$$

After dropping additive constants independent of the saddle variables, the zero-temperature limit becomes

$$\begin{aligned}
&\frac{1}{\beta}\left[\frac{C}{2}\log(q - Q) + \frac{C}{2}\frac{Q - r_d^2 - (C-1)r_o^2}{q - Q} - \frac{\beta C\lambda_w}{2}q\right] \\
&= \frac{C}{2\beta}\log\frac{\chi}{\beta} + \frac{C}{2\chi}\left[q - r_d^2 - (C-1)r_o^2 - \frac{\chi}{\beta}\right] - \frac{C\lambda_w}{2}q + o(1) \\
&\xrightarrow{\beta\to\infty} \frac{C}{2\chi}\left[q - r_d^2 - (C-1)r_o^2\right] - \frac{C\lambda_w}{2}q.
\end{aligned} \tag{80}$$

Here we use $Q = q - \chi/\beta + o(\beta^{-1})$ and $\beta^{-1}\log(\chi/\beta) \to 0$.

For the energetic contribution, define $\chi_\beta = \beta(q - Q) \to \chi$. Let $\boldsymbol{u}_{c,\beta}$, $\boldsymbol{u}^+_{c,\beta}$, and $\boldsymbol{u}^-_{c^-_k,\beta}$ denote the finite-temperature field centers. They are defined by the limiting expressions below with q replaced by Q and the finite-temperature overlaps retained. The private fields then admit the exact representations

$$\boldsymbol{h} = \boldsymbol{u}_{c,\beta} + \sqrt{\frac{(1-\gamma)\chi_\beta}{\beta}}\,\boldsymbol{g}, \tag{81}$$

$$\boldsymbol{h}^+ = \boldsymbol{u}^+_{c,\beta} + \sqrt{\frac{(1-\gamma)\chi_\beta}{\beta}}\,\boldsymbol{g}^+, \tag{82}$$

$$\boldsymbol{h}^-_k = \boldsymbol{u}^-_{c^-_k,\beta} + \sqrt{\frac{(1-\gamma)\chi_\beta}{\beta}}\,\boldsymbol{g}^-_k. \tag{83}$$

The centers converge to

$$\boldsymbol{u}_c = \sqrt{\kappa}\left[r_o\mathbf{1}_C + (r_d - r_o)\boldsymbol{e}^{(c)}\right] + \sqrt{(1-\gamma)q}\,\boldsymbol{g}_0, \tag{84}$$

$$\boldsymbol{u}^+_c = \sqrt{\kappa}\left[r_o\mathbf{1}_C + (r_d - r_o)\boldsymbol{e}^{(c)}\right] + \sqrt{(1-\gamma)q}\,\boldsymbol{g}^+_0, \tag{85}$$

$$\boldsymbol{u}^-_{c^-_k} = \sqrt{\kappa}\left[r_o\mathbf{1}_C + (r_d - r_o)\boldsymbol{e}^{(c^-_k)}\right] + \sqrt{(1-\gamma)q}\,\boldsymbol{g}^-_{0,k}. \tag{86}$$

For fixed negative classes and shared Gaussian fields, the private-field contribution satisfies

$$
\begin{aligned}
&\frac{1}{\beta}\log \mathbb{E}_{\boldsymbol{g},\boldsymbol{g}^+,\{\boldsymbol{g}_k^-\}} \exp\left[-\beta \log\left(1+\sum_{k=1}^{K} e^{\Delta_k}\right)\right] \\
&= \frac{1}{\beta}\log\left[\left(\frac{\beta}{2\pi(1-\gamma)\chi_\beta}\right)^{C(K+2)/2} \int d\boldsymbol{h}\, d\boldsymbol{h}^+ \prod_{k=1}^{K} d\boldsymbol{h}_k^- \right. \\
&\qquad \left. \times \exp\left\{-\beta\left[\log\left(1+\sum_{k=1}^{K} e^{\Delta_k}\right) + \frac{\|\boldsymbol{h}-\boldsymbol{u}_{c,\beta}\|_2^2 + \|\boldsymbol{h}^+-\boldsymbol{u}_{c,\beta}^+\|_2^2 + \sum_{k=1}^{K}\|\boldsymbol{h}_k^- - \boldsymbol{u}_{c_k^-,\beta}^-\|_2^2}{2(1-\gamma)\chi_\beta}\right]\right\}\right] \\
&\xrightarrow{\beta\to\infty} -\min_{\boldsymbol{h},\boldsymbol{h}^+,\{\boldsymbol{h}_k^-\}} \left\{\log\left(1+\sum_{k=1}^{K} e^{\Delta_k}\right) + \frac{\|\boldsymbol{h}-\boldsymbol{u}_c\|_2^2 + \|\boldsymbol{h}^+-\boldsymbol{u}_c^+\|_2^2 + \sum_{k=1}^{K}\|\boldsymbol{h}_k^- - \boldsymbol{u}_{c_k^-}^-\|_2^2}{2(1-\gamma)\chi}\right\} \\
&\equiv -V_\chi^*.
\end{aligned}
\tag{87}
$$

The integral representation follows from an exact change of variables from the private Gaussian fields to the preactivations. The InfoNCE loss is nonnegative, and the quadratic penalty is positive for $\chi > 0$, making the effective objective coercive. The Laplace principle then yields the stated minimum. Let $\boldsymbol{h}^*$, $\boldsymbol{h}^{+,*}$, and $\{\boldsymbol{h}_k^{-,*}\}_{k=1}^K$ denote minimizers attaining V_χ^*.

Assuming that the $\beta \to \infty$ limit can be interchanged with the outer expectation, class-permutation symmetry allows us to fix the reference class c and write

$$
\begin{aligned}
\lim_{\beta\to\infty} \frac{\psi_E}{\beta} &= -\alpha\, \mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[V_\chi^*\right] \\
&\equiv -\mathcal{G}_\alpha(r_d, r_o, q, \chi).
\end{aligned}
\tag{88}
$$

Combining Eq. (80) and Eq. (88) yields the reduced zero-temperature free-energy density

$$
f_0(r_d, r_o, q, \chi) = \frac{C}{2\chi}\left[r_d^2 + (C-1)r_o^2 - q\right] + \frac{C\lambda_w}{2} q + \mathcal{G}_\alpha(r_d, r_o, q, \chi). \tag{89}
$$

At differentiable interior stationary points with $q > 0$ and $\chi > 0$, stationarity of Eq. (89) gives

$$
\begin{cases}
0 = \dfrac{\chi}{C}\,\partial_{r_d} f_0 = r_d + \dfrac{\chi}{C}\,\partial_{r_d}\mathcal{G}_\alpha, \\
0 = \dfrac{\chi}{C(C-1)}\,\partial_{r_o} f_0 = r_o + \dfrac{\chi}{C(C-1)}\,\partial_{r_o}\mathcal{G}_\alpha, \\
0 = -\dfrac{2}{C}\,\partial_q f_0 = \dfrac{1}{\chi} - \lambda_w - \dfrac{2}{C}\,\partial_q \mathcal{G}_\alpha, \\
0 = \dfrac{2\chi^2}{C}\,\partial_\chi f_0 = q - r_d^2 - (C-1)r_o^2 + \dfrac{2\chi^2}{C}\,\partial_\chi \mathcal{G}_\alpha.
\end{cases}
\tag{90}
$$

When the minimizing fields satisfy the required differentiability conditions and differentiation can be interchanged with the outer expectation, these equations reduce to

$$\begin{cases} r_d = \dfrac{\alpha\sqrt{\kappa}}{C(1-\gamma)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[(\boldsymbol{h}^* - \boldsymbol{u}_c)_c + (\boldsymbol{h}^{+,*} - \boldsymbol{u}_c^+)_c + \sum_{k=1}^{K}(\boldsymbol{h}_k^{-,*} - \boldsymbol{u}^-_{c_k^-})_{c_k^-}\right], \\ r_o = \dfrac{\alpha\sqrt{\kappa}}{C(C-1)(1-\gamma)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\sum_{i\neq c}\left[(\boldsymbol{h}^* - \boldsymbol{u}_c)_i + (\boldsymbol{h}^{+,*} - \boldsymbol{u}_c^+)_i\right] + \sum_{k=1}^{K}\sum_{i\neq c_k^-}(\boldsymbol{h}_k^{-,*} - \boldsymbol{u}^-_{c_k^-})_i\right], \\ q = r_d^2 + (C-1)r_o^2 + \dfrac{\alpha}{C(1-\gamma)}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\|\boldsymbol{h}^* - \boldsymbol{u}_c\|_2^2 + \|\boldsymbol{h}^{+,*} - \boldsymbol{u}_c^+\|_2^2 + \sum_{k=1}^{K}\|\boldsymbol{h}_k^{-,*} - \boldsymbol{u}^-_{c_k^-}\|_2^2\right], \\ 1 = \lambda_w\chi - \dfrac{\alpha}{C\sqrt{(1-\gamma)q}}\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\boldsymbol{g}_0^\top(\boldsymbol{h}^* - \boldsymbol{u}_c) + (\boldsymbol{g}_0^+)^\top(\boldsymbol{h}^{+,*} - \boldsymbol{u}_c^+) + \sum_{k=1}^{K}(\boldsymbol{g}_{0,k}^-)^\top(\boldsymbol{h}_k^{-,*} - \boldsymbol{u}^-_{c_k^-})\right]. \end{cases} \tag{91}$$

These equations form a closed self-consistent system for (r_d, r_o, q, χ).

A.8 Replica Theory for Unconditioned Negative Sampling

We finally consider unconditioned negative sampling, in which each negative class is drawn independently from the full class set,

$$c_k^- \overset{\text{i.i.d.}}{\sim} \text{Unif}\left(\{1,\dots,C\}\right), \qquad k = 1,\dots,K. \tag{92}$$

The anchor class remains uniformly distributed over $\{1,\dots,C\}$, and class-permutation symmetry again allows us to fix an arbitrary reference class c. The overlap structure and Gaussian fields are unchanged; only the negative-class sampling law differs from the class-excluded case. In particular, a negative may satisfy $c_k^- = c$. It then has the same class-dependent mean as the anchor, while its replica-shared and private Gaussian fields remain independent of those of the anchor.

Replacing the class-excluded sampling law in Eq. (62) by Eq. (92) gives

$$\begin{aligned} \Psi_E^{\text{unc}}|_{\text{RS}} &= \alpha \log\left\{\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[\frac{1}{\Gamma(\beta)}\int_0^\infty dt\, t^{\beta-1}e^{-t}\mathbb{E}_{\boldsymbol{g},\boldsymbol{g}^+,\{\boldsymbol{g}_k^-\}}\exp\left(-t\sum_{k=1}^{K}e^{\Delta_k}\right)\right]^n\right\} \\ &= n\alpha\,\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\log\left[\frac{1}{\Gamma(\beta)}\int_0^\infty dt\, t^{\beta-1}e^{-t}\mathbb{E}_{\boldsymbol{g},\boldsymbol{g}^+,\{\boldsymbol{g}_k^-\}}\exp\left(-t\sum_{k=1}^{K}e^{\Delta_k}\right)\right] + O(n^2). \end{aligned} \tag{93}$$

Here, $\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}$ uses the distribution in Eq. (92), so the negative-class average corresponds to the factor C^{-K} rather than $(C-1)^{-K}$ in the class-excluded case. The second equality follows from $\log\mathbb{E}[X^n] = n\,\mathbb{E}[\log X] + O(n^2)$.

Dividing by n and taking $n \to 0$ yields

$$\begin{aligned} \psi_E^{\text{unc}} &\equiv \lim_{n\to 0}\frac{1}{n}\,\Psi_E^{\text{unc}}|_{\text{RS}} \\ &= \alpha\,\mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\log\left[\frac{1}{\Gamma(\beta)}\int_0^\infty dt\, t^{\beta-1}e^{-t}\mathbb{E}_{\boldsymbol{g},\boldsymbol{g}^+,\{\boldsymbol{g}_k^-\}}\exp\left(-t\sum_{k=1}^{K}e^{\Delta_k}\right)\right]. \end{aligned} \tag{94}$$

Because the overlap–conjugate and weight contributions are unchanged, the finite-temperature free energy becomes

$$\begin{aligned} -\beta f_\beta^{\text{RS,unc}} = \underset{r_d,r_o,q,Q,\hat{r}_d,\hat{r}_o,\hat{q},\hat{Q}}{\text{Extr}}\Bigg\{ & C\left[\hat{r}_d r_d + (C-1)\hat{r}_o r_o + \hat{q}q - \frac{\hat{Q}Q}{2}\right] \\ & + \psi_S(\hat{r}_d,\hat{r}_o,\hat{q},\hat{Q}) + \psi_E^{\text{unc}}(r_d,r_o,q,Q)\Bigg\}. \end{aligned} \tag{95}$$

The finite-temperature saddle-point equations therefore retain the form of Eq. (67), with ψ_E replaced by ψ_E^{unc}.

Under the scaling $q - Q = \chi/\beta + o(\beta^{-1})$, the private-field integral follows the same zero-temperature reduction as in Sec. A.7. Thus,

$$\begin{aligned}\lim_{\beta\to\infty} \frac{\psi_E^{\mathrm{unc}}}{\beta} &= -\alpha\, \mathbb{E}_{\{c_k^-\},\boldsymbol{G}_0}\left[V_\chi^*\right] \\ &\equiv -\mathcal{G}_\alpha^{\mathrm{unc}}(r_d, r_o, q, \chi),\end{aligned} \tag{96}$$

where the expectation uses $c_k^- \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{1,\dots,C\})$. For fixed class labels and shared Gaussian fields, the minimization defining V_χ^* is unchanged. When $c_k^- = c$, the corresponding negative has the same class-dependent mean as the anchor, while its replica-shared and private Gaussian fields remain independent of those of the anchor.

The zero-temperature free energy is therefore

$$f_0^{\mathrm{unc}}(r_d, r_o, q, \chi) = \frac{C}{2\chi}\left[r_d^2 + (C-1)r_o^2 - q\right] + \frac{C\lambda_w}{2}q + \mathcal{G}_\alpha^{\mathrm{unc}}(r_d, r_o, q, \chi). \tag{97}$$

The zero-temperature saddle-point equations retain the form of Eq. (90), with $\mathcal{G}_\alpha$ replaced by $\mathcal{G}_\alpha^{\mathrm{unc}}$.

A.9 Population Linear-Probe Generalization

Consider an independent test sample from class c,

$$\boldsymbol{x} = \sqrt{\gamma}\,\boldsymbol{\xi}_c + \sqrt{1-\gamma}\,\boldsymbol{z}, \qquad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_N). \tag{98}$$

Under the reduced overlap ansatz, the corresponding encoder preactivation satisfies

$$\boldsymbol{h} \mid c \overset{d}{=} \sqrt{\kappa}\left[r_o \mathbf{1}_C + (r_d - r_o)\boldsymbol{e}^{(c)}\right] + \sqrt{(1-\gamma)q}\,\boldsymbol{Z}, \qquad \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_C). \tag{99}$$

Hence,

$$h_c = \sqrt{\kappa}\,r_d + \sqrt{(1-\gamma)q}\,Z_c, \qquad h_j = \sqrt{\kappa}\,r_o + \sqrt{(1-\gamma)q}\,Z_j, \quad j \neq c. \tag{100}$$

Let $\boldsymbol{v} = \phi(\boldsymbol{h})$ coordinatewise. We consider an unregularized population linear probe with logits $\boldsymbol{A}\boldsymbol{v}$, where $\boldsymbol{A} \in \mathbb{R}^{C\times C}$ minimizes the population cross-entropy

$$\mathcal{L}_{\mathrm{probe}}(\boldsymbol{A}) = \mathbb{E}_{c,\boldsymbol{h}|c}\left[\log \sum_{j=1}^{C} \exp((\boldsymbol{A}\boldsymbol{v})_j) - (\boldsymbol{A}\boldsymbol{v})_c\right]. \tag{101}$$

We assume that $q > 0$, that ϕ is strictly increasing with finite second moments under the Gaussian field law, and that the population minimum is attained. The objective is invariant under simultaneous permutations of class labels and representation coordinates and is convex in $\boldsymbol{A}$. Averaging any minimizer over this symmetry group therefore yields another minimizer of the form $\boldsymbol{A} = a_{\mathrm{p}}\boldsymbol{I}_C + b_{\mathrm{p}}\mathbf{1}_C\mathbf{1}_C^\top$. The second term adds the same quantity to every class score and hence cancels in the softmax. If a class-specific bias is included, the same symmetrization reduces it to a common offset.

First consider $r_d > r_o$. At $a_{\mathrm{p}} = 0$,

$$\begin{aligned}\partial_{a_{\mathrm{p}}}\mathcal{L}_{\mathrm{probe}}\big|_{a_{\mathrm{p}}=0} &= \mathbb{E}_{c,\boldsymbol{h}|c}\left[-v_c + \frac{1}{C}\sum_{i=1}^{C} v_i\right] \\ &= -\frac{C-1}{C}\Bigg\{\mathbb{E}_Z\left[\phi\Big(\sqrt{\kappa}\,r_d + \sqrt{(1-\gamma)q}\,Z\Big)\right] \\ &\qquad - \mathbb{E}_Z\left[\phi\Big(\sqrt{\kappa}\,r_o + \sqrt{(1-\gamma)q}\,Z\Big)\right]\Bigg\} < 0.\end{aligned} \tag{102}$$

Convexity therefore implies $a_{\mathrm{p}}^* > 0$. Since ϕ is strictly increasing, the resulting decision rule is

$$\hat{c} = \arg\max_i (\boldsymbol{A}\boldsymbol{v})_i = \arg\max_i v_i = \arg\max_i h_i. \tag{103}$$

For $r_d < r_o$, the derivative has the opposite sign, so $a_{\mathrm{p}}^* < 0$ and $\hat{c} = \arg\min_i h_i$.

Define the signed specialization signal-to-noise ratio

$$\Delta_{\mathrm{LP}} = \sqrt{\frac{\kappa}{(1-\gamma)q}}\,(r_d - r_o). \tag{104}$$

For $\Delta_{\mathrm{LP}} > 0$, conditioning on $Z_c = z$ gives

$$\begin{aligned} \hat{c} = c \iff Z_j < z + \Delta_{\mathrm{LP}} \quad \text{for all } j \neq c, \\ \Pr(\hat{c} = c \mid Z_c = z) = \prod_{j \neq c} \Pr(Z_j < z + \Delta_{\mathrm{LP}}) \\ = \Phi(z + \Delta_{\mathrm{LP}})^{C-1}, \end{aligned} \tag{105}$$

where Φ denotes the standard normal cumulative distribution function. For $\Delta_{\mathrm{LP}} < 0$, the argmin rule and Gaussian symmetry yield the same population accuracy after replacing Δ_{LP} by $|\Delta_{\mathrm{LP}}|$. The population classification error is therefore

$$\epsilon_{\mathrm{gen}} = 1 - \int Dz\, \Phi(z + |\Delta_{\mathrm{LP}}|)^{C-1}, \qquad Dz = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz. \tag{106}$$

At $\Delta_{\mathrm{LP}} = 0$, the class-conditional representation distributions coincide, giving the chance-level error $1 - 1/C$. When $q = 0$, the representation is constant and the error is again $1 - 1/C$.

For the linear activation $\phi(h) = h$, the class-explained variance fraction satisfies

$$\begin{aligned} \eta_{\mathrm{cls}} &= \frac{(C-1)\kappa(r_d - r_o)^2}{C^2(1-\gamma)q + (C-1)\kappa(r_d - r_o)^2} \\ &= \frac{(C-1)\Delta_{\mathrm{LP}}^2}{C^2 + (C-1)\Delta_{\mathrm{LP}}^2}. \end{aligned} \tag{107}$$

Solving for the signal-to-noise magnitude gives

$$|\Delta_{\mathrm{LP}}| = \frac{C}{\sqrt{C-1}} \left(\frac{\eta_{\mathrm{cls}}}{1 - \eta_{\mathrm{cls}}}\right)^{1/2}. \tag{108}$$

Substituting this relation into Eq. (106) yields

$$\epsilon_{\mathrm{gen}}(\eta_{\mathrm{cls}}) = 1 - \int Dz\, \Phi\left[z + \frac{C}{\sqrt{C-1}} \left(\frac{\eta_{\mathrm{cls}}}{1 - \eta_{\mathrm{cls}}}\right)^{1/2}\right]^{C-1}. \tag{109}$$

For strictly increasing nonlinear activations satisfying the assumptions above, including smooth ReLU, Eq. (106) remains valid because the activation preserves coordinate ordering. However, η_{cls} also depends on the nonlinear operating point and is therefore not determined by Δ_{LP} alone. These results characterize the population-optimal linear probe under the reduced ansatz. They do not describe a probe estimated from finitely many labeled samples.

B Class-Informative Geometry and Spectral Effective Rank

This section relates the specialization amplitude $\Delta r = r_d - r_o$ to representation geometry under the reduced overlap ansatz. Near the unspecialized reference state, the class-mean contrast scales linearly with Δr, the class-explained variance fraction scales as $O((\Delta r)^2)$, and centered effective-rank deficits scale as $O((\Delta r)^4)$. These scalings assume that the within-class variance has a finite, strictly positive limit.

B.1 Class-Mean Separation and Explained Variance

Using the class-conditional field distribution in Eq. (99), define the matched–unmatched representation contrast

$$\Delta m = \mathbb{E}_Z \left[\phi\Big(\sqrt{\kappa}\, r_d + \sqrt{(1-\gamma)q}\, Z\Big) - \phi\Big(\sqrt{\kappa}\, r_o + \sqrt{(1-\gamma)q}\, Z\Big)\right], \qquad Z \sim \mathcal{N}(0,1). \tag{110}$$

By permutation symmetry, the class means defined in Section 2 satisfy

$$\boldsymbol{m}_c - \bar{\boldsymbol{m}} = \Delta m \left(\boldsymbol{e}^{(c)} - \frac{1}{C} \mathbf{1}_C \right). \tag{111}$$

Hence, $\|\boldsymbol{m}_c - \boldsymbol{m}_{c'}\|_2 = \sqrt{2}\,|\Delta m|$ for $c \neq c'$. The between-class covariance is therefore

$$\begin{aligned} \boldsymbol{\Sigma}_B &= \frac{1}{C} \sum_{c=1}^{C} (\boldsymbol{m}_c - \bar{\boldsymbol{m}})(\boldsymbol{m}_c - \bar{\boldsymbol{m}})^\top \\ &= \frac{(\Delta m)^2}{C} \left(\boldsymbol{I}_C - \frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top \right), \end{aligned} \tag{112}$$

with trace

$$\operatorname{Tr}(\boldsymbol{\Sigma}_B) = \frac{C-1}{C} (\Delta m)^2. \tag{113}$$

Along a continuous path approaching $(\bar{r}, \Delta r, q) = (\bar{r}_0, 0, q_0)$, differentiation under the Gaussian expectation for the linear and smooth-ReLU activations considered here gives

$$\begin{aligned} \Delta m &= \sqrt{\kappa}\, \mathbb{E}_Z \left[\phi' \Big(\sqrt{\kappa}\, \bar{r}_0 + \sqrt{(1-\gamma) q_0}\, Z \Big) \right] \Delta r + o(\Delta r) \\ &= O(\Delta r). \end{aligned} \tag{114}$$

This expansion is taken about the exactly symmetric state $\Delta r = 0$. Numerical states with small nonzero overlap contrast lie in its neighborhood. It follows that

$$\operatorname{Tr}(\boldsymbol{\Sigma}_B) = O\big((\Delta r)^2\big). \tag{115}$$

If $\operatorname{Tr}(\boldsymbol{\Sigma}_W)$ converges to a finite, strictly positive value, the class-explained variance fraction satisfies

$$\eta_{\text{cls}} = \frac{\operatorname{Tr}(\boldsymbol{\Sigma}_B)}{\operatorname{Tr}(\boldsymbol{\Sigma})} = \frac{\operatorname{Tr}(\boldsymbol{\Sigma}_B)}{\operatorname{Tr}(\boldsymbol{\Sigma}_B) + \operatorname{Tr}(\boldsymbol{\Sigma}_W)} = O\big((\Delta r)^2\big). \tag{116}$$

For both activations, the coefficient in Eq. (114) is positive. Consequently, Δm and η_{cls} have nonzero linear and quadratic leading terms, respectively.

B.2 Effective-Rank Response

Under the coordinate-diagonal overlap ansatz, the Gaussian preactivation coordinates are conditionally independent given the class. Their activated coordinates are therefore conditionally independent as well. Define the matched and unmatched conditional variances as

$$\begin{aligned} \nu_d &= \operatorname{Var}_Z \left[\phi \Big(\sqrt{\kappa}\, r_d + \sqrt{(1-\gamma) q}\, Z \Big) \right], \\ \nu_o &= \operatorname{Var}_Z \left[\phi \Big(\sqrt{\kappa}\, r_o + \sqrt{(1-\gamma) q}\, Z \Big) \right]. \end{aligned} \tag{117}$$

Averaging over the balanced classes gives

$$\boldsymbol{\Sigma}_W = \bar{\nu}\, \boldsymbol{I}_C, \qquad \bar{\nu} = \frac{\nu_d + (C-1)\nu_o}{C} > 0. \tag{118}$$

Combining this expression with Eq. (112) yields the centered representation covariance

$$\boldsymbol{\Sigma} = \bar{\nu}\, \boldsymbol{I}_C + \frac{(\Delta m)^2}{C} \left(\boldsymbol{I}_C - \frac{1}{C} \mathbf{1}_C \mathbf{1}_C^\top \right). \tag{119}$$

The eigenvalue along $\mathbf{1}_C$ is $\bar{\nu}$, whereas the eigenvalue on its $(C-1)$-dimensional orthogonal complement is $\bar{\nu} + (\Delta m)^2 / C$.

Define the participation ratio as

$$r_{\text{PR}} = \frac{[\operatorname{Tr}(\boldsymbol{\Sigma})]^2}{\operatorname{Tr}(\boldsymbol{\Sigma}^2)} \tag{120}$$

and its deficit from the isotropic value as

$$D_{\mathrm{PR}} = C - r_{\mathrm{PR}}. \tag{121}$$

From the two-eigenvalue spectrum,

$$\eta_{\mathrm{cls}} = \frac{\frac{C-1}{C}(\Delta m)^2}{C\bar{\nu} + \frac{C-1}{C}(\Delta m)^2}, \tag{122}$$

and

$$r_{\mathrm{PR}} = \frac{\left[C\bar{\nu} + \frac{C-1}{C}(\Delta m)^2\right]^2}{\bar{\nu}^2 + (C-1)\left(\bar{\nu} + \frac{(\Delta m)^2}{C}\right)^2}. \tag{123}$$

To express the spectral response in terms of class organization, normalize the eigenvalues by $\mathrm{Tr}(\mathbf{\Sigma})$. The normalized weights along $\mathbf{1}_C$ and its orthogonal complement are

$$w_{\parallel} = \frac{1-\eta_{\mathrm{cls}}}{C}, \qquad w_{\perp} = \frac{1+\eta_{\mathrm{cls}}/(C-1)}{C}, \tag{124}$$

where $w_{\perp}$ has multiplicity $C-1$. It follows that

$$r_{\mathrm{PR}} = \frac{1}{w_{\parallel}^2 + (C-1)w_{\perp}^2} = \frac{C(C-1)}{C-1+\eta_{\mathrm{cls}}^2}, \tag{125}$$

and hence

$$D_{\mathrm{PR}} = \frac{C\eta_{\mathrm{cls}}^2}{C-1+\eta_{\mathrm{cls}}^2}. \tag{126}$$

For weak class organization,

$$D_{\mathrm{PR}} = \frac{C}{C-1}\eta_{\mathrm{cls}}^2 + O\big(\eta_{\mathrm{cls}}^4\big). \tag{127}$$

Together with Eq. (116), this gives

$$D_{\mathrm{PR}} = O\big((\Delta r)^4\big). \tag{128}$$

Thus, the participation-ratio deficit responds quadratically to weak class organization.

We next consider the population analogue of RankMe for centered representations, without numerical stabilization terms (Roy & Vetterli, 2007; Garrido et al., 2023). For centered covariance eigenvalues $\{\lambda_j\}_{j=1}^{C}$, define the normalized singular-value weights as

$$p_j^{\mathrm{RM}} = \frac{\sqrt{\lambda_j}}{\sum_{\ell=1}^{C}\sqrt{\lambda_\ell}}. \tag{129}$$

The corresponding effective rank is

$$r_{\mathrm{RM}} = \exp\left(-\sum_{j=1}^{C} p_j^{\mathrm{RM}} \log p_j^{\mathrm{RM}}\right). \tag{130}$$

For the two-eigenvalue spectrum, the singular-value weights have the expansions

$$\begin{aligned} p_{\parallel}^{\mathrm{RM}} &= \frac{\sqrt{w_{\parallel}}}{\sqrt{w_{\parallel}} + (C-1)\sqrt{w_{\perp}}} = \frac{1}{C} - \frac{\eta_{\mathrm{cls}}}{2C} + O(\eta_{\mathrm{cls}}^2), \\ p_{\perp}^{\mathrm{RM}} &= \frac{\sqrt{w_{\perp}}}{\sqrt{w_{\parallel}} + (C-1)\sqrt{w_{\perp}}} = \frac{1}{C} + \frac{\eta_{\mathrm{cls}}}{2C(C-1)} + O(\eta_{\mathrm{cls}}^2). \end{aligned} \tag{131}$$

Writing $\delta p_j^{\mathrm{RM}} = p_j^{\mathrm{RM}} - 1/C$ and expanding the entropy about the uniform spectrum gives

$$\begin{aligned} \log C - \log r_{\mathrm{RM}} &= \frac{C}{2}\sum_{j=1}^{C}(\delta p_j^{\mathrm{RM}})^2 + O\big(\|\delta \boldsymbol{p}^{\mathrm{RM}}\|_2^3\big) \\ &= \frac{\eta_{\mathrm{cls}}^2}{8(C-1)} + O\big(\eta_{\mathrm{cls}}^3\big). \end{aligned} \tag{132}$$

Equivalently,

$$C - r_{\mathrm{RM}} = \frac{C}{8(C-1)}\eta_{\mathrm{cls}}^2 + O\big(\eta_{\mathrm{cls}}^3\big)\,. \tag{133}$$

RankMe therefore has the same leading-order response,

$$C - r_{\mathrm{RM}} = O\big(\eta_{\mathrm{cls}}^2\big) = O\big((\Delta r)^4\big)\,. \tag{134}$$

More generally, let $\mathcal{R}$ be a twice continuously differentiable, permutation-invariant function of normalized spectral weights in a neighborhood of the uniform spectrum. Write

$$\boldsymbol{p} = \frac{1}{C}\mathbf{1}_C + \delta\boldsymbol{p}, \qquad \mathbf{1}_C^\top \delta\boldsymbol{p} = 0. \tag{135}$$

Permutation symmetry at the uniform spectrum implies

$$\nabla\mathcal{R}\left(\frac{1}{C}\mathbf{1}_C\right) \propto \mathbf{1}_C. \tag{136}$$

Therefore,

$$\begin{aligned}\mathcal{R}(\boldsymbol{p}) - \mathcal{R}\left(\frac{1}{C}\mathbf{1}_C\right) &= \nabla\mathcal{R}\left(\frac{1}{C}\mathbf{1}_C\right)^\top \delta\boldsymbol{p} + O\big(\|\delta\boldsymbol{p}\|_2^2\big) \\ &= O\big(\|\delta\boldsymbol{p}\|_2^2\big)\,.\end{aligned} \tag{137}$$

For the spectrum above, both the normalized eigenvalue weights and normalized singular-value weights satisfy $\|\delta\boldsymbol{p}\|_2 = O(\eta_{\mathrm{cls}})$. Hence,

$$\mathcal{R}(\boldsymbol{p}) - \mathcal{R}\left(\frac{1}{C}\mathbf{1}_C\right) = O\big(\eta_{\mathrm{cls}}^2\big) = O\big((\Delta r)^4\big)\,. \tag{138}$$

Together, normalization and permutation symmetry eliminate the first-order spectral response at isotropy. If the quadratic coefficient also vanishes, the leading change can occur at still higher order.

B.3 Centering and Scope of the Spectral Relation

The preceding relations are derived for the centered representation covariance. Under balanced class-permutation symmetry, $\mathbb{E}[\boldsymbol{v}] \propto \mathbf{1}_C$, whereas the uncentered second moment decomposes as

$$\mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = \boldsymbol{\Sigma} + \mathbb{E}[\boldsymbol{v}]\mathbb{E}[\boldsymbol{v}]^\top. \tag{139}$$

The rank-one mean term acts only along the class-symmetric direction and can alter an uncentered effective-rank statistic even when $\Delta r = 0$ and $\eta_{\mathrm{cls}} = 0$. Centering removes this class-independent common mode.

The exact relation in Eq. (126) is specific to the covariance structure in Eq. (119). Within this structure, the participation-ratio deficit is a monotone function of η_{cls}, and its quadratic response near $\eta_{\mathrm{cls}} = 0$ reflects reduced local sensitivity rather than a distinct onset of class organization. Beyond this covariance structure, class-neutral anisotropy can alter spectral effective ranks independently of class organization. For $\bar{\nu} > 0$, the centered covariance has full algebraic rank C, while specialization changes its effective rank through spectral anisotropy.

C Experimental Validation and Extensions

This section describes the numerical protocols and evaluates the theoretical predictions beyond the canonical setting. We first present the finite-N training protocol and convergence diagnostics (Sec. C.1), followed by a comparison between trained networks and the reduced replica theory (Sec. C.2). We then examine configuration dependence (Sec. C.3), downstream generalization (Sec. C.4), and unconditioned negative sampling (Sec. C.5). Finally, experiments on CIFAR-10 (Krizhevsky, 2009) compare class organization with linear-probe performance and spectral effective rank in a SimCLR pipeline (Chen et al., 2020) (Sec. C.6).

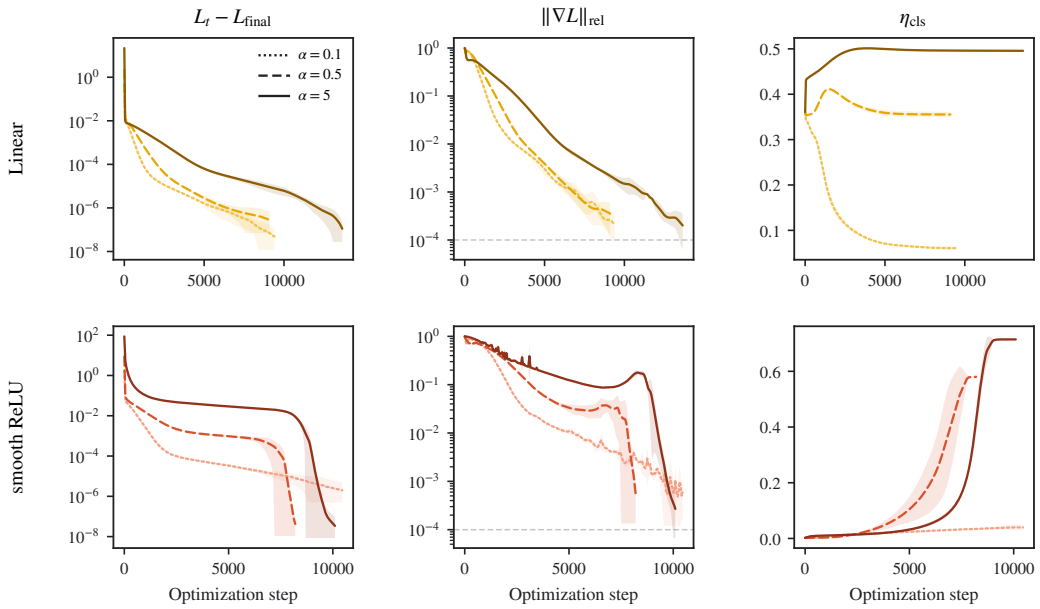

Figure 4: Selected optimization trajectories for linear (top) and smooth-ReLU (bottom) encoders at $\alpha \in \{0.1, 0.5, 5.0\}$. The columns show the objective-density difference $[\mathcal{L}(\boldsymbol{W}_t) - \mathcal{L}(\boldsymbol{W}_{\text{final}})]/N$, relative gradient ϵ_{grad}, and population class-explained variance fraction η_{cls}. The dashed line marks 10^{-5} for reference, while these diagnostic runs use a stopping threshold of 10^{-4}. Curves show means and shading shows one sample standard deviation across three selected training-data realizations evaluated at shared checkpoints.

C.1 Finite-N Simulation Protocol and Convergence Diagnostics

The finite-N simulations in Fig. 2 use $N = 10{,}000$, $P = \alpha N$, $K = 5$, $\kappa = 15$, $\tau = 0.1$, and $\lambda_w = 0.01$, with $M = C$ and class-excluded negatives. We consider a linear encoder with $C = 10$ and smooth-ReLU encoders with $C \in \{2, 10\}$, using $\phi(t) = \frac{1}{2}(t + \sqrt{t^2 + 0.01})$. The linear scan covers $\alpha \in [0.03, 0.30]$, while both smooth-ReLU scans cover $\alpha \in [0.01, 0.30]$, with spacing 0.01. The entries of $\boldsymbol{\Xi}$ are independent standard Gaussian variables, so $N^{-1}\boldsymbol{\Xi}\boldsymbol{\Xi}^\top \to \boldsymbol{I}_C$ as $N \to \infty$ at fixed C. For each anchor, the positive is drawn independently from the same class, while each negative is sampled uniformly from a uniformly chosen different class. Negative samples are drawn from the shared anchor pool and held fixed throughout training. This preserves the theoretical class-sampling rule while introducing sample sharing across contrastive groups.

At each α, we train 50 networks from independent initializations whose entries are drawn from $\mathcal{N}(0, 1)$. The class directions, training set, and negative assignments are shared across restarts. All computations use `float64`. Optimization uses at most $30{,}000$ full-batch Adam steps (Kingma & Ba, 2015) with learning rate 0.03, $(\beta_1, \beta_2) = (0.9, 0.999)$, and numerical stability constant 10^{-8}. This is followed by at most 100 L-BFGS steps (Liu & Nocedal, 1989) with history size 8, root-mean-square update cap 0.25, and Armijo backtracking (Armijo, 1966) using sufficient-decrease coefficient 10^{-4} and at most 12 step halvings. The implementation minimizes $\mathcal{L}(\boldsymbol{W})/N$, where $\mathcal{L}$ is the summed regularized objective in Eq. (1). We assess stationarity through the relative gradient

$$\epsilon_{\text{grad}} = \frac{\|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W})\|_F}{\max\{\|\nabla_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) - \lambda_w \boldsymbol{W}\|_F + \|\lambda_w \boldsymbol{W}\|_F, 10^{-20}\}}. \tag{140}$$

Training terminates early when $\epsilon_{\text{grad}} \le 10^{-5}$, and only terminal states satisfying this criterion are retained. This leaves $1{,}381$ of $1{,}400$ linear $C = 10$ runs, $1{,}400$ of $1{,}500$ smooth-ReLU $C = 10$ runs, and all $1{,}500$ smooth-ReLU $C = 2$ runs.

For each retained encoder, we evaluate population representation statistics under the corresponding Gaussian-mixture input distribution, using analytic moments for the linear activation and Gauss–Hermite quadrature for smooth ReLU. Each point in Fig. 2 therefore represents a trained encoder

Algorithm 1 Solving the zero-temperature saddle-point equations at fixed α

Require: Initial state $\boldsymbol{\theta}^{(0)} = (r_d^{(0)}, r_o^{(0)}, q^{(0)}, \chi^{(0)})$, fixed Monte Carlo samples $\mathcal{S}$, independent validation samples $\mathcal{S}_{\rm val}$, and numerical tolerances
Ensure: A terminal candidate and its numerical validation status
Map $\boldsymbol{\theta}^{(0)}$ to unconstrained coordinates $\boldsymbol{z}^{(0)}$ while enforcing the physical constraints
function EVALUATE($\boldsymbol{z}, \mathcal{S}$)
    Recover $\boldsymbol{\theta} = (r_d, r_o, q, \chi)$ from $\boldsymbol{z}$
    for each Gaussian-field and class-label sample in $\mathcal{S}$ **do**
        Solve the effective single-group minimization by adaptive multistart L-BFGS
        Accumulate the quantities entering the saddle-point equations in Eq. (90)
    end for
    Average over $\mathcal{S}$ and return the four saddle residuals
end function
Apply damped fixed-point iterations of Eq. (90) to obtain a warm start
Refine the solution by minimizing the squared residual norm on the fixed samples $\mathcal{S}$ with Levenberg–Marquardt
Retain the visited state with the smallest residual norm on $\mathcal{S}$
Re-evaluate the retained state on $\mathcal{S}_{\rm val}$ and check the prescribed convergence and validation tolerances
return (r_d, r_o, q, χ) and its numerical validation status

satisfying the numerical stationarity criterion, with its statistics computed from population moments rather than a finite test set. Separate random-number streams use seed 40 for the class directions and training data, 4000 for initialization, and 1040 for validation data. Restart-specific initialization keys are obtained by folding the restart index into the base initialization key.

The diagnostic runs in Fig. 4 use $N = 5{,}000$, $\alpha \in \{0.1, 0.5, 5.0\}$, at most $20{,}000$ Adam steps and 200 L-BFGS steps, and a stopping threshold of 10^{-4}. For each encoder, we select the three training-data seeds with the smallest late-stage drift in the log relative gradient and $\eta_{\rm cls}$ among those available at all three α values. The selected seeds are $\{41, 42, 48\}$ for the linear encoder and $\{41, 45, 46\}$ for smooth ReLU. The class-direction and initialization seeds are fixed to 40, and each validation seed is obtained by adding 1000 to the corresponding training-data seed. Checkpoints are stored every 50 optimization steps, including the initial and final states.

At each checkpoint, we evaluate $[\mathcal{L}(\boldsymbol{W}_t) - \mathcal{L}(\boldsymbol{W}_{\rm final})]/N$, $\epsilon_{\rm grad}$, and population $\eta_{\rm cls}$, using 64-point Gauss–Hermite quadrature for smooth ReLU. For logarithmic plotting, negative objective differences are clipped to zero and a positive plotting floor is applied. Means and sample standard deviations are computed across the three selected realizations at shared checkpoints, without interpolation or smoothing. These trajectories illustrate the approach to numerical stationarity together with the evolution of class organization.

C.2 Finite-N Validation of the Reduced Replica Theory

The comparison in Fig. 5 uses simulations distinct from those in Fig. 2, with $N = 5{,}000$, $M = C = 10$, $K = 5$, $\kappa = 15$, $\lambda_w = 0.01$, and $\tau = 0.1$. Smooth ReLU uses $\phi(t) = \frac{1}{2}(t + \sqrt{t^2 + 0.01})$. Both encoders are evaluated at 17 sample ratios,

$$\alpha \in \{0.1, 0.15, 0.18, 0.2, 0.22, 0.25, 0.28, 0.3, 0.35, 0.4, 0.5, 0.7, 1, 1.5, 2, 3, 5\}.$$

For each α, we use ten training-data realizations with seeds 41–50, yielding 170 retained results per encoder. The class directions are fixed across runs. Training data vary across seeds and are shared or nested across α within each seed. Negative sampling follows Sec. C.1, with assignments held fixed throughout training.

Training uses at most $20{,}000$ full-batch Adam steps followed by 200 L-BFGS steps. The remaining optimizer settings follow Sec. C.1, with a stopping criterion of $\epsilon_{\rm grad} \le 10^{-4}$. For the linear encoder, we compare one random initialization with one strongly specialized initialization using an independent validation objective. Smooth ReLU uses one random initialization. Simulation $\eta_{\rm cls}$ is estimated from $1{,}000$ balanced held-out samples using centered empirical covariances. Within each data seed, the same held-out sample set is used across α.

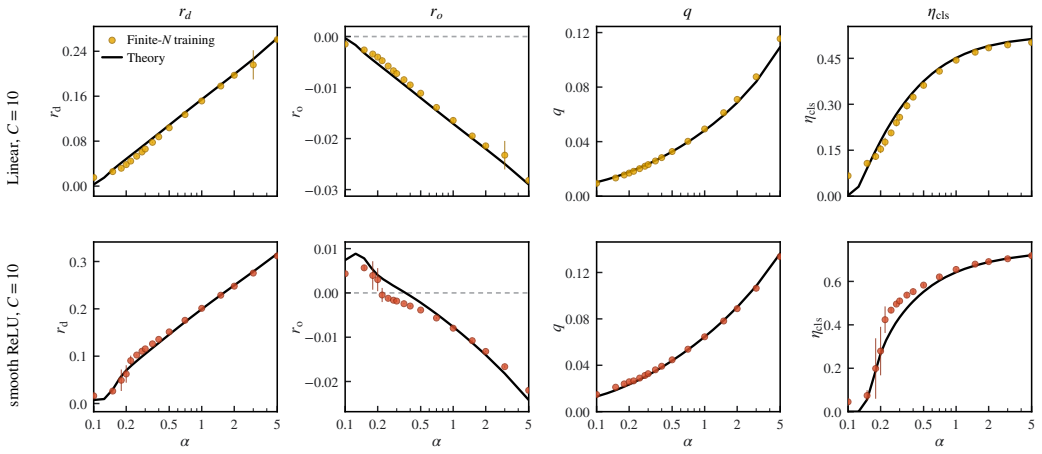

Figure 5: Quantitative theory–simulation correspondence. The top and bottom rows correspond to the linear and smooth-ReLU encoders with $C = 10$, respectively. The columns show the matched overlap r_d, unmatched overlap r_o, weight self-overlap q, and class-explained variance fraction η_{cls}. Solid black curves denote mean theoretical predictions across five solver seeds. Colored markers denote means across ten training-data realizations, with error bars showing one sample standard deviation.

Table 2: RMSE between mean theoretical predictions and mean finite-N statistics at the 17 shared values of α, computed without interpolation.

	Linear	Smooth ReLU
r_d	7.01×10^{-3}	6.74×10^{-3}
r_o	9.82×10^{-4}	2.09×10^{-3}
q	2.13×10^{-3}	1.54×10^{-3}
η_{cls}	2.42×10^{-2}	4.18×10^{-2}

The theoretical predictions are obtained from the four zero-temperature saddle-point equations in Eq. (90), solved using Algorithm 1. For each of five solver seeds, we select at each α the validated candidate with the lowest zero-temperature free energy and then average the resulting predictions across seeds. Theoretical η_{cls} is evaluated using analytic Gaussian moments for the linear encoder and 96-point Gauss–Hermite quadrature for smooth ReLU.

Theory and simulation are compared at 17 shared values of α, with the corresponding RMSE values reported in Table 2. For both encoders, the mean theoretical predictions closely track r_d, r_o, q, and η_{cls} at moderate and high α (Fig. 5). The main deviations occur at small α, near the onset of specialization. For the linear encoder, the mismatch is concentrated around $\alpha \simeq 0.1$. For smooth ReLU, variability across training-data realizations is larger for $\alpha \lesssim 0.3$, where training reaches weakly or strongly organized states. At $N = 5{,}000$, this regime corresponds to $P \lesssim 1{,}500$. Finite-N representations are typically more strongly organized than the mean theoretical prediction in this regime. Taken together, the agreement in the overlap statistics and η_{cls} supports the quantitative accuracy of the reduced theory over a broader range than the scans in Fig. 2.

To assess the overlap structure underlying the scalar reduction, we examine the full encoder–class and weight self-overlap matrices from the retained smooth-ReLU simulations across the (C, K, α) scans. We first permute the representation coordinates to maximize the sum of matched encoder–class overlaps. We then fit $\boldsymbol{R}_{\mathrm{fit}} = r_o \mathbf{1}_C \mathbf{1}_C^\top + (r_d - r_o)\boldsymbol{I}_C$, taking r_d and r_o as the mean diagonal and off-diagonal entries of $\boldsymbol{R}^{\mathrm{sim}}$, respectively. The overlap-fit residual and off-diagonal weight

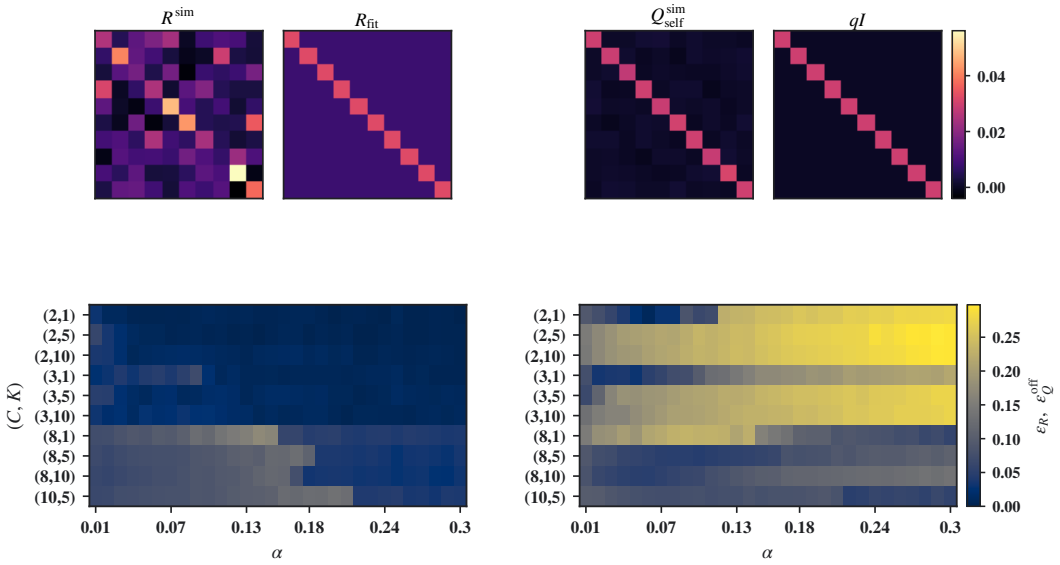

Figure 6: Overlap structure in finite-N smooth-ReLU simulations. For each trained encoder, $\boldsymbol{R}^{\mathrm{sim}} = N^{-1}\boldsymbol{W}\boldsymbol{\Xi}^\top$ and $\boldsymbol{Q}^{\mathrm{sim}}_{\mathrm{self}} = N^{-1}\boldsymbol{W}\boldsymbol{W}^\top$. After permutation alignment, $r_d = C^{-1}\,\mathrm{Tr}\,\boldsymbol{R}^{\mathrm{sim}}$, $r_o = [C(C-1)]^{-1}\sum_{i=1}^{C}\sum_{c\neq i}(\boldsymbol{R}^{\mathrm{sim}})_{ic}$, and $q = C^{-1}\,\mathrm{Tr}\,\boldsymbol{Q}^{\mathrm{sim}}_{\mathrm{self}}$. The top row compares representative empirical matrices with $\boldsymbol{R}_{\mathrm{fit}}$ and $q\boldsymbol{I}_C$. The bottom-left and bottom-right maps show the median ϵ_R and $\epsilon_Q^{\mathrm{off}}$ across retained runs for each (C, K, α) configuration.

self-overlap are quantified by

$$\epsilon_R = \frac{\|\boldsymbol{R}^{\mathrm{sim}} - \boldsymbol{R}_{\mathrm{fit}}\|_F}{\sqrt{Cq}},$$
$$\epsilon_Q^{\mathrm{off}} = \frac{\|\boldsymbol{Q}^{\mathrm{sim}}_{\mathrm{self}} - \mathrm{diag}(\boldsymbol{Q}^{\mathrm{sim}}_{\mathrm{self}})\|_F}{\|\boldsymbol{Q}^{\mathrm{sim}}_{\mathrm{self}}\|_F}. \tag{141}$$

Here, ϵ_R is normalized by the weight scale $\sqrt{Cq}$, while $\mathrm{diag}(\boldsymbol{Q})$ denotes the matrix obtained by retaining only the diagonal entries of $\boldsymbol{Q}$.

The representative matrices in Fig. 6 exhibit a dominant matched–unmatched structure in $\boldsymbol{R}^{\mathrm{sim}}$ and a predominantly diagonal structure in $\boldsymbol{Q}^{\mathrm{sim}}_{\mathrm{self}}$. Across the scans, ϵ_R remains small for $C = 2$ and $C = 3$ over most of the range. For $C = 8$ and $C = 10$, deviations are broader at small and intermediate α but decrease as sample complexity increases. The off-diagonal residual $\epsilon_Q^{\mathrm{off}}$ is largest mainly at smaller C and larger K, particularly for $(C, K) \in \{(2, 5), (2, 10), (3, 5), (3, 10)\}$. Several configurations with $C = 8$ and $C = 10$ become increasingly diagonal as α grows. Taken together, these diagnostics support the matched–unmatched encoder–class overlap structure and the approximate diagonality of the weight self-overlap, while also revealing configuration-dependent finite-N deviations.

C.3 Dependence of the Emergence Route on Contrastive Configuration

We scan smooth ReLU over $C \in \{2, 3, 8, 10\}$ and $K \in \{1, 5, 10\}$ with $M = C$, $N = 10{,}000$, $\epsilon_{\mathrm{sm}} = 0.01$, $\kappa = 15$, $\lambda_w = 0.01$, $\tau = 0.1$, and class-excluded negatives. Forward continuation starts from low-α, low-specialization solutions, whereas backward continuation starts from high-α, high-specialization solutions. In both cases, each converged solution initializes the next continuation point. Most scans use $\Delta\alpha = 0.005$ up to $\alpha = 0.3$. For $(C, K) = (3, 5)$, $(8, 5)$, and $(10, 5)$, we additionally use finer backward scans with resolutions 10^{-4}, 10^{-3}, and 10^{-4} in the respective refined regions. A failed continuation is treated as a numerical endpoint rather than a physical spinodal.

Each continuation point is solved using four Monte Carlo realizations and validated using eight independent realizations, each with 2048 outer samples. Common random numbers are used across forward and backward continuations when their sampling dimensions match. Retained points must pass both convergence and validation checks. When a continuation contains duplicate α values, we retain the point with the smaller scaled saddle residual. Population η_{cls} is computed from (r_d, r_o, q) using 80-point Gauss–Hermite quadrature.

For α values retained by both continuations, we compare their free energies through

$$\begin{aligned} \Delta f(\alpha) &= f_0^{\mathrm{backward}}(\alpha) - f_0^{\mathrm{forward}}(\alpha), \\ \sigma_{\Delta f}^2 &= \sigma_{\mathrm{backward}}^2 + \sigma_{\mathrm{forward}}^2 - 2\,\mathrm{Cov}\left(\widehat{f}_0^{\mathrm{backward}}, \widehat{f}_0^{\mathrm{forward}}\right), \end{aligned} \tag{142}$$

where the hats denote validation estimates and $\sigma_{\mathrm{forward}}$ and $\sigma_{\mathrm{backward}}$ denote their standard errors. The covariance term vanishes when the validation estimates are independent. Within the reduced RS ansatz, positive Δf favors the forward continuation, whereas negative Δf favors the backward continuation. We regard this ordering as resolved when $|\Delta f| > 2\sigma_{\Delta f}$. We further regard a crossing between distinctly organized states as resolved when neighboring sampled points have opposite resolved signs and the between-branch difference in η_{cls} exceeds 0.02 at both endpoints. For $C = 3$, the 2σ criterion is not met, so we report only an unresolved sign change. For a resolved crossing bracketed by α_0 and α_1, linear interpolation gives

$$\alpha_{\mathrm{th}} = \alpha_0 - \Delta f(\alpha_0)\frac{\alpha_1 - \alpha_0}{\Delta f(\alpha_1) - \Delta f(\alpha_0)}. \tag{143}$$

For $C = 2$, the forward and backward continuations overlap, and class organization grows smoothly from nearly zero at the smallest sampled α (Fig. 7). An exactly unspecialized stationary solution can still be recovered through separate constrained or random-initialization searches, but these continuations do not follow that solution. For $C = 3$, the forward continuation remains near zero specialization throughout the scanned range, whereas the backward continuation grows smoothly from near-zero specialization at small α. Thus, a continuously emerging specialized branch coexists with a persistent low-specialization branch.

For $C \in \{8, 10\}$, the backward continuation appears at finite class organization while the low-specialization forward continuation remains available (Fig. 7). The finite-amplitude branch is already present below the free-energy crossing, where the lower-free-energy branch switches discontinuously from the forward to the backward continuation (Fig. 8). Across the sampled values of K, the distinction between continuous specialization for smaller C and finite-amplitude competition for larger C remains unchanged, while the characteristic α values shift. Increasing K generally shifts both the lowest-α points reached by the finite-amplitude continuation and the free-energy crossing toward smaller α. This trend is clearest for $C = 8$ and between $K = 1$ and $K = 5$ for $C = 10$.

Across the scanned configurations, the observed emergence route differs with C, whereas varying K primarily shifts the characteristic sample complexities. Within the sampled values of K, more negatives generally shift specialization toward lower sample complexity without changing the observed route.

C.4 Downstream Generalization from Replica Order Parameters

We evaluate downstream generalization using the lower-free-energy smooth-ReLU replica candidates for $C \in \{2, 3, 8\}$ and $K \in \{5, 10\}$. At each retained saddle, the order parameters (r_d, r_o, q) determine the separation Δ_{LP}. This separation is then mapped to the population-optimal linear-probe error derived in Sec. A.9. The Gaussian expectation is evaluated by 96-point Gauss–Hermite quadrature. We compare the resulting prediction with the test error of a calibrated localist classifier applied to finite-N representations.

For each $\alpha \in \{0.01, 0.02, \ldots, 0.30\}$ at $N = 10{,}000$, we train ten independently initialized networks at that value of α, without continuation across sample ratios. Each network is initialized with isotropic Gaussian weights and optimized with Adam for at most $30{,}000$ iterations at learning rate 0.03, followed by at most 100 L-BFGS iterations with relative-gradient tolerance 10^{-5}. Among the converged restarts, we retain the network with the lowest independent-validation energy. Hidden units are matched to classes using an independent calibration set of 1000 samples, after which the resulting localist classifier is evaluated on a disjoint test set of 2000 samples.

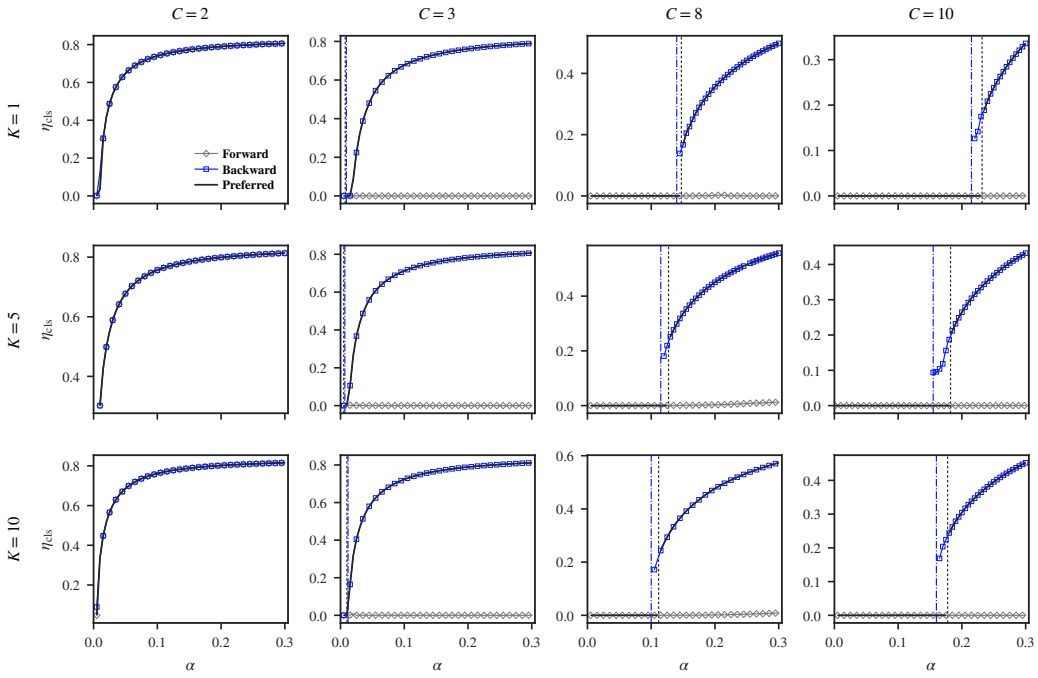

Figure 7: Class-organization branches across smooth-ReLU configurations. Gray diamonds and blue squares denote forward and backward continuations of η_{cls}; black segments indicate the branch with lower estimated free energy. Blue dash-dotted and black dashed vertical lines mark backward-continuation termination and estimated free-energy crossings, respectively.

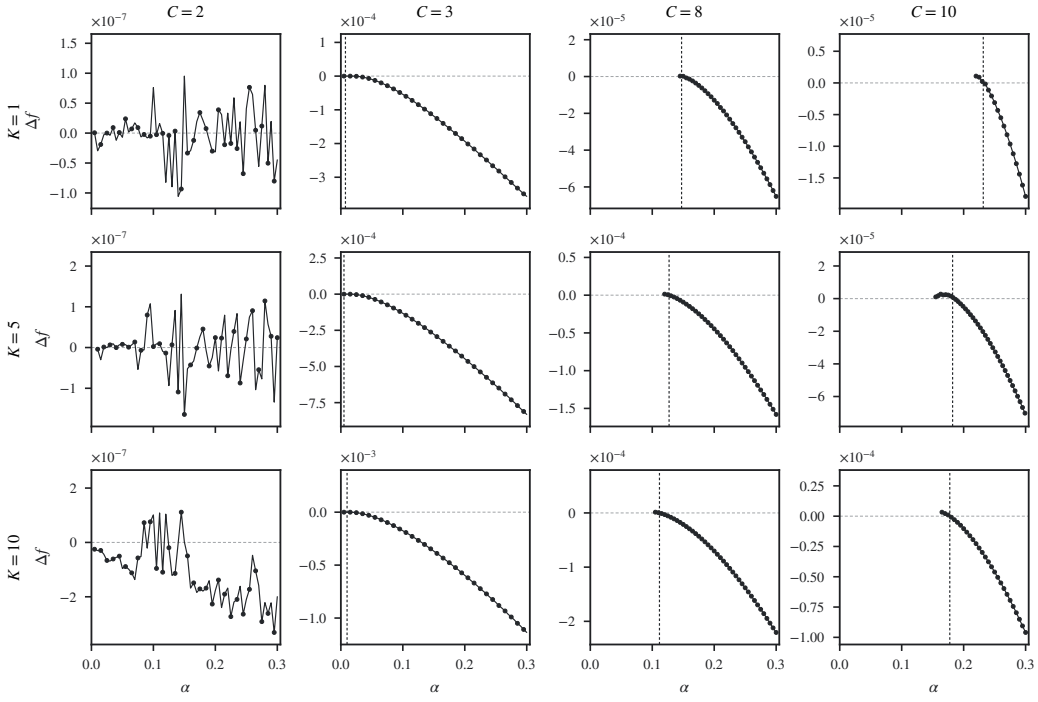

Figure 8: Free-energy comparison across smooth-ReLU configurations. The panel arrangement matches Fig. 7. Positive and negative $\Delta f = f_0^{\text{backward}} - f_0^{\text{forward}}$ indicate lower free energy for the forward and backward continuations, respectively. Black dashed lines mark estimated crossings; for $C = 3$, the observed sign changes do not satisfy the 2σ resolution criterion.

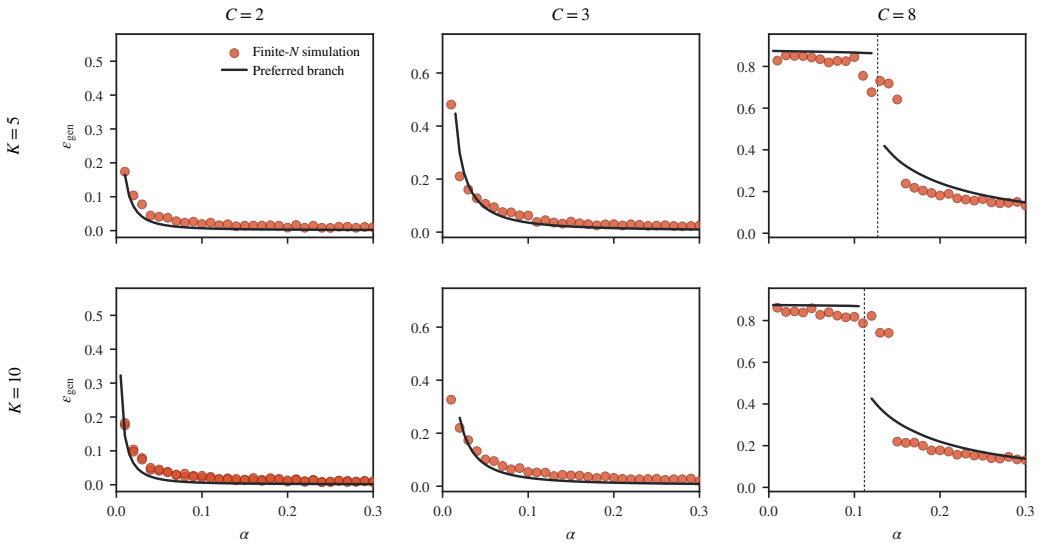

Figure 9: Replica prediction of downstream generalization. Solid black curves show predictions from the lower-free-energy replica candidates, while colored markers show selected finite-N networks. For each replica saddle, $\Delta_{\text{LP}} = \sqrt{\kappa/[(1-\gamma)q]}\,(r_d - r_o)$ and $\epsilon_{\text{gen}} = 1 - \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\Phi(Z + |\Delta_{\text{LP}}|)^{C-1}]$, where Φ is the standard normal CDF and the expectation is evaluated by 96-point Gauss–Hermite quadrature. Each finite-N marker corresponds to the lowest-validation-energy converged network among ten independent random initializations, with $\epsilon_{\text{gen}}^{\text{sim}} = 1 - \text{Acc}_{\text{localist}}$.

The replica prediction closely tracks the finite-N generalization error across the six configurations without fitting to the simulation results (Fig. 9). For $C = 2$, the theory reproduces the smooth decrease in error associated with continuous specialization. For $C = 3$, the forward and backward continuations remain close in generalization error, and the simulations follow the predicted smooth improvement. For $C = 8$, the finite-amplitude backward continuation corresponds to a distinct lower-error regime that is also observed in the trained networks. The largest discrepancies occur near branch selection, where the selected finite-N networks exhibit a smoother and slightly shifted crossover.

Taken together, these results support the connection between specialization and downstream performance through $|\Delta_{\text{LP}}|$. At fixed C, changing K affects the predicted error through the learned saddle, whereas C also enters the mapping from Δ_{LP} to ϵ_{gen} explicitly.

C.5 Robustness to Unconditioned Negative Sampling

Under unconditioned negative sampling (UCL), each negative class is drawn independently from the uniform prior over all C classes, allowing same-class negatives (Sec. A.8). For the zero-temperature analysis, we use the model parameters and numerical procedures of Sec. C.3, replacing class-excluded sampling with UCL. The forward and backward continuation procedures, numerical validation, and free-energy comparison are otherwise unchanged. We consider linear $C = 10$, smooth ReLU $C = 10$, and smooth ReLU $C = 2$, all with $M = C$ and $K = 5$.

For the finite-N comparison, we focus on smooth ReLU with $M = C = 10$ and $K = 5$. Negatives are sampled uniformly from the training anchor pool, excluding the anchor itself while allowing same-class samples, and their assignments are held fixed throughout training. For random initialization, we use $N = 10{,}000$ and ten Gaussian restarts at each $\alpha \in \{0.01, 0.02, \ldots, 0.30\}$. For strongly specialized initialization, we use the same N and five restarts for each of two training-data realizations with shared class directions. The initialization targets $(r_d, r_o, q) = (0.32, -0.025/(C-1), 0.25)$, with random components orthogonal to the class directions. Both protocols use at most 30,000 Adam steps followed by 200 L-BFGS steps, with relative-gradient thresholds of 10^{-6} and 10^{-5} for random and strongly specialized initialization, respectively. We retain 269 of 300 random-

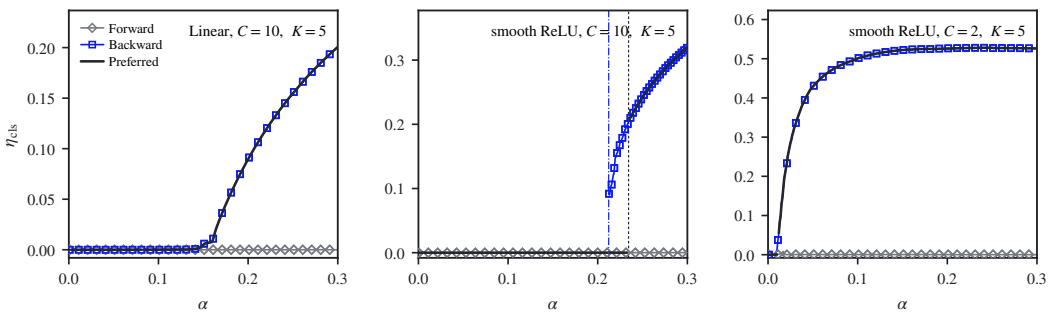

Figure 10: Specialization routes under unconditioned negative sampling. Gray diamonds and blue squares show forward and backward continuations, respectively, while black segments indicate the lower-free-energy branch. Linear $C = 10$ and smooth ReLU $C = 2$ exhibit continuous specialization, whereas smooth ReLU $C = 10$ exhibits a distinct finite-amplitude backward branch.

initialization runs and all 300 strongly initialized runs. Population $\eta_{\rm cls}$ is evaluated for each retained network using 64-point Gauss–Hermite quadrature.

Across the three configurations examined here, the UCL calculation reproduces the same qualitative distinction in emergence routes observed under class-excluded sampling. Linear $C = 10$ and smooth ReLU $C = 2$ specialize continuously, whereas smooth ReLU $C = 10$ exhibits a distinct finite-amplitude backward-continuation branch. Within the reduced RS theory, this branch becomes lower in free energy at $\alpha_{\rm th} \simeq 0.2344$ (Fig. 10).

Random initialization predominantly reaches weakly organized states, whereas strongly specialized initialization reaches strongly organized states associated with the finite-amplitude backward-continuation branch (Fig. 11). The two protocols also differ in stopping tolerance and in how training-data realizations are allocated across restarts, so this comparison does not isolate the effect of initialization. Nevertheless, the results show that lower RS free energy does not imply that finite-N optimization from the random-initialization protocol studied here reaches the strongly organized state.

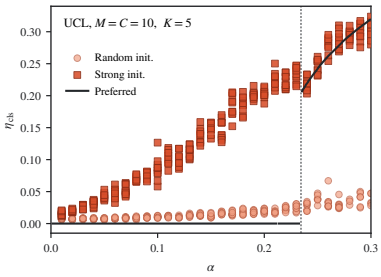

Figure 11: Finite-N UCL outcomes for smooth ReLU. Light and dark red markers denote random and strongly specialized initializations, respectively, while the black curve shows the lower-RS-free-energy theoretical branch.

C.6 Class Organization and Linear-probe Performance on CIFAR-10

The preceding analysis distinguishes class organization from spectral effective rank and relates $\eta_{\rm cls}$ to downstream generalization in the solvable model. We next examine whether $\eta_{\rm cls}$ also tracks linear-probe performance in a standard contrastive-learning pipeline.

We train ResNet-18 (He et al., 2016) with SimCLR (Chen et al., 2020) on CIFAR-10 (Krizhevsky, 2009), using a CIFAR stem and a two-layer projector with a 512-dimensional hidden layer, batch normalization, ReLU, and output dimension $d \in \{10, 128\}$. Training uses NT-Xent with temperature 0.1, two augmented views per image, batch size 256, and 30,000 SGD steps with momentum 0.9, weight decay 10^{-4}, and mixed precision. The learning rate increases linearly to 0.3 over the first 500 warmup steps and then follows a cosine decay. Augmentations include random resized crops with scale range $[0.2, 1]$, horizontal flips with probability 0.5, color jitter with strengths $(0.8, 0.8, 0.8, 0.2)$ and probability 0.8, grayscale conversion with probability 0.2, and Gaussian blur with probability 0.5, kernel size 3, and standard deviation in $[0.1, 2]$. CIFAR-10 channel normalization is applied after augmentation.

We vary the number of unique unlabeled images P over 29 values from 500 to 50,000, while keeping the training-step budget fixed. The subsets are class balanced and nested within each seed, and the

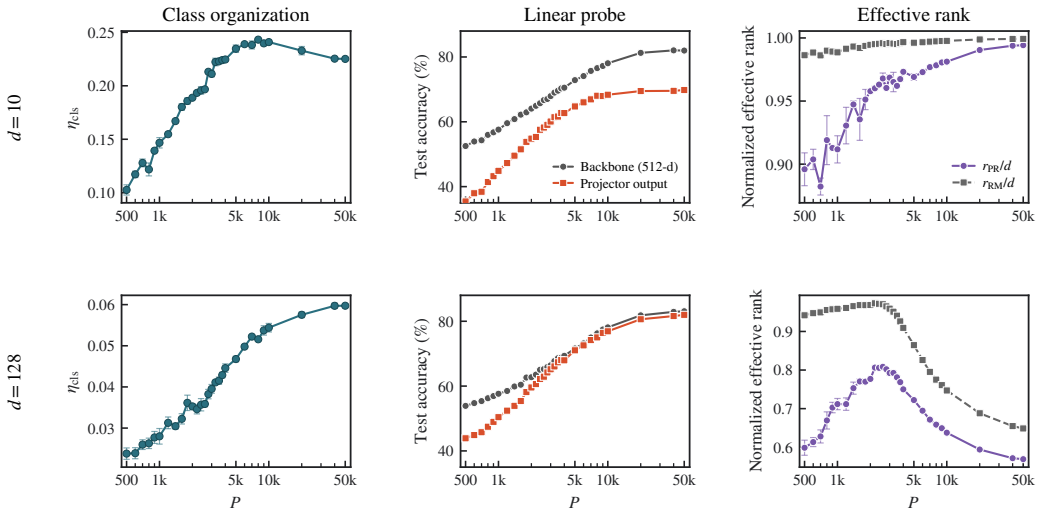

Figure 12: Class organization, linear-probe performance, and spectral effective rank on CIFAR-10. Rows correspond to projector dimensions $d = 10$ and $d = 128$, while columns show η_{cls}, frozen linear-probe accuracy, and normalized effective rank as the pretraining-set size P increases. For the ℓ_2-normalized projector representation, $\eta_{\text{cls}} = \operatorname{Tr} \widehat{\boldsymbol{\Sigma}}_B / \operatorname{Tr} \widehat{\boldsymbol{\Sigma}}$, where $\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_i (\boldsymbol{v}_i - \bar{\boldsymbol{v}})(\boldsymbol{v}_i - \bar{\boldsymbol{v}})^\top$ and $\widehat{\boldsymbol{\Sigma}}_B = \sum_c p_c (\bar{\boldsymbol{v}}_c - \bar{\boldsymbol{v}})(\bar{\boldsymbol{v}}_c - \bar{\boldsymbol{v}})^\top$. Here, n is the evaluation-set size, p_c is the empirical class fraction, and $\bar{\boldsymbol{v}}$ and $\bar{\boldsymbol{v}}_c$ denote the overall and class means. The spectral measures are r_{PR}/d, with $r_{\text{PR}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$, and the centered RankMe effective rank r_{RM}/d (Roy & Vetterli, 2007; Garrido et al., 2023), with $r_{\text{RM}} = \exp[-\sum_i \rho_i \log \rho_i]$ and $\rho_i = \sqrt{\lambda_i} / \sum_j \sqrt{\lambda_j}$. The λ_i are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, with $0 \log 0 = 0$. The middle column reports test accuracy for separate probes trained on the frozen backbone and projector representations. Panels show means and SEM across the contributing seeds, without smoothing or interpolation.

same initialization is reused across P within a seed. Completed runs for both projector dimensions use seeds $\{41, 42, 43\}$. In Fig. 12, the $d = 10$ representation statistics aggregate seeds $\{41, 42\}$, whereas the $d = 128$ representation statistics and the linear-probe results for both dimensions aggregate all three seeds.

After pretraining, the encoder and projector are frozen. Representation statistics are computed from the ℓ_2-normalized projector outputs on the full CIFAR-10 test set. Separate affine linear probes are trained on the raw 512-dimensional backbone features and the normalized d-dimensional projector features using cross-entropy loss. All probes use the same stratified split of the CIFAR-10 training set, consisting of 45,000 training images and 5,000 validation images with 500 validation images per class, independently of P. Each probe is trained with SGD and momentum 0.9 for 100 epochs using batch size 1024, zero weight decay, and initial learning rates 0.1 and 1.0 for the backbone and projector, respectively. The learning rate follows a half-cosine decay. Validation accuracy selects the checkpoint that is subsequently evaluated on the 10,000 test images.

Across both projector dimensions, η_{cls} and projector linear-probe accuracy follow similar trends as P increases (Fig. 12). For $d = 10$, both rise rapidly and approach a plateau, while backbone probe accuracy continues to improve. For $d = 128$, both increase over most of the sampled range, and projector probe accuracy approaches backbone probe accuracy at large P. Across these settings, η_{cls} therefore tracks the qualitative trend in the linear-probe performance of class information in the representation optimized by the contrastive objective.

The spectral measures follow different trends. Effective-rank statistics characterize spectral dimensionality and representation collapse in self-supervised learning (Jing et al., 2022; Garrido et al., 2023), but they need not track class organization. For $d = 10$, both normalized ranks continue to increase after η_{cls} and projector accuracy have largely saturated. For $d = 128$, both ranks peak and then decrease while η_{cls} and projector accuracy continue to increase. Across these settings, η_{cls}

follows the qualitative trend in projector linear-probe accuracy, whereas the effective-rank measures can evolve differently.

D Stability and Branch Selection

This section distinguishes specialization, local stability, and branch selection. We first derive the specialization stiffness of the unspecialized branch (Sec. D.1) and then characterize the nonlinear response associated with the observed emergence routes through a Landau-type expansion (Sec. D.2). We next compare the free energies of competing branches (Sec. D.3) and assess their local stability within the reduced overlap ansatz (Sec. D.4). Finally, we evaluate a scalar zero-temperature response diagnostic (Sec. D.5).

D.1 Local Specialization Stability of the Unspecialized Branch

We parameterize the matched and unmatched overlaps by

$$\begin{aligned} \bar{r} &= \frac{r_d + (C-1)r_o}{C}, \quad \Delta r = r_d - r_o, \\ r_d &= \bar{r} + \frac{C-1}{C}\Delta r, \qquad r_o = \bar{r} - \frac{\Delta r}{C}. \end{aligned} \tag{144}$$

Let $\boldsymbol{s} = (\bar{r}, q, \chi)$ denote the remaining class-symmetric variables. Assuming a smooth interior stationary branch with an invertible Hessian block $\partial_{\boldsymbol{s}}^2 f_0$, define the profiled free energy by

$$F(\Delta r; \alpha) = f_0(\Delta r, \boldsymbol{s}^\star(\Delta r; \alpha); \alpha)\,, \qquad \partial_{\boldsymbol{s}} f_0|_{\boldsymbol{s}^\star} = \mathbf{0}. \tag{145}$$

Let $\boldsymbol{s}_0 = (\bar{r}_0, q_0, \chi_0)$ denote the unspecialized stationary point, and use $\cdot|_0$ for evaluation at $\Delta r = 0$ and $\boldsymbol{s} = \boldsymbol{s}_0$. Differentiating the stationarity conditions gives

$$\frac{d\boldsymbol{s}^\star}{d\Delta r} = -\left(\partial_{\boldsymbol{s}}^2 f_0\right)^{-1} \partial_{\boldsymbol{s}} \partial_{\Delta r} f_0. \tag{146}$$

Substituting this response into the second derivative of F yields

$$\begin{aligned} \frac{d^2F}{d(\Delta r)^2} &= \partial_{\Delta r}^2 f_0 - \partial_{\Delta r}\partial_{\boldsymbol{s}} f_0 \left(\partial_{\boldsymbol{s}}^2 f_0\right)^{-1} \partial_{\boldsymbol{s}} \partial_{\Delta r} f_0, \\ \Lambda_{\mathrm{sp}} &\equiv \left.\frac{d^2F}{d(\Delta r)^2}\right|_0 = \frac{C-1}{\chi_0} + \left.\frac{\partial^2 \mathcal{G}_\alpha}{\partial(\Delta r)^2}\right|_0. \end{aligned} \tag{147}$$

Permutation symmetry implies $\partial_{\Delta r}\partial_{\boldsymbol{s}} f_0|_0 = 0$, so the specialization mode decouples from the class-symmetric variables to quadratic order. Along this direction, the explicit weight term in Eq. (89) contributes $(C-1)(\Delta r)^2/(2\chi_0)$. The unspecialized branch is locally stable against specialization when $\Lambda_{\mathrm{sp}} > 0$ and marginal when $\Lambda_{\mathrm{sp}} = 0$. Negative stiffness signals local instability.

D.2 Landau Expansion and Local Specialization Response

The specialization stiffness determines local stability but does not by itself distinguish continuous specialization from finite-amplitude competition. Assuming sufficient smoothness, we expand the profiled free energy around $\Delta r = 0$ as

$$F(\Delta r; \alpha) = F_0 + \frac{a_2}{2}(\Delta r)^2 + \frac{a_3}{3}(\Delta r)^3 + \frac{a_4}{4}(\Delta r)^4 + O((\Delta r)^5), \tag{148}$$

where $F_0 = F(0; \alpha)$ and the coefficients depend on α and the model parameters. The symmetry decoupling above implies

$$\left.\frac{d\boldsymbol{s}^\star}{d\Delta r}\right|_0 = \mathbf{0}, \qquad \left.\frac{d^2\boldsymbol{s}^\star}{d(\Delta r)^2}\right|_0 = -\left[\left(\partial_{\boldsymbol{s}}^2 f_0\right)^{-1} \partial_{\boldsymbol{s}} \partial_{\Delta r}^2 f_0\right]_0. \tag{149}$$

The resulting coefficients are

$$\begin{aligned} a_2 &= \Lambda_{\mathrm{sp}} = \frac{C-1}{\chi_0} + \left.\frac{\partial^2 \mathcal{G}_\alpha}{\partial(\Delta r)^2}\right|_0, \\ a_3 &= \frac{1}{2}\left.\frac{\partial^3 \mathcal{G}_\alpha}{\partial(\Delta r)^3}\right|_0, \\ a_4 &= \frac{1}{6}\left[\partial_{\Delta r}^4 f_0 - 3\,\partial_{\Delta r}^2 \partial_{\boldsymbol{s}} f_0 \left(\partial_{\boldsymbol{s}}^2 f_0\right)^{-1} \partial_{\boldsymbol{s}} \partial_{\Delta r}^2 f_0\right]_0. \end{aligned} \tag{150}$$

Thus, the stationary response of $\boldsymbol{s}$ does not contribute to the cubic coefficient at this order, whereas its relaxation enters the quartic coefficient.

Within the quartic approximation, stationary points satisfy

$$\begin{aligned} 0 &= \Delta r \left[a_2 + a_3 \Delta r + a_4 (\Delta r)^2\right], \\ \Delta r_\pm &= \frac{-a_3 \pm \sqrt{a_3^2 - 4a_2 a_4}}{2a_4}, \qquad a_4 > 0, \quad a_3^2 - 4a_2 a_4 \geq 0. \end{aligned} \tag{151}$$

For $a_3 = 0$ and $a_4 > 0$, the specialized minimum emerges continuously as a_2 becomes negative, with $(\Delta r)^2 = -a_2/a_4$ in the quartic truncation. For $a_3 < 0$ and $a_4 > 0$, the interval $0 < a_2 < a_3^2/(4a_4)$ permits a finite positive minimum at Δr_+ while $\Delta r = 0$ remains locally stable. These conditions characterize the local quartic landscape but do not establish competing branches of the full profiled free energy.

For the linear encoder, the overlap matrix can be written as

$$\begin{aligned} \boldsymbol{R} &= \bar{r}\, \boldsymbol{J}_C + \Delta r \left(\boldsymbol{I}_C - \frac{1}{C}\boldsymbol{J}_C\right), \\ \left(\frac{2}{C}\boldsymbol{J}_C - \boldsymbol{I}_C\right) \boldsymbol{R} &= \bar{r}\, \boldsymbol{J}_C - \Delta r \left(\boldsymbol{I}_C - \frac{1}{C}\boldsymbol{J}_C\right). \end{aligned} \tag{152}$$

The transformation

$$\boldsymbol{W} \mapsto \left(\frac{2}{C}\boldsymbol{J}_C - \boldsymbol{I}_C\right) \boldsymbol{W} \tag{153}$$

therefore reverses the sign of Δr while preserving $\bar{r}$, the scalar weight overlaps, and the regularized linear-encoder objective. It follows that

$$F_{\mathrm{lin}}(\Delta r; \alpha) = F_{\mathrm{lin}}(-\Delta r; \alpha) = F_0 + \frac{a_2}{2}(\Delta r)^2 + \frac{a_4}{4}(\Delta r)^4 + O((\Delta r)^6). \tag{154}$$

All odd coefficients therefore vanish. When $a_4 > 0$, this even local structure permits continuous specialization as a_2 changes sign.

For a nonlinear encoder, the cubic term need not vanish. To identify its leading dependence at high τ, define the scalar reference field

$$H_0 = \sqrt{\kappa}\, \bar{r}_0 + \sqrt{(1-\gamma) q_0}\, Z, \qquad Z \sim \mathcal{N}(0,1). \tag{155}$$

At fixed $(\bar{r}, q) = (\bar{r}_0, q_0)$, the class-conditional representation mean expands as

$$\begin{aligned} \mathbb{E}[\boldsymbol{v} \mid c] = \mathbb{E}[\phi(H_0)]\, \mathbf{1}_C + \sqrt{\kappa}\, \mathbb{E}[\phi'(H_0)]\, \Delta r \left(\boldsymbol{e}^{(c)} - \frac{1}{C}\mathbf{1}_C\right) \\ + \frac{\kappa}{2}\mathbb{E}[\phi''(H_0)]\, (\Delta r)^2 \left(\boldsymbol{e}^{(c)} - \frac{1}{C}\mathbf{1}_C\right)^{\odot 2} + O((\Delta r)^3), \end{aligned} \tag{156}$$

where $\odot 2$ denotes coordinate-wise squaring. Let m_d and m_o denote the matched and unmatched components of this mean. Their difference satisfies

$$m_d - m_o = \sqrt{\kappa}\, \mathbb{E}[\phi'(H_0)] \Delta r + \frac{\kappa}{2}\frac{C-2}{C}\mathbb{E}[\phi''(H_0)](\Delta r)^2 + O((\Delta r)^3). \tag{157}$$

Writing $s^+ = \boldsymbol{v}^\top \boldsymbol{v}^+$ and $s_k^- = \boldsymbol{v}^\top \boldsymbol{v}_k^-$, conditional independence together with class-excluded sampling gives

$$\begin{aligned} \mathbb{E}[s_k^- - s^+] &= -(m_d - m_o)^2 \\ &= -\kappa \mathbb{E}[\phi'(H_0)]^2 (\Delta r)^2 \\ &\quad - \frac{C-2}{C}\kappa^{3/2}\mathbb{E}[\phi'(H_0)]\mathbb{E}[\phi''(H_0)](\Delta r)^3 + O((\Delta r)^4). \end{aligned} \tag{158}$$

The local InfoNCE loss has the expansion

$$\ell_\tau = \log(K+1) + \frac{1}{(K+1)\tau}\sum_{k=1}^{K}(s_k^- - s^+) + O(\tau^{-2}). \tag{159}$$

Table 3: Local Landau diagnostic for the linear encoder. Λ_{sp} denotes the specialization stiffness of the unspecialized branch, while $\widehat{a}_2^{\mathrm{prof}}$ and $\widehat{a}_4^{\mathrm{prof}}$ are jointly fitted from its symmetric profiled free energy. Parentheses report estimates obtained with independent validation disorder. Near the zero crossing, small sign differences reflect numerical uncertainty in locating the onset rather than qualitatively different behavior.

α	Λ_{sp}	$\widehat{a}_2^{\mathrm{prof}}$	$\widehat{a}_4^{\mathrm{prof}}$
0.1215	5.36×10^{-5} (3.93×10^{-4})	2.87×10^{-5} (4.65×10^{-5})	9.17 (9.17)
0.1225	-4.65×10^{-4} $(+1.93\times10^{-5})$	-4.92×10^{-4} (-4.73×10^{-4})	9.11 (9.11)
0.1240	-4.88×10^{-4} (-8.63×10^{-4})	-5.18×10^{-4} (-5.31×10^{-4})	9.16 (9.16)

To connect this expansion to the effective minimum entering $\mathcal{G}_\alpha$, stack the local fields and their centers into $\boldsymbol{h}$ and $\boldsymbol{u}$. At a regular background with finite positive χ_0, local stationarity gives

$$\begin{aligned} \boldsymbol{h}^\star - \boldsymbol{u} &= -(1-\gamma)\chi_0 \nabla_{\boldsymbol{h}} \ell_\tau(\boldsymbol{h}^\star) = O(\tau^{-1}), \\ V^\star_{\chi_0} &= \ell_\tau(\boldsymbol{u}) + O(\tau^{-2}). \end{aligned} \tag{160}$$

Assuming that this expansion remains valid after averaging and differentiation through third order in Δr, the energetic contribution at $\boldsymbol{s}_0 = (\bar{r}_0, q_0, \chi_0)$ satisfies

$$\begin{aligned} \mathcal{G}_\alpha(\Delta r; \boldsymbol{s}_0) - \mathcal{G}_\alpha(0; \boldsymbol{s}_0) = &-\alpha \frac{K}{K+1} \frac{\kappa}{\tau} \mathbb{E}[\phi'(H_0)]^2 (\Delta r)^2 \\ &- \alpha \frac{K}{K+1} \frac{C-2}{C} \frac{\kappa^{3/2}}{\tau} \mathbb{E}[\phi'(H_0)] \mathbb{E}[\phi''(H_0)] (\Delta r)^3 \\ &+ O\big(\tau^{-1}(\Delta r)^4\big) + O(\tau^{-2}). \end{aligned} \tag{161}$$

Comparing the cubic terms with Eq. (148) gives

$$a_3 = -3\alpha \frac{K}{K+1} \frac{C-2}{C} \frac{\kappa^{3/2}}{\tau} \mathbb{E}[\phi'(H_0)] \mathbb{E}[\phi''(H_0)] + O(\tau^{-2}). \tag{162}$$

The leading contribution vanishes for a linear activation and for $C = 2$. For smooth ReLU with $C > 2$, both derivative expectations are positive, resulting in a negative leading cubic coefficient. This provides a local source of asymmetry. Full branch coexistence and relative preference must instead be established by continuation and free-energy comparison.

For the linear encoder with $M = C = 10$, $K = 5$, $\kappa = 15$, $\lambda_w = 0.01$, and $\tau = 0.1$, we jointly fit the quadratic and quartic coefficients of the symmetric profiled free energy:

$$\frac{F(\Delta r) + F(-\Delta r)}{2} - F(0) \simeq \frac{\widehat{a}_2^{\mathrm{prof}}}{2} (\Delta r)^2 + \frac{\widehat{a}_4^{\mathrm{prof}}}{4} (\Delta r)^4, \qquad |\Delta r| \in \{0.004, 0.006, 0.008\}. \tag{163}$$

Table 3 reports the fitted coefficients together with an independently evaluated specialization stiffness.

The stiffness estimate on the solving disorder changes sign between $\alpha = 0.1215$ and 0.1225, whereas the independent validation estimate changes sign between 0.1225 and 0.124. The fitted quadratic coefficient changes sign in the same neighborhood, while the fitted quartic coefficient remains positive. Together with the exact even symmetry, these results support continuous local specialization, with numerical variation in the estimated threshold.

D.3 Finite-Amplitude Branch Competition and Thermodynamic Selection

The Landau expansion characterizes only the neighborhood of the unspecialized reference state. To examine the full stationary solutions, we trace branches by forward continuation from low α and backward continuation from high α. Where both branches are available, their reduced RS free energies are compared through

$$\Delta f(\alpha) = f_0^{\mathrm{backward}}(\alpha) - f_0^{\mathrm{forward}}(\alpha), \qquad \Delta f(\alpha_{\mathrm{th}}) = 0. \tag{164}$$

Table 4: Free-energy selection between smooth-ReLU branches. $\Delta r^{\mathrm{backward}}$ denotes the specialization amplitude of the backward branch, and $\Lambda_{\mathrm{sp}}^{\mathrm{forward}}$ is evaluated on the constrained unspecialized branch with $r_d = r_o$. The free-energy difference is $\Delta f = f_0^{\mathrm{backward}} - f_0^{\mathrm{forward}}$. Parentheses give estimates obtained with independent validation disorder.

α	$\Delta r^{\mathrm{backward}}$	$\Lambda_{\mathrm{sp}}^{\mathrm{forward}}$	Δf
0.160	3.70×10^{-2}	2.10×10^{-2} (2.08×10^{-2})	1.10×10^{-6} (1.14×10^{-6})
0.170	4.26×10^{-2}	1.90×10^{-2} (1.94×10^{-2})	2.19×10^{-6} (2.24×10^{-6})
0.180	5.44×10^{-2}	1.75×10^{-2} (1.76×10^{-2})	-3.67×10^{-7} (-3.16×10^{-7})
0.200	6.63×10^{-2}	1.47×10^{-2} (1.48×10^{-2})	-6.87×10^{-6} (-6.81×10^{-6})
0.250	8.74×10^{-2}	8.67×10^{-3} (7.39×10^{-3})	-3.43×10^{-5} (-3.42×10^{-5})
0.300	1.03×10^{-1}	3.09×10^{-3} (2.50×10^{-3})	-7.49×10^{-5} (-7.48×10^{-5})
0.330	1.11×10^{-1}	3.37×10^{-4} (7.60×10^{-4})	-1.03×10^{-4} (-1.03×10^{-4})
0.340	1.14×10^{-1}	-6.44×10^{-4} (-2.16×10^{-4})	-1.10×10^{-4} (-1.10×10^{-4})

Thus, $\Delta f > 0$ favors the forward branch, whereas $\Delta f < 0$ favors the backward branch.

For smooth ReLU with $M = C = 10$, $K = 5$, $\kappa = 15$, $\lambda_w = 0.01$, and $\tau = 0.1$, we compare the two branches and evaluate the specialization stiffness of the constrained unspecialized branch with $r_d = r_o$.

A finite-amplitude backward branch is already present at $\alpha = 0.16$, while the unspecialized branch retains positive specialization stiffness. The free-energy difference changes sign between $\alpha = 0.17$ and 0.18. By contrast, the unspecialized branch loses local specialization stability only between $\alpha = 0.33$ and 0.34 (Table 4). Finite-amplitude branch existence, free-energy selection, and local specialization instability are therefore distinct events.

D.4 Local Stability of Competing Branches

At finite Δr, the class-common and class-contrast directions can couple. We evaluate their curvature in $(\bar{r}, \Delta r)$ after allowing (q, χ) to satisfy the corresponding stationarity conditions. Assuming that the (q, χ) Hessian block is invertible, the relaxed Hessian is

$$\begin{aligned} \boldsymbol{H}_{\mathrm{phys}} &= \nabla^2_{(\bar{r},\Delta r)} f_0 - \nabla^2_{(\bar{r},\Delta r),(q,\chi)} f_0 \left[\nabla^2_{(q,\chi)} f_0\right]^{-1} \nabla^2_{(q,\chi),(\bar{r},\Delta r)} f_0, \\ \lambda_{\min}^{\mathrm{phys}} &= \lambda_{\min}(\boldsymbol{H}_{\mathrm{phys}}). \end{aligned} \tag{165}$$

A stationary solution is locally stable along these profiled directions when $\lambda_{\min}^{\mathrm{phys}} > 0$. At $\Delta r = 0$, permutation symmetry decouples the specialization mode, whose curvature reduces to Λ_{sp}.

For the smooth-ReLU coexistence regime, we evaluate the Hessian at the numerical forward and backward solutions, including the small but nonzero specialization of the forward branch. Central differences of the stationary-action gradient use base steps $(2\times10^{-4}, 2\times10^{-4}, 2\times10^{-5}, 0.05)$ in $(\bar{r}, \Delta r, q, \chi)$ and a second set of steps twice as large. We symmetrize the resulting Hessian and evaluate the Schur complement using a numerical pseudoinverse with relative cutoff 10^{-10}, while also checking the invertibility of the nuisance block. Independent validation disorder provides a second estimate.

All reported relaxed-curvature estimates are positive (Table 5). These results support local stability of both branches along the profiled overlap directions on either side of the free-energy crossing.

D.5 Scalar Zero-Temperature Response Diagnostic

We complement the overlap-curvature analysis with a scalar response diagnostic motivated by Almeida–Thouless stability analysis (de Almeida & Thouless, 1978; Engel & Van den Broeck, 2001). The diagnostic is computed directly from the zero-temperature effective single-group problem.

Table 5: Relaxed-curvature estimates for competing smooth-ReLU branches. Intervals span the smallest-eigenvalue estimates obtained with two finite-difference scales and are not confidence intervals. The validation column reports estimates obtained with independent Monte Carlo disorder.

α	branch	Δr	$\lambda_{\min}^{\mathrm{phys}}$	validation
0.171	forward	1.38×10^{-3}	$[1.52, 1.57]\times10^{-2}$	$[1.68, 1.85]\times10^{-2}$
0.171	backward	4.63×10^{-2}	$[6.40, 7.85]\times10^{-3}$	$[4.63, 6.99]\times10^{-3}$
0.174	forward	2.89×10^{-4}	$[2.25, 2.29]\times10^{-2}$	$[2.10, 2.19]\times10^{-2}$
0.174	backward	4.91×10^{-2}	$[1.04, 1.14]\times10^{-2}$	$[7.86, 9.82]\times10^{-3}$
0.178	forward	1.04×10^{-3}	$[1.87, 1.89]\times10^{-2}$	$[1.55, 1.57]\times10^{-2}$
0.178	backward	5.21×10^{-2}	$[9.82, 11.13]\times10^{-3}$	$[1.30, 1.39]\times10^{-2}$

Table 6: Scalar zero-temperature response diagnostics. The maximum ρ_{AT} and minimum λ_{AT} are reported separately over each sampled range using the mean of estimates from the solving and independent validation samples.

encoder	branch	α range	$\max \rho_{\mathrm{AT}}$	$\min \lambda_{\mathrm{AT}}$
linear	forward	0.119–0.126	0.408	3.73×10^{-5}
linear	backward	0.119–0.126	0.405	3.76×10^{-5}
smooth ReLU	forward	0.160–0.250	0.740	1.98×10^{-5}
smooth ReLU	backward	0.160–0.250	0.506	4.03×10^{-5}

Stack the anchor, positive, and negative fields into $\boldsymbol{h} \in \mathbb{R}^D$, with $D = C(K+2)$, and denote the corresponding stacked centers by $\boldsymbol{u}$. Write $v = 1 - \gamma$. At a smooth local minimizer,

$$\nabla_{\boldsymbol{h}} \ell_\tau(\boldsymbol{h}^\star) + \frac{\boldsymbol{h}^\star - \boldsymbol{u}}{v\chi} = \boldsymbol{0}. \tag{166}$$

Let $\boldsymbol{H}_\ell = \nabla^2_{\boldsymbol{h}} \ell_\tau(\boldsymbol{h}^\star)$. When $\boldsymbol{I}_D + v\chi \boldsymbol{H}_\ell$ is invertible, differentiation with respect to $\boldsymbol{u}$ gives

$$\begin{aligned} \boldsymbol{J} &\equiv \frac{\partial \boldsymbol{h}^\star}{\partial \boldsymbol{u}} = (\boldsymbol{I}_D + v\chi \boldsymbol{H}_\ell)^{-1}, \\ \boldsymbol{B} &\equiv \nabla^2_{\boldsymbol{u}} V^\star_\chi = \frac{\boldsymbol{I}_D - \boldsymbol{J}}{v\chi} = \boldsymbol{H}_\ell \, (\boldsymbol{I}_D + v\chi \boldsymbol{H}_\ell)^{-1}. \end{aligned} \tag{167}$$

Thus, $\boldsymbol{B}$ measures the curvature of the optimized local objective with respect to its Gaussian field centers.

Partition $\boldsymbol{B}$ into $C\times C$ blocks indexed by the $K+2$ local fields. For $\boldsymbol{f} \in \mathcal{F} = \{\boldsymbol{h}, \boldsymbol{h}^+, \boldsymbol{h}_1^-, \dots, \boldsymbol{h}_K^-\}$, define

$$\boldsymbol{\mathcal{X}}_{\boldsymbol{f}} = \boldsymbol{B}_{\boldsymbol{f}\boldsymbol{f}} \in \mathbb{R}^{C\times C}. \tag{168}$$

The scalar diagnostic used below is

$$\lambda_{\mathrm{AT}} \equiv \frac{1}{\chi^2} - \frac{\alpha(1-\gamma)^2}{C} \mathbb{E}\left[\sum_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{\mathcal{X}}_{\boldsymbol{f}}\|_F^2 \right], \tag{169}$$

where the expectation averages over the sampled class labels and Gaussian field centers. Equivalently,

$$\begin{aligned} \rho_{\mathrm{AT}} &= \frac{\alpha(1-\gamma)^2\chi^2}{C} \mathbb{E}\left[\sum_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{\mathcal{X}}_{\boldsymbol{f}}\|_F^2 \right], \\ \lambda_{\mathrm{AT}} &= \frac{1 - \rho_{\mathrm{AT}}}{\chi^2}. \end{aligned} \tag{170}$$

Although the full local Hessian enters the matrix inverse, the final norm includes only the same-field blocks. We therefore treat this quantity as a scalar response diagnostic rather than a complete replicon-stability criterion. Its threshold is $\rho_{\mathrm{AT}} = 1$, equivalently $\lambda_{\mathrm{AT}} = 0$.

For $M = C = 10$, $K = 5$, $\kappa = 15$, $\lambda_w = 0.01$, and $\tau = 0.1$, all sampled branches satisfy $\rho_{\mathrm{AT}} < 1$ and $\lambda_{\mathrm{AT}} > 0$ under this diagnostic (Table 6). At $\alpha = 0.174$, near the smooth-ReLU free-energy crossing, the forward and backward branches have $\rho_{\mathrm{AT}} = 0.740$ and 0.445, respectively. No threshold crossing is detected over the sampled ranges.