EXPLANATION SELECTION IN ABDUCTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Abductive learning (ABL) offers a promising framework that connects machine learning and logical reasoning in a mutually beneficial loop. However, this bidirectional feedback can also reinforce incorrect concept semantics, since different interpretations of intermediate concepts may explain the same observed reasoning outcomes. Static analyses characterize this ambiguity but leave open which explanation training selects and whether its labels can still be corrected. We develop a dynamical analysis of a fixed-feature binary ABL model that links the local stability of competing explanations to differences in their concept labels. We prove that Gram matrices with identical spectra can select different explanations even under the same supervision, objective, and initialization. We further identify conditions under which a selected explanation is guaranteed to persist despite subsequent changes in optimization geometry. Changing its labels would require crossing a region with lower probability of satisfying all constraints, creating an evidence barrier. Even when the correct explanation is the only realizable one, the updates can converge to a finite, strictly suboptimal state with incorrect concept labels. Experiments on visual ABL tasks support this distinction between selecting among compatible explanations and recovering a uniquely realizable correct explanation.

1 INTRODUCTION

Unifying machine learning and logical reasoning is a central ambition of artificial intelligence (Marra et al., 2024). Abductive learning (ABL) offers a promising framework for this goal through a reciprocal interaction between machine learning and logical reasoning (Zhou, 2019; Zhou & Huang, 2022). In ABL, the learner first predicts concept labels from data, and logical rules connect these labels to observed reasoning outcomes. Abductive reasoning then uses those outcomes to revise the predictions, producing labels that supervise the next round of concept learning. Repeating this cycle allows domain knowledge to shape concept learning without requiring a differentiable reasoner.

Theoretical studies have exposed a fundamental ambiguity in learning concepts from reasoning outcomes. The same rules and observed outcomes may admit several complete sets of concept labels for the training data, each constituting a *compatible explanation*. Reproducing the observed outcomes with incorrect concept labels gives rise to *reasoning shortcuts* (Marconato et al., 2023). Existing theoretical analyses characterize the generalization risk associated with reasoning shortcuts (Yang et al., 2024), establish conditions for concept identifiability (Bortolotti et al., 2025), and describe the geometry of model predictions that satisfy logical constraints (van Krieken et al., 2024). We also review how logical constraints supply supervision for concept learning, when the correct concept labels are identifiable, and how optimization geometry influences solution selection (Appendix A).

Knowing which explanations are compatible with supervision leaves open which one, if any, a training trajectory will reach. Because each reasoning step depends on the current predictions, the feedback can progressively favor one explanation over equally compatible alternatives. A dynamical account must explain which explanations attract nearby states, how optimization selects among them from a common initialization, and when subsequent updates must preserve the selected labels.

We analyze these questions in a fixed-feature binary ABL model based on the formal product-posterior update introduced in A^{3}BL (He et al., 2024). At each iteration, we compute an exact posterior over all compatible explanations and use its marginal label probabilities as fixed soft targets for one full-batch cross-entropy update. The resulting dynamics follow preconditioned gradient ascent on a common potential derived from the *legal evidence*, the probability that the model's predicted labels jointly satisfy all constraints. Local analysis then yields stability criteria based on the number and overlap of label differences between compatible explanations. Several explanations can be locally attracting, so

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.

stability alone does not determine which one training selects. That choice can depend on the Gram matrix, which maps the potential gradient to an update direction for the model's concept predictions. We isolate orientation by holding the Gram eigenvalues fixed and varying the eigenvectors. Such Grams can select arbitrarily many legal explanations from a common open set of initializations.

This sensitivity to optimization geometry need not persist throughout training. We establish a sufficient *capture* certificate in terms of the model's current legal evidence. Once the certificate holds, any change in the predicted labels would require crossing a region of lower legal evidence. Our spectral and step-size bounds prevent such a crossing and guarantee convergence to the selected explanation, even when the Gram matrix is chosen adaptively. Training can therefore entrench an explanation with incorrect concept labels, placing a limit on what changes in optimization alone can correct. This raises a natural question: if these incorrect alternatives are ruled out and only the correct legal explanation is realizable, is recovery guaranteed? We show, however, that native updates can still converge to finite, strictly suboptimal states with incorrect concept labels.

Our contributions are summarized as follows:

- **Local stability.** We characterize local asymptotic stability in terms of label differences among compatible explanations and establish convergence rates for the discrete updates (Sections 2–3).
- **Explanation selection.** We show that Grams with identical spectra can robustly select arbitrarily many distinct locally attracting explanations from a common open set of initializations (Section 4).
- **Certified capture.** We derive a legal-evidence certificate for label preservation and convergence under admissible updates. Furthermore, finite suboptimal attractors can prevent capture even when the correct explanation is the only legal one realizable by the shared predictor (Section 5).

Section 6 examines these dynamics in visual ABL tasks, showing how explanation selection, persistence under later interventions, and finite suboptimal attraction shape the learned concept labels.

2 Training Dynamics and a Common Potential

We study a soft-label formulation of ABL over propositional variables with a fixed feature representation. The binary concept labels are unknown, and background logical rules relate them to the observed reasoning outcomes. For N concept labels, let $\mathcal{C} \subseteq \{\pm 1\}^N$ be the finite set of *complete codes*, the full label assignments compatible with these rules and observations. We encode the intended semantic label on each coordinate as $+1$ and orient the margins $y \in \mathbb{R}^N$ so that $y_i > 0$ favors that label. We assume $\mathbf{1}_N \in \mathcal{C}$ and $|\mathcal{C}| \geq 2$, so the correct explanation is legal but has at least one alternative. The semantic labels are used only for analysis and are unavailable during training.

2.1 Legal evidence and posterior refresh

To form soft targets, we condition the model's independent label predictions on joint satisfaction of the logical constraints (He et al., 2024). With $\sigma(s) = (1 + e^{-s})^{-1}$, the code likelihood, legal evidence and candidate posterior are

$$L_z(y) = \prod_{i=1}^{N} \sigma(z_i y_i), \qquad \mathcal{E}(y) = \sum_{z \in \mathcal{C}} L_z(y), \qquad q_z(y) = \frac{L_z(y)}{\mathcal{E}(y)}. \tag{1}$$

The *legal evidence* $\mathcal{E}$ is the probability that the predicted labels jointly satisfy all constraints. The posterior q normalizes likelihoods directly, corresponding to uniform priors and unit temperature. Its marginals $p_i(y) = \Pr_q(z_i = +1)$ sum the posterior weights of codes assigning $+1$ to coordinate i.

At each iteration, we refresh q at the current margins and use its marginals as soft targets for one full-batch cross-entropy update, holding them fixed during differentiation. The sample Gram matrix K maps the residual R between targets and predictions to the margin update:

$$y_{t+1} = y_t + \eta K R(y_t), \qquad R_i(y) = 2[p_i(y) - \sigma(y_i)]. \tag{2}$$

Here $\eta > 0$ absorbs the loss normalization, and the factor two fixes our residual convention. This posterior and update define the *native* model used unless stated otherwise. For a fixed feature matrix $X \in \mathbb{R}^{N \times d}$ and a shared linear head $w \in \mathbb{R}^d$, parameter updates with a positive-definite preconditioner M induce the following margin dynamics:

$$y = Xw, \qquad w^+ = w + \eta M X^\top R(y), \qquad K_M = X M X^\top. \tag{3}$$

The parameter metric M controls the induced sample Gram matrix K_M for a fixed representation. Ordinary parameter gradient descent uses $M = I$ and hence $K_M = XX^\top$, which generally differs from $K = I$ in sample space. We first analyze positive-definite Gram matrices $K \succ 0$, each realizable by $X = K^{1/2}$. Corollary 1 extends capture to arbitrary-rank X, where K_M may be singular.

2.2 What remains fixed when the Gram changes

With exact posterior refresh, the residual of the soft-label update is the gradient of a common potential:

$$\Phi(y) = 2 \log \mathcal{E}(y), \qquad \nabla \Phi(y) = R(y). \tag{4}$$

This potential is a rescaled negative semantic loss (Xu et al., 2018). The Gram determines the ascent direction but, for $K \succ 0$, leaves finite fixed points unchanged because $KR = 0$ if and only if $R = 0$. A singular Gram may also map a nonzero residual to zero, creating additional fixed points.

For $K \succ 0$, ascent on the common potential is guaranteed by the sufficient step-size condition

$$\eta \lambda_{\max}(K)(N + 1) < 4. \tag{5}$$

Under this condition, the smoothness bound in Appendix B gives $\Phi(y_{t+1}) > \Phi(y_t)$ whenever $R(y_t) \neq 0$. We assume (5) for all fixed-Gram updates unless stated otherwise.

3 Which Explanations Are Locally Stable?

A trajectory approaches the *code endpoint* z when $z_i y_{i,t} \to +\infty$ for every coordinate, so its independent label predictions concentrate on the complete assignment z. Local stability depends on which concept labels differ between a code and its competitors. Fix $z \in \mathcal{C}$ and call every other code a *rival*. Write $d_{\min}(z) = \min_{w \neq z} d_{\mathrm{H}}(w, z)$, where Hamming distance counts the labels on which two codes differ. To analyze stability at infinite signed margins, we express the dynamics in compactified coordinates that place this endpoint at a finite boundary point:

$$a_i = e^{-z_i y_i}, \qquad r = \sum_i a_i. \tag{6}$$

a_i is the ratio of the model probabilities for labels $-z_i$ and z_i, so small a_i means confident agreement with z_i. The endpoint lies at $a = 0$, with r its ℓ_1 distance. A coordinate $a_i = 0$ represents a prediction already saturated at label z_i. The closed orthant includes these saturated predictions, so stability must cover perturbations on its boundary faces as well as states with finite logits. Formally, the endpoint is *locally asymptotically stable* if the continuous extension of the update to $[0, \infty)^N$ is Lyapunov stable at 0 and attracts a relative neighborhood. All proofs are in Appendix C.

3.1 Local stability in the native model

Theorem 1 (Native local endpoints)**.** *Use uniform priors, unit temperature and no direct semantic supervision. For every fixed $K \succ 0$ and $\eta > 0$ satisfying (5), a code z is locally asymptotically stable in the closed compactified orthant if and only if $d_{\min}(z) \geq 2$. In the stable case, a sufficiently small neighborhood is trapping and $r_t = O(1/t)$.*

The distance condition identifies when logical conditioning can resolve uncertainty in one label. If flipping that label yields another legal code, either value satisfies the constraints when all other labels are fixed. Conditioning leaves the prediction unchanged, producing a flat zero-loss edge between the codes (van Krieken et al., 2024). Without a restoring signal along this edge, the endpoint cannot attract a whole neighborhood, although finite-logit trajectories may still approach it (Appendix C.1).

Without a one-bit rival, a single-label deviation is illegal, and every competing legal assignment requires several deviations whose probabilities multiply. Near the endpoint, most probability lost from the current code therefore goes to illegal assignments. Conditioning removes this mass and reinforces the code's labels, whether or not they are semantically correct. The Gram mixes these corrections, so individual signed margins may temporarily decrease. Near the endpoint, the evidence deficit $1 - \mathcal{E}(y)$ equals r to first order. Theorem 1 shows that ascent of the common potential drives r to zero under every admissible fixed positive-definite Gram. This attraction persists under time-varying posterior-refresh errors if their effect on the residual remains sufficiently small relative to the exact residual, with the same $O(1/t)$ rate under a tighter step-size bound (details in Appendix C.10).

3.2 WHY DISTANCE, TEMPERATURE AND OVERLAP MATTER

Priors and temperature modify how the posterior distributes mass between a code and its rivals. Allow positive priors π_z and temperature $\tau > 0$, and set $q_z(y) \propto [\pi_z L_z(y)]^{1/\tau}$. We use the resulting marginals in the same detached update, with no direct semantic supervision. The corresponding potential is $\Phi(y) = 2\tau \log \sum_{z \in \mathcal{C}} [\pi_z L_z(y)]^{1/\tau}$. For a rival w, the *defect set* $D(w, z) = \{i : w_i \neq z_i\}$ records the coordinates on which it differs from z. In the posterior ratio, the likelihood factors for agreeing labels cancel, leaving only the prior ratio and the predictions on the defect set:

$$\frac{q_w(y)}{q_z(y)} = \left(\frac{\pi_w}{\pi_z}\right)^{1/\tau} \prod_{i \in D(w,z)} a_i^{1/\tau}. \tag{7}$$

As confidence in z increases, each differing label contributes another vanishing factor to a rival's posterior odds. Higher temperature reduces these exponents, slowing the loss of rival support, whereas priors change only the coefficients. Since the code's confidence loss is first order in r, $d_{\min}(z)/\tau$ determines whether rival support vanishes faster or slower than this loss.

At $d_{\min}(z) = \tau$, rival support and the loss of confidence in z have the same order. Stability now depends on how uncertainty is distributed across labels: uncertainty on a shared coordinate can support several rivals at once. Write $a = ru$, where $u \in \Delta^{N-1}$ is nonnegative and sums to one. Its entries allocate the total r across coordinates, including boundary directions. Define

$$H_z(u) = \tau \sum_{\substack{w \neq z \\ d_{\mathrm{H}}(w,z) = \tau}} \left(\frac{\pi_w}{\pi_z}\right)^{1/\tau} \prod_{i \in D(w,z)} u_i^{1/\tau}, \qquad \chi_z = \max_{u \in \Delta^{N-1}} H_z(u). \tag{8}$$

Here $H_z(u)$ sums the nearest rivals' contributions, and χ_z maximizes this sum over allocations. The unit threshold compares this rival support with the current code's first-order confidence loss. χ_z can be computed by concave maximization over the simplex (Appendix C.6). Direct semantic supervision increases the restoring coefficient for the correct code but leaves the rival exponents unchanged, shifting the critical balance without changing the distance–temperature comparison (Corollary 2).

Proposition 2 (Local classification with priors and temperature)**.** *With positive priors, no direct semantic supervision, and $\eta \lambda_{\max}(K)(N/\tau + 1) < 4$, local asymptotic stability holds precisely when*

$$d_{\min}(z) > \tau, \qquad \text{or} \qquad d_{\min}(z) = \tau \ \text{ and } \ \chi_z < 1,$$

apart from the equality case $d_{\min}(z) = \tau$, $\chi_z = 1$. Every stable case has a trapping neighborhood and $r_t = O(1/t)$. The equality case requires higher-order or exact analysis.

Example 1 (Rivals can jointly defeat local attraction). Let $N = 4$, $z = (+, +, +, +)$, $\tau = 2$, and give two distance-two rivals the same prior ratio $\pi_w/\pi_z = 0.64$. Either rival alone leaves z locally attracting. With disjoint defect sets $\{1, 2\}$ and $\{3, 4\}$, uncertainty that supports one rival cannot support the other, and attraction persists. Changing the second defect set to $\{1, 3\}$ lets uncertainty in coordinate 1 support both rivals at once, and z loses local attraction (see Appendix C.6 for details).

4 SAME SPECTRUM, DIFFERENT EXPLANATIONS

We have proved that several explanations can be locally attracting. In this section, we study how Gram orientation affects selection among them while holding the spectra fixed. Proofs are in Appendix D.

4.1 A SINGLE SLOW DIRECTION INSIDE THE CANDIDATE SUBSPACE

For any $m \geq 2$, partition $N = 2m$ coordinates into pairs and let $\mathcal{C}_m = \{z_0, \ldots, z_m\}$, where $z_0 = \mathbf{1}_{2m}$ and z_j flips only pair j. In the native model, each code has minimum rival distance two and is locally asymptotically stable by Theorem 1. All codes give the two coordinates of a pair the same label, so their relative posterior weights depend on the sum of its logits. Equal and opposite logit changes leave these weights unchanged. Equal changes within each pair follow the directions

$$c_j = (e_{2j-1} + e_{2j})/\sqrt{2}, \qquad B = [c_1, \ldots, c_m].$$

Here e_i is the ith standard basis vector. More generally, relative posterior weights depend only on logits projected onto the *candidate-active subspace* $S_{\mathcal{C}} = \operatorname{span}\{z - w : z, w \in \mathcal{C}\}$. For $\mathcal{C}_m$, $B^\top K B$ is the Gram in the orthonormal basis B, and its eigenvalues form the active spectrum.

Choose a slow eigenvalue $\alpha \in (0,1)$ and a unit vector $v \in \mathbb{R}^m$ orthogonal to $\mathbf{1}_m$, and set

$$H_0 = I_m - (1-\alpha)vv^\top, \qquad H_j = I_m - (1-\alpha)e_j e_j^\top \quad (1 \le j \le m),$$
$$K_j = BH_jB^\top + I_{2m} - BB^\top, \qquad j = 0, \dots, m. \tag{9}$$

Each K_j acts as H_j on the candidate-active subspace and as the identity on its orthogonal complement. The slow direction stays inside this subspace, so rotating it preserves both the full and active spectra. Consider the common symmetric initialization $y_0 = s_0\mathbf{1}_{2m}$ with $s_0 < 0$. Under K_0, the slow direction is orthogonal to this initialization, so all pairs move together toward positive labels. Under K_j, pair j initially moves toward positive labels too, but remains negative while the faster pairs cross zero. Rivals requiring negative labels on those other pairs lose posterior weight, whereas z_j remains supported by pair j. Once the other pairs are confidently positive, z_0 is the main remaining competitor, and the target pair's negative margins favor z_j. With the rate separation in Theorem 3, the resulting feedback reverses the target pair's update, reinforcing its negative labels. Gram orientation thus changes which explanation the inferred targets reinforce, rather than merely the speed of learning.

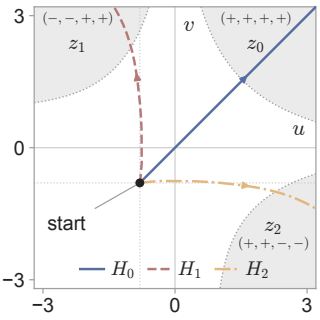


Figure 1: Isospectral selection. Gray shading marks $\mathcal{E} > 1/2$.

Figure 1 visualizes the case $N = 4$ ($m = 2$) with $\alpha = 1/6$, $\eta = 0.02$, and $y_0 = -0.8\mathbf{1}_4$. The trajectories follow the reduced H_j dynamics in the plane $y = (u, u, v, v)$, so the two axes track the common margin in each pair. The posterior refresh and all three Grams preserve equality within each pair. The Grams share full spectrum $\{1/6, 1, 1, 1\}$ and active spectrum $\{1/6, 1\}$. Each trajectory enters a different shaded region, meeting the capture criterion that will be introduced in Section 5.

4.2 Robust routing from a common open set

The preceding analysis starts from a symmetric initial state. We now show that this symmetry is not essential: each Gram retains its selected endpoint under small perturbations of both the initial predictions and the matrix itself. Below, $\mathbb{S}^{2m}_{++}$ denotes the cone of symmetric positive-definite $2m \times 2m$ matrices, and bracketed superscripts record eigenvalue multiplicities.

Theorem 3 (Robust full- and active-isospectral routing)**.** *For every $m \ge 2$ and $s_0 < 0$, there exist $\alpha_m \in (0,1)$ and a common $\eta_m > 0$ such that the Grams in (9) satisfy the ascent condition and*

$$\operatorname{spec}(K_j) = \{\alpha_m, 1^{[2m-1]}\}, \qquad \operatorname{spec}(B^\top K_j B) = \{\alpha_m, 1^{[m-1]}\}. \tag{10}$$

There are a common open neighborhood $\mathcal{U}_m \subset \mathbb{R}^{2m}$ of $y_\star = (s_0, \dots, s_0)$ and open Gram neighborhoods $\mathcal{V}_{m,j} \subset \mathbb{S}^{2m}_{++}$ of K_j such that

$$y_0 \in \mathcal{U}_m, \quad K \in \mathcal{V}_{m,j} \quad \Longrightarrow \quad F_K^t(y_0) \to z_j, \qquad j = 0, \dots, m, \tag{11}$$

where $F_K(y) = y + \eta_m K R(y)$ and convergence is in the endpoint sense.

The open neighborhoods allow all initial margins to vary, so exact symmetry is unnecessary. Each route enters a region that attracts its selected code, and small perturbations of the preceding trajectory preserve this outcome. The proof in Appendix D makes this argument uniform over nearby initial states and Gram matrices. General Gram perturbations may change the spectra, but small rotations of the slow direction within the active subspace preserve both spectra and the endpoint (Appendix D.8).

5 Capture and the Limits of Recovery

Section 4 shows that Gram orientation can change the selected explanation. But can later Gram changes correct a mistaken choice? We study when labels become fixed and whether excluding incorrect legal alternatives from the shared model ensures recovery. All proofs are in Appendix E.

5.1 Certifying the Selected Explanation

For native updates, legal evidence above one half at a finite state certifies label preservation and endpoint convergence under every subsequent Gram sequence satisfying the spectral and step-size bounds below. We assume that candidate codes differ pairwise in at least two labels.

Theorem 4 (Half-evidence capture). *Let $|\mathcal{C}| \geq 2$ and $d_{\mathrm{H}}(z, w) \geq 2$ for all distinct $z, w \in \mathcal{C}$. Use the native exact residual. After any finite prefix, allow the continuation*

$$y_{t+1} = y_t + \eta K_t R(y_t), \qquad 0 < \kappa_- I \preceq K_t \preceq \kappa_+ I, \qquad \eta\kappa_+(N+1) < 4, \tag{12}$$

with fixed $\eta > 0$ and $0 < \kappa_- \leq \kappa_+ < \infty$. If the finite state y_T satisfies

$$\mathcal{E}(y_T) > \tfrac{1}{2}, \tag{13}$$

then $z = \operatorname{sign}(y_T)$ is well defined and belongs to $\mathcal{C}$. Every continuation obeying (12) preserves $\operatorname{sign}(y_t) = z$ for $t \geq T$, and

$$r_t = \sum_i e^{-z_i y_{i,t}} = O((t - T + 1)^{-1}).$$

The Gram matrices K_t may depend arbitrarily on the training history. Conversely, every trajectory approaching a code endpoint eventually satisfies (13).

Along a continuous path, a hard label changes only when its logit crosses zero, where both labels are equally likely. With the other labels fixed, at most one choice can complete a legal code, since legal codes differ in at least two coordinates. Averaging over the other labels therefore gives $\mathcal{E} \leq 1/2$ at every such tie. These ties form an *evidence barrier* between different hard codes.

Applying this argument to discrete updates requires more than an increase in evidence at successive iterates, since a step could jump across a label tie. Under the step-size bound, evidence throughout each update segment stays at or above its starting value. An update from $\mathcal{E} > 1/2$ therefore preserves the selected labels, and the spectral bounds ensure convergence to the corresponding endpoint. In particular, an admissible Gram that selects another explanation from the initial state must preserve the captured code when introduced later (Appendix E.1). This guarantee extends to controlled errors in posterior updates under the same spectral bounds and a tighter step-size bound (Appendix E.2).

Theorem 4 requires positive-definite sample Grams. For a shared head on fixed features, $K_t = XM_tX^\top$ is singular whenever X lacks full row rank. The head then cannot choose each sample's prediction freely, because some predictions are determined by the others. The following corollary establishes capture under this restriction, with spectral bounds on the parameter metrics.

Corollary 1 (Capture with limited control over individual predictions). *Under the candidate and native-update assumptions of Theorem 4, let $X \in \mathbb{R}^{N\times d}$ have arbitrary rank. From a finite head $w_T \in \mathbb{R}^d$, use fixed $\eta > 0$, constants $0 < m_- \leq m_+ < \infty$, and*

$$y_t = Xw_t, \qquad w_{t+1} = w_t + \eta M_t X^\top R(Xw_t), \qquad 0 < m_- I_d \preceq M_t \preceq m_+ I_d. \tag{14}$$

If $\eta m_+ \|X\|_{\mathrm{op}}^2 (N+1) < 4$, then $\mathcal{E}(Xw_T) > 1/2$ implies the same code preservation and $O((t - T + 1)^{-1})$ compactified rate for every history-dependent parameter-metric continuation.

For a shared head, the residuals from all samples enter the parameter gradient $X^\top R$, where opposing contributions can cancel and halt training. Capture prevents this cancellation: the head w_T already separates the selected labels, so moving along it increases all their signed margins. Since subsequent residuals favor these labels, each sample contributes positively to the gradient along w_T, keeping the parameter gradient nonzero. The metric bounds then ensure continued progress toward the selected endpoint (Appendix E.3). This guarantee includes ordinary full-batch parameter gradient descent ($M_t = I$), even when the selected labels are semantically incorrect.

5.2 Identifiability Does Not Ensure Recovery

Capture preserves the selected explanation under admissible updates, even when its concept labels are incorrect. A natural question, then, is whether ruling out these incorrect legal alternatives guarantees recovery. The following proposition shows, however, that recovery can still fail: training may converge to a finite suboptimal state whose predicted labels violate the constraints.

Proposition 5 (A suboptimal attracting state). *Let $\mathcal{C}_X = \{z \in \mathcal{C} : \exists w,\ \min_i z_i(Xw)_i > 0\}$ denote the legal codes realizable by the shared predictor Xw. There exist X and a candidate set $\mathcal{C}$ with pairwise Hamming distance at least two and $\mathcal{C}_X = \{\mathbf{1}_N\}$, for which ordinary native updates $w^+ = w + \eta X^\top R(Xw)$ admit a finite, strictly suboptimal, locally asymptotically stable fixed point for every sufficiently small $\eta > 0$. Its hard predictions violate some constraints, and $\mathcal{E}(Xw) < 1/2$.*

To explain Proposition 5, consider a binary classification task where each training example gives a pair of inputs and their label sum. The two inputs x_0 and x_1 have true labels 0 and 1. The data contain (x_0, x_0) with sum zero, (x_1, x_1) with sum two, and mixed pairs (x_0, x_1) with sum one. Zero forces both copies of x_0 to have label 0, while two forces both copies of x_1 to have label 1. Since the classifier uses one logit per input across all pairs, only the correct legal code is realizable.

The posterior, however, uses a separate label variable for each input occurrence and enforces only the observed sums. A candidate can thus label mixed pairs as $(1, 0)$ and the others correctly, satisfying every sum while giving the same input conflicting labels. If the classifier favors reversed labels, mixed pairs supply targets that reinforce them, while the other pairs supply corrective targets. As the classifier's probabilities for the reversed labels approach one, the mixed pairs' targets nearly match its predictions, weakening reinforcement while correction persists. With enough mixed pairs, these signals can balance at finite confidence, forming a strict local minimum of the negative log evidence that attracts nearby updates back to the wrong predictions. Appendix E.4 extends this example to multiple inputs, and Section 6.3 verifies these wrong attractors on CIFAR-10.

6 Empirical Evaluation

Fitting logical supervision need not recover the correct concept labels. We examine how update geometry selects an explanation and when later metric changes can no longer revise it. We then test whether recovery can still fail when supervision uniquely determines the correct labels.[1]

6.1 Different explanations under the same counts

We consider the task of distinguishing birds from cats in CIFAR-10 images (Krizhevsky, 2009). Training images are grouped into bags, and the learner receives only the number of cats in each bag (Baručić & Kybic, 2022). Each training set contains 24 sampled images in six disjoint bags of four, with cat counts $(1, 2, 3, 1, 2, 3)$. These counts admit $\left[\binom{4}{1}\binom{4}{2}\binom{4}{3}\right]^2 = 9{,}216$ complete label assignments. Any two legal assignments differ in at least two labels, so all satisfy the local-attraction criterion of Theorem 1. The one-cat and three-cat bags exclude a global swap of the class labels.

We train a shared linear head on frozen ImageNet-supervised ResNet-18 features (He et al., 2016). Each update follows Section 2 with refreshed exact posterior targets. We compare four fixed, positive-definite Grams (feature, feature-reverse, random 1 and random 2) with identical full and candidate-active spectra, using a common step size satisfying the ascent condition. We construct these Grams from input features, bag counts, and fixed randomness, without per-image labels or a target explanation. Appendix F.1 details the construction and verifies these conditions.

We use six disjoint training sets with four random initial heads per set, giving $6 \times 4 = 24$ comparison groups. Within each group, we train four copies of the same initial head on the same images and counts, changing only the Gram orientation. Each training set uses the same four Grams for all initial heads. In every group, the four Grams select distinct complete label assignments. All $24 \times 4 = 96$ runs satisfy every count, assign high probability to the counts, and retain confident, stable image labels over the final evaluation window. Figure 2(a) shows two runs from one group: both satisfy every bag count, yet label different images as cats. Full results are in Appendix F.2.

Within each group, we compare the four final assignments in all $\binom{4}{2} = 6$ pairs. This gives $24 \times 6 = 144$ pairs of runs that share the images and initial head but use different Gram orientations. The assignments differ on a median of 10 of the 24 images per pair (range: 2–14, Figure 2(b)). Thus, matching both spectra does not fix the labels learned from the same counts.

6.2 Certifying the selected explanations in the same learner

The preceding experiment shows that changing the Gram at initialization can select different label assignments. We now test whether the same change can alter labels after capture. For the six disjoint bags, the legal evidence $\mathcal{E}_{\text{all}}(y) = \prod_{b=1}^{6} \mathcal{E}_b(y_b)$ is the model's probability of satisfying all six counts jointly. Here $\mathcal{E}_b(y_b)$ is the probability of satisfying bag b's count. All 96 fitted runs cross $\mathcal{E}_{\text{all}} > 1/2$ (Figure 2(c)). Theorem 4 therefore certifies label preservation and convergence to the selected endpoint under every continuation satisfying its spectral and step-size bounds.

[1] Code is available at: https://anonymous.4open.science/r/abl-selection.

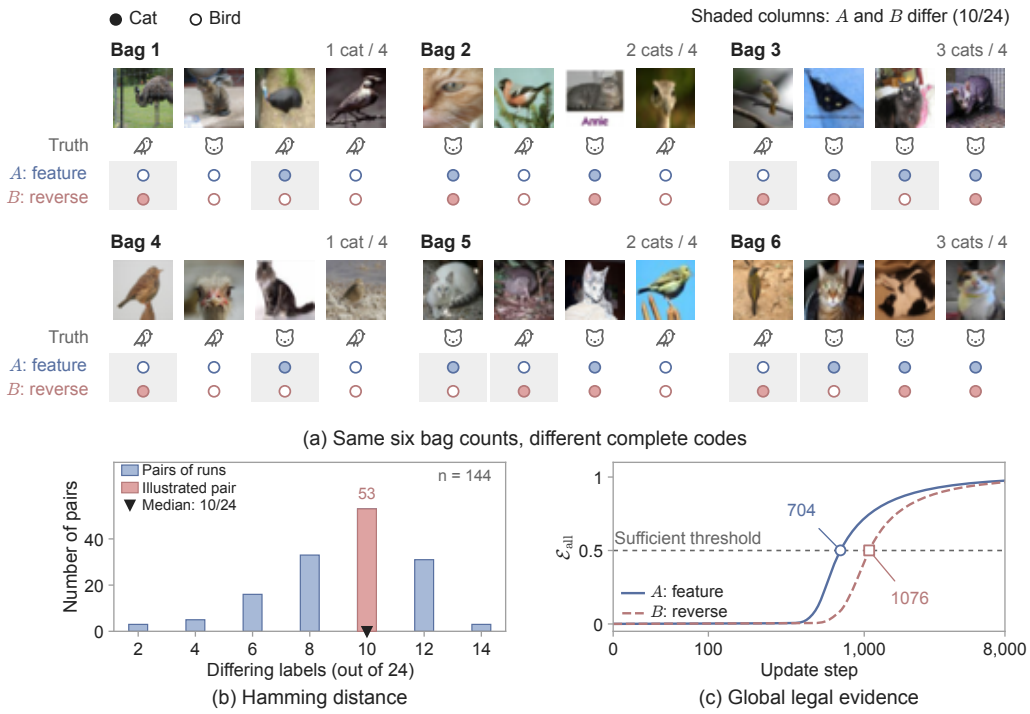


Figure 2: Gram matrices with matched spectra select different captured assignments under the same CIFAR-10 bag counts. (a) Cat/bird icons give ground truth, and filled/open circles are cat/bird in the two learned codes. Shaded columns mark the 10 labels on which the two codes differ. (b) Number of differing labels across 144 pairs of runs with the same data and initial head but different Gram orientations. (c) Both runs cross the global evidence threshold $1/2$, certifying capture. The first crossing is marked, on a linear step axis through 100 and a logarithmic axis thereafter.

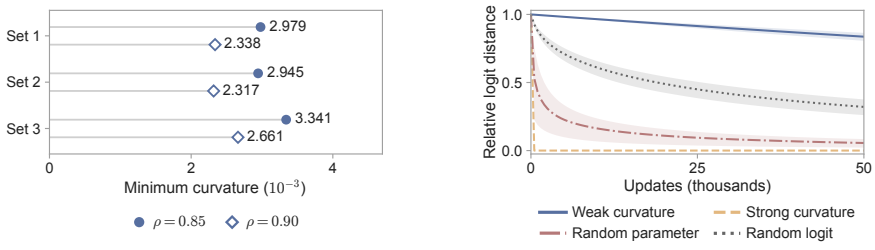
At each run's first crossing, we save the head and train three copies for a further 4,000 updates. One copy switches to a Gram that selected a different assignment from the same initial head. The other two use metrics chosen to slow confidence growth on the model's least certain captured label, with one fixed at capture and the other recomputed at each update. All three interventions obey the theorem's bounds, although only the swap preserves the original spectrum. All $96 \times 3 = 288$ continuations preserve every captured label. The swap thus shows that the same Gram can select one assignment from initialization yet preserve another after capture. Full results are in Appendix F.3.

Capture with a restricted shared head. The preceding head can realize any combination of prediction probabilities on its 24 training images. To test whether capture requires this flexibility, we retain the frozen encoder but restrict the head to 33 trainable parameters for 96 images in 24 bags. We use four disjoint training sets and four random initial heads per set, giving $4 \times 4 = 16$ comparison groups. Each group trains three copies of the same initial head on the same data, using ordinary parameter gradient descent or one of two fixed parameter metrics. All $16 \times 3 = 48$ runs fit the counts and cross the half-evidence threshold for all 24 bags jointly.

From each captured head, we run three 4,000-update continuations using ordinary gradient descent, a fixed metric, or an adaptive metric chosen to slow confidence growth on the least certain captured label. All $48 \times 3 = 144$ continuations preserve every captured label. The disjoint-bag refinement of Corollary 1 certifies all 48 runs, although the head cannot freely set all training predictions. The parameter metrics meet its bounds, while the sample Grams are singular with different spectra. Capture also preserves concept errors: 93 of the first experiment's 96 assignments and all 48 restricted-head assignments mislabel at least one training image. Full results are in Appendix F.4.

6.3 Suboptimal attraction under identifiable counts

The preceding experiments allow several label assignments to satisfy the same counts. We now test whether training can remain trapped at incorrect predictions when the correct assignment is uniquely determined and realizable, as predicted by Proposition 5.



(a) Local curvature (b) Return to wrong predictions

Figure 3: Wrong attractors despite identifiable counts. (a) Minimum curvature at six stationary heads that mislabel every image, over directions that change predictions. (b) RMS logit distance from these wrong predictions after perturbation, relative to its initial value 0.01. Lines and bands show medians and full ranges over three training sets and two loss weights for each perturbation type.

Each training set contains 256 CIFAR-10 images, with 128 birds and 128 cats. We form pairs and supervise the head with the number of cats in each pair. Every image appears in one same-class pair and eight mixed-class pairs. A count of zero identifies both images as birds, while a count of two identifies both as cats. Because these same-class pairs cover every image, only the correct labels satisfy all counts. The shared linear head can realize this assignment on frozen ResNet-18 features.

To understand how errors can persist, consider the targets supplied by these pairs. The head gives each image the same prediction wherever it appears, but posterior targets are computed separately for each pair. In a mixed pair, the count of one permits either image to be the cat. If the head favors reversed labels, the posterior favors that ordering and supplies targets that reinforce the error. The same-class pairs supply corrective targets for those same images. The mean mixed-pair loss receives weight $\rho \in \{0.85, 0.9\}$, and the mean same-class loss receives weight $1 - \rho$.

Using the true labels, we construct $3 \times 2 = 6$ heads, one for each of three training sets and each weight. Each head mislabels every image, assigning probability 0.8 to the wrong label at $\rho = 0.85$ and approximately 0.885 at $\rho = 0.9$. Reinforcement and correction cancel at these heads, making them stationary. We use them as centers for our local tests. Figure 3(a) shows positive minimum curvature at all six centers: any sufficiently small change to their predictions increases the loss. The wrong predictions therefore form local loss minima, despite violating every same-class count.

We then perturb each center in four ways: along the directions in which the loss rises least or most sharply, and randomly in the parameters or logits. All four perturbations start at the same root-mean-square (RMS) logit distance of 0.01. The $6 \times 4 = 24$ heads receive 50,000 ordinary parameter updates, refreshing posterior targets at every step. Figure 3(b) plots the distance from each center, so a falling curve means a return toward the wrong predictions. Every trajectory moves toward its center while retaining all wrong labels, most slowly along the weakest-curvature direction.

For comparison, we train 12 random initial heads per set and weight on four training sets, including the three above, with $\rho \in \{0.5, 0.7, 0.85, 0.9\}$. All $4 \times 12 \times 4 = 192$ runs recover every training label after 30,000 updates. Correct recovery from these random starts thus coexists with local attraction toward wrong predictions. Uniquely determined, realizable labels do not ensure that every training trajectory recovers them. Full results are in Appendix F.5.

7 Conclusion

Our dynamical analysis explains how ABL can select and entrench an incorrect concept interpretation while fitting its logical supervision. The pattern of differences in concept labels can make several legal explanations locally attracting. Among these alternatives, Gram orientation can change the selected explanation even under matched spectra and a common initialization. This influence depends on the training state: once capture is certified, subsequent changes in geometry within our bounds must preserve the selected labels. The experiments illustrate this limit: a metric that selects another explanation from initialization preserves the captured labels when applied later. Even if these incorrect alternatives are excluded so that only the correct legal explanation is realizable, finite suboptimal attractors can still prevent recovery. A limitation of this account is that it characterizes local attraction and capture within a fixed representation, leaving the outcome from an arbitrary initialization unresolved. In future work, we will explore how initial conditions shape full training trajectories and whether representation learning can enable recovery from persistent concept errors.

AI-USE DISCLOSURE

We used generative AI tools as assistants for theoretical analysis, experimental work, and manuscript preparation. Their assistance included literature search, formulation and refinement of mathematical claims, development and checking of proof arguments, experimental design and protocol refinement, construction of synthetic test cases, implementation and debugging of research and verification code, and interpretation of numerical results. These tools also assisted with manuscript polishing and the preparation of figures and tables. The authors evaluated the suggestions, reviewed mathematical arguments against their stated assumptions, and checked reported numerical results against experimental records. The authors made all final decisions on the research methodology, claims, and presentation and take full responsibility for the paper and its accompanying artifacts.

REPRODUCIBILITY STATEMENT

The model and assumptions are specified in Sections 2–5, with proofs provided in Appendices B–E. Experimental protocols, data preprocessing, and full results are documented in Appendix F. The anonymous supplementary repository linked in Section 6 contains the experiment and plotting code, precomputed features, experiment configurations, and reference results. Its README provides environment setup instructions and commands for reproducing the experiments and figures.

REFERENCES

Denis Baručić and Jan Kybic. Fast learning from label proportions with small bags. In *Proceedings of the 2022 IEEE International Conference on Image Processing*, pp. 3156–3160, 2022.

Silvia Bonettini, Ignace Loris, Federica Porta, and Marco Prato. Variable metric inexact line-search-based methods for nonsmooth optimization. *SIAM Journal on Optimization*, 26(2):891–921, 2016.

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. In *Advances in Neural Information Processing Systems*, pp. 173204–173243, 2025.

Yunuo Cen, Zixuan Wang, Jintao Zhang, Zhiwei Zhang, and Xuanyao Fong. FourierCSP: Differentiable constraint satisfaction problem solving by Walsh-Fourier expansion. arXiv preprint arXiv:2510.04480v2, 2026.

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *Advances in Neural Information Processing Systems*, pp. 21061–21071, 2020.

Joshua Cooper and Aaron Dutle. Spectra of uniform hypergraphs. *Linear Algebra and its Applications*, 436(9):3268–3292, 2012.

Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging machine learning and logical reasoning by abductive learning. In *Advances in Neural Information Processing Systems*, pp. 2815–2826, 2019.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

Bernold Fiedler, Mohamed Belhaq, and Mohamed Houssni. Basins of attraction in strongly damped coupled mechanical oscillators: A global example. *Zeitschrift für angewandte Mathematik und Physik*, 50(2):282–300, 1999.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1832–1841, 2018.

Hao-Yuan He and Ming Li. A learnability analysis on neuro-symbolic learning. In *Advances in Neural Information Processing Systems*, pp. 84923–84956, 2025.

Hao-Yuan He, Hui Sun, Zheng Xie, and Ming Li. Ambiguity-aware abductive learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 18019–18042, 2024.

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.

Wen-Chao Hu, Wang-Zhou Dai, Yuan Jiang, and Zhi-Hua Zhou. Efficient rectification of neuro-symbolic reasoning inconsistencies by abductive reflection. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pp. 17333–17341, 2025a.

Wen-Chao Hu, Qi-Jie Li, Lin-Han Jia, Cunjing Ge, Yu-Feng Li, Yuan Jiang, and Zhi-Hua Zhou. Curriculum abductive learning. In *Advances in Neural Information Processing Systems*, pp. 174455–174483, 2025b.

Lin-Han Jia, Si-Yu Han, Lan-Zhe Guo, Zhi Zhou, Zhao-Long Li, Yu-Feng Li, and Zhi-Hua Zhou. A smooth transition between induction and deduction: Fast abductive learning based on probabilistic symbol perception. arXiv preprint arXiv:2502.12919, 2025a.

Lin-Han Jia, Wen-Chao Hu, Jie-Jing Shao, Lan-Zhe Guo, and Yu-Feng Li. Verification learning: Make unsupervised neuro-symbolic system feasible. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 27292–27306, 2025b.

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

Thomas Lartigue, Stanley Durrleman, and Stéphanie Allassonnière. Deterministic approximate EM algorithm; application to the Riemann approximation EM and the tempered EM. *Algorithms*, 15 (3):78, 2022.

Harald Leisenberger, Christian Knoll, Richard Seeber, and Franz Pernkopf. Convergence behavior of belief propagation: Estimating regions of attraction via Lyapunov functions. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, pp. 1863–1873, 2021.

Jaron Maene, Vincent Derkinderen, and Luc De Raedt. On the hardness of probabilistic neurosymbolic learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34203–34218, 2024.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, pp. 3749–3759, 2018.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. In *Advances in Neural Information Processing Systems*, pp. 72507–72539, 2023.

Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, 328:104062, 2024.

Samet Oymak and Talha Cihad Gulcu. A theoretical characterization of semi-supervised learning with self-training for Gaussian mixture models. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pp. 3601–3609, 2021.

Hrithik Ravi, Clayton Scott, Daniel Soudry, and Yutong Wang. The implicit bias of gradient descent on separable multiclass data. In *Advances in Neural Information Processing Systems*, pp. 81324–81359, 2024.

Mitchell T. Scott, Tianshi Xu, Ziyuan Tang, Alexandra Pichette-Emmons, Qiang Ye, Yousef Saad, and Yuanzhe Xi. Design criteria for SGD preconditioners: Local conditioning, noise floors, and basin stability. *Transactions on Machine Learning Research*, 2026.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Akihiro Takemura, Katsumi Inoue, and Masaaki Nishino. Constraint-based analysis of reasoning shortcuts in neurosymbolic learning. In *Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning*, pp. 1132–1142, 2026.

Lue Tao, Yu-Xuan Huang, Wang-Zhou Dai, and Yuan Jiang. Deciphering raw data in neuro-symbolic learning with provable guarantees. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 15310–15318, 2024.

Efthymia Tsamoura, Kaifu Wang, and Dan Roth. Imbalances in neurosymbolic learning: Characterization and mitigating strategies. In *Advances in Neural Information Processing Systems*, pp. 152087–152123, 2025.

Emile van Krieken, Pasquale Minervini, Edoardo M. Ponti, and Antonio Vergari. On the independence assumption in neurosymbolic learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 49078–49097, 2024.

Jun-Kun Wang and Andre Wibisono. Towards understanding GD with hard and conjugate pseudo-labels for test-time adaptation. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Kaifu Wang, Efthymia Tsamoura, and Dan Roth. On learning latent models with multi-instance weak supervision. In *Advances in Neural Information Processing Systems*, pp. 9661–9694, 2023.

Wen-Da Wei, Xiao-Wen Yang, Jie-Jing Shao, and Lan-Zhe Guo. Curriculum abductive learning for mitigating reasoning shortcuts. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pp. 6533–6541, 2025.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, pp. 74229–74256, 2023.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5502–5511, 2018.

Xin Xu. Reasoning shortcuts and value symmetries: What symmetry permits, architecture realizes, and optimization selects. arXiv preprint arXiv:2608.10420v2, 2026.

Xiao-Wen Yang, Wen-Da Wei, Jie-Jing Shao, Yu-Feng Li, and Zhi-Hua Zhou. Analysis for abductive learning and neural-symbolic reasoning shortcuts. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 56524–56541, 2024.

Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. How unlabeled data improve generalization in self-training? A one-hidden-layer theoretical analysis. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Zhi-Hua Zhou. Abductive learning: Towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62(7):076101, 2019.

Zhi-Hua Zhou and Yu-Xuan Huang. Abductive learning. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pp. 353–369. IOS Press, 2022.

Zsolt Zombori, Agapi Rissaki, Kristóf Szabó, Wolfgang Gatterbauer, and Michael Benedikt. Towards unbiased exploration in partial label learning. *Journal of Machine Learning Research*, 25(412): 1–56, 2024.

A Related Work

Reasoning and label revision in ABL. ABL uses logical knowledge to revise perceptual predictions and guide further learning through a cycle of perception, abduction, and label revision (Zhou, 2019; Dai et al., 2019). Methods refine this cycle through candidate weighting (He et al., 2024), curricula (Wei et al., 2025; Hu et al., 2025b), and probabilistic symbol perception (Jia et al., 2025a). Abductive Reflection learns to locate inconsistent predictions before symbolic repair (Hu et al., 2025a). Consistency-distance analysis relates scoring and pseudo-label sampling to generalization risk and shows how abductive consistency can reduce it (Yang et al., 2024). These methods change how training obtains or uses labels. We fix the complete-code posterior rule and analyze how Gram orientation affects selection among compatible explanations on a common evidence potential.

Identifiability, learnability and shortcut risk. Identifiability and constraint-based analyses characterize which concept interpretations the knowledge and observations can distinguish (Marconato et al., 2023; Bortolotti et al., 2025; Takemura et al., 2026). Learnability results give conditions for recovering latent labels from multi-instance supervision or logical rules (Wang et al., 2023; Tao et al., 2024; He & Li, 2025; Jia et al., 2025b). Risk analyses show how the symbolic component can create unequal class-specific errors (Tsamoura et al., 2025). The distinction between symbolic ambiguity, architectural realizability and optimization selection is explicit in Xu (2026, Sections 6.3–6.4), which also studies how absorbing arithmetic values suppress correction gradients. When multiple explanations remain compatible with supervision, we ask which attract nearby training states and when their labels become captured under fixed logical constraints and exact posterior updates.

Learning from label proportions. Small-bag count supervision uses a posterior over all binary assignments consistent with each count (Baručić & Kybic, 2022). Its exact EM formulation alternates product-posterior normalization with classifier training against marginal label probabilities, including experiments on CIFAR-10 bird/cat images. We use this likelihood and posterior training method with a fixed pretrained representation, refreshing the posterior after every shared-head update. For disjoint bags, global evidence factors into bag evidences, connecting the count likelihood to our selection and capture analysis under sample-space or fixed-linear parameter updates.

Semantic loss and probabilistic neuro-symbolic learning. In the native model, $\Phi = 2\log\mathcal{E}$ is twice the negative semantic loss of Xu et al. (2018), where $\mathcal{E}$ is the probability that independent Bernoulli predictions satisfy the logical constraints. This probability is a weighted model count, connecting our objective to probabilistic neuro-symbolic learning (Manhaeve et al., 2018). The posterior refresh connects this gradient update to latent-variable evidence optimization and generalized EM (Dempster et al., 1977). Exact normalization has computational costs characterized by hardness results for probabilistic neuro-symbolic learning (Maene et al., 2024). Deterministic approximate-EM analysis gives convergence conditions for controlled E-step surrogates (Lartigue et al., 2022). Our error bounds on posterior approximation are relative to the exact residual, yielding quantitative ascent and preserving local stability and sample-SPD capture (Appendices C.10 and E.2).

van Krieken et al. (2024, Theorems 4.3 and 4.10) characterize the zero-loss independent distributions as a union of implicant cubes. Their Theorem 4.13 describes its connected components. In our native regime this implies that a valid one-bit neighbor creates a flat zero-loss edge, whereas a vertex without such a neighbor is isolated in the zero-loss set. Under its candidate and step-size conditions, Theorem 9 establishes a trapping neighborhood and an $O(1/t)$ compactified rate for discrete logit updates with any fixed positive-definite Gram. The critical functional describes how priors and temperature affect stability, and the inexact extension controls posterior-refresh errors. Appendix C.1 distinguishes closed-face stability from attraction of finite initializations.

Ambiguous-target optimization. Softmax regression with partial-label likelihood exhibits winner-take-all behavior among initially maximal acceptable labels (Zombori et al., 2024, Theorem 4), while changing the loss can preserve their probability ratios. Self-training analyses study confidence thresholds and fixed points (Oymak & Gulcu, 2021), generalization from unlabeled data (Zhang et al., 2022), and spurious-feature removal under domain shift (Chen et al., 2020). Test-time adaptation theory studies gradient descent with hard and conjugate pseudo-labels (Wang & Wibisono, 2023). With a complete-code posterior, we derive stability criteria involving defect overlap and temperature and construct isospectral Grams that select different endpoints from a common initialization.

Implicit bias and optimizer geometry. Implicit-bias theory asks which solution gradient methods select and how optimizer geometry affects that selection (Soudry et al., 2018; Gunasekar et al.,

2018; Ravi et al., 2024; Wu et al., 2023). Metric-dependent basins for a fixed potential are classical (Fiedler et al., 1999). Selection of different minimizers with a fixed spectrum also follows from optimization-geometry theory (Gunasekar et al., 2018, Theorem 1). We establish this dependence for selection among discrete legal explanations of the same logical observations. Our construction routes to arbitrarily many such endpoints of one evidence potential from a common open initialization neighborhood, with matched full and candidate-active spectra.

Regions of attraction and certificates. Lyapunov sublevel sets and optimization-based searches can certify regions of attraction for nonlinear discrete systems (Leisenberger et al., 2021). Bounded variable metrics and controlled inexactness are standard tools for descent and stationarity (Bonettini et al., 2016). For Hamming-separated binary candidates, legal evidence provides an explicit barrier: it is at most one half on every hard-sign boundary. Bounding evidence along the entire update segment turns that static barrier into hard-code invariance under arbitrary admissible metric changes. Residual bounds give corner convergence with an explicit compactified rate. The certificate can be checked at positive loss by evaluating legal evidence, without a separate search for a region of attraction.

Local basin stability for preconditioned SGD has also been studied around minimizer sets, including flat sets (Scott et al., 2026, Section 3.2, Theorem 3.5). That analysis uses a fixed SPD parameter metric, local Polyak–Łojasiewicz and quadratic-growth conditions, and stochastic-gradient and overshoot bounds to control the probability of remaining in a basin up to a finite horizon. For our exact binary updates, half evidence guarantees code preservation and endpoint convergence under any subsequent history-dependent metric sequence obeying the uniform spectral and step-size bounds.

Continuous constraint solving. FourierCSP minimizes the negative expected number of satisfied constraints as a multilinear function on a product of probability simplices (Cen et al., 2026, Section 3). It studies Euclidean projected gradient descent and negative-entropy mirror descent, with boundary-optimality and critical-point convergence results in Theorems 3.8–3.10. Our objective is the log probability of joint legality, and updates act on logits or shared linear parameters. The half-evidence result guarantees persistence of a specific complete code throughout every admissible continuation. The metric may vary during training, whereas FourierCSP establishes convergence for specified descent rules.

B Model, Proof Roadmap, and Common Tools

B.1 Core proof dependencies

Table 1: Proof locations and principal dependencies for the main results.

Result	Main dependency	Proof
Local classifications (Theorem 1, Proposition 2)	Anchored factorization, gap bounds, discrete trapping	Appendix C.8
Isospectral routing (Theorem 3)	Slow–fast construction and local attraction	Appendix D.8
Half-evidence capture (Theorem 4)	Disjoint slices, segment ascent, gap bounds	Appendix E, p. 31
Capture with limited control over individual predictions (Corollary 1)	Evidence barrier and separation by the attained head	Appendix E.3
Suboptimal attraction (Propositions 5 and 11)	Prototype stationary point, full Hessian, occurrence mapping	Appendix E.4

B.2 Parameterized posterior and direct supervision

The native update of Section 2 uses uniform priors, unit temperature, and no direct supervision. The local theory also allows positive priors π_z, temperature $\tau > 0$, and direct semantic supervision of relative weight $\lambda \geq 0$. The extended posterior is

$$q_z(y) = \frac{[\pi_z \prod_i \sigma(z_i y_i)]^{1/\tau}}{\sum_{w \in \mathcal{C}} [\pi_w \prod_i \sigma(w_i y_i)]^{1/\tau}}, \qquad p_i(y) = \Pr_q(z_i = +1), \tag{15}$$

and a detached, full-batch cross-entropy step gives

$$y^+ = y + \eta K R(y), \qquad R_i(y) = 2[p_i(y) + \lambda - (1+\lambda)\sigma(y_i)]. \tag{16}$$

Its potential is

$$\Phi_{\mathcal{C},\lambda}(y) = 2\tau \log \sum_{z\in\mathcal{C}} \pi_z^{1/\tau} \prod_i \sigma(z_i y_i)^{1/\tau} + 2\lambda \sum_i \log \sigma(y_i), \tag{17}$$

with $\nabla\Phi_{\mathcal{C},\lambda} = R$ and $\Phi_{\mathcal{C},\lambda} \le 2\tau \log \sum_z \pi_z^{1/\tau}$. For normalized uniform priors, $\tau = 1$, and $\lambda = 0$, $\Phi_{\mathcal{C},0} = 2\log \mathcal{E} - 2\log|\mathcal{C}|$. Thus it differs from the native Φ in (4) by an additive constant and induces the same update. In the parameterized local-stability and posterior-error analyses, Φ abbreviates $\Phi_{\mathcal{C},\lambda}$. Potential deficits subtract the potential at the corresponding code corner.

Lemma 1 (Defect-direction odds)**.** *For all $z, w \in \mathcal{C}$ and all $y \in \mathbb{R}^N$,*

$$\log \frac{q_w(y)}{q_z(y)} = \frac{1}{\tau}\left[\log \frac{\pi_w}{\pi_z} - \sum_{i\in D(w,z)} z_i y_i\right].$$

Relative to an anchor z, q is therefore a softmax over affine log-odds with coefficient $-z_i/\tau$ on each defect coordinate $i \in D(w,z)$ and zero elsewhere. The candidate set enters through these defect sets and the prior ratios.

The identity follows from $\log\sigma(-t) - \log\sigma(t) = -t$ on each defect coordinate.

Proposition 6 (Gram invariance)**.** *Let $K \succ 0$. The potential and its finite critical points are independent of K. A sufficient global condition for strict ascent whenever $R(y) \neq 0$ is*

$$\eta\lambda_{\max}(K)\left(\frac{N}{\tau} + 1 + \lambda\right) < 4, \tag{18}$$

because $\|\nabla^2\Phi\|_{\mathrm{op}} \le L_\Phi$. Finite critical-point invariance follows from $KR = 0 \iff R = 0$. This invariance does not make the Jacobian $I + \eta K\nabla^2\Phi(y^\star)$ or its discrete multipliers Gram-independent.

B.3 Common notation

The proofs use the following notation from the main text:

$$\begin{aligned} D(w,z) &= \{i : w_i \neq z_i\}, \quad \text{the defect set between codes } w, z,\\ a_i &= e^{-z_i y_i}, \quad \text{compactified coordinate},\\ r &= \sum_{i=1}^N a_i, \quad \text{compactified radius},\\ K_z &= D_z K D_z, \quad \text{where } D_z = \mathrm{diag}(z),\\ F_K(y) &= y + \eta K R(y), \quad \text{the one-step map}. \end{aligned}$$

For an SPD matrix K, write $\|v\|_K^2 = v^\top K v$. A trajectory approaches the code corner z exactly when $a_t \to 0$, or equivalently when every signed margin $z_i y_{i,t}$ diverges to $+\infty$.

B.4 Common analytic tools

Splitting off the semantic factor gives the equivalent potential

$$\Phi_{\mathcal{C},0}(y) = 2\sum_i \log\sigma(y_i) + 2\tau \log \sum_{z\in\mathcal{C}} \pi_z^{1/\tau} e^{-\langle \mathbf{1}_{D(z,\mathbf{1}_N)}, y\rangle/\tau}.$$

The first term is separable and strictly concave. The second is a convex log-sum-exp of the defect functionals, with weights determined by the candidate priors and temperature.

Lemma 2 (Hessian bound and global smoothness)**.** *The potential $\Phi : \mathbb{R}^N \to \mathbb{R}$ is twice continuously differentiable with gradient $\nabla\Phi(y) = R(y)$. Its Hessian satisfies*

$$\nabla^2\Phi(y) = \frac{2}{\tau}\,\mathrm{Cov}_{q(y)}[b(z)] - 2(1+\lambda)\,\mathrm{diag}(\sigma'(y)),$$

where $b(z) = (z + \mathbf{1})/2 \in \{0, 1\}^N$ and $q(y)$ is the complete-code posterior. Consequently, for any $y \in \mathbb{R}^N$,

$$\|\nabla^2 \Phi(y)\|_{\mathrm{op}} \le L_\Phi := \frac{1}{2}\left(\frac{N}{\tau} + 1 + \lambda\right).$$

The gradient $\nabla\Phi$ is globally Lipschitz with constant L_Φ.

Proof. For the unnormalized posterior weight in (15), the y_i-derivative of its log is $(b_i(z) - \sigma(y_i))/\tau$. Softmax differentiation therefore gives $\nabla p(y) = \mathrm{Cov}_{q(y)}[b(z)]/\tau$. Differentiating (17) gives $\nabla\Phi = R$, and differentiating $R(y) = 2[p(y) + \lambda\mathbf{1} - (1 + \lambda)\sigma(y)]$ gives the stated Hessian. The covariance term has operator norm at most $N/4$ (uniform bound over all distributions on $\{0, 1\}^N$), and $\sigma'(t) \le 1/4$ uniformly. The Lipschitz property follows from the mean-value theorem. □

Lemma 3 (Ascent inequality)**.** *Let $K \succ 0$ and $y \in \mathbb{R}^N$ with $R(y) \neq 0$. Then*

$$\Phi(F_K(y)) - \Phi(y) \ge \eta R(y)^\top K R(y) - \frac{\eta^2 L_\Phi}{2}\|KR(y)\|^2.$$

Consequently,

$$\Phi(F_K(y)) - \Phi(y) \ge \eta\left(1 - \frac{\eta L_\Phi \lambda_{\max}(K)}{2}\right) R(y)^\top K R(y).$$

If the step condition

$$\eta\lambda_{\max}(K)\left(\frac{N}{\tau} + 1 + \lambda\right) < 4$$

holds, the coefficient in the preceding display is positive. Consequently, $\Phi(F_K(y)) > \Phi(y)$ whenever $R(y) \neq 0$.

Proof. Set $\Delta = \eta K R(y)$. The lower smoothness inequality from Lemma 2 gives

$$\Phi(F_K(y)) - \Phi(y) \ge \eta R(y)^\top K R(y) - \frac{\eta^2 L_\Phi}{2}\|KR(y)\|^2.$$

Because $K \succ 0$, spectral calculus gives $K^2 \preceq \lambda_{\max}(K)K$, and hence

$$\|KR(y)\|^2 = R(y)^\top K^2 R(y) \le \lambda_{\max}(K) R(y)^\top K R(y).$$

Substitution gives the second inequality. Since $L_\Phi = \frac{1}{2}(N/\tau + 1 + \lambda)$, the step condition is equivalent to $\eta L_\Phi \lambda_{\max}(K) < 2$, making the coefficient $1 - \eta L_\Phi \lambda_{\max}(K)/2$ positive. Positive definiteness also gives $R(y)^\top K R(y) > 0$ whenever $R(y) \neq 0$, proving strict ascent at every nonstationary state. □

Lemma 4 (SPD spectral neighborhood)**.** *Let $K_0 \succ 0$ with eigenvalues $\lambda_{\min}(K_0) > 0$ and $\lambda_{\max}(K_0)$. For any $0 < \delta < \frac{1}{2}\lambda_{\min}(K_0)$, the set*

$$\mathcal{V}_\delta = \{K \succ 0 : \|K - K_0\|_F < \delta\}$$

is an open neighborhood of K_0 in the cone of SPD matrices, and every $K \in \mathcal{V}_\delta$ satisfies

$$\lambda_{\min}(K) \ge \tfrac{1}{2}\lambda_{\min}(K_0), \qquad \lambda_{\max}(K) \le 2\lambda_{\max}(K_0).$$

Proof. Since $\|K - K_0\|_{\mathrm{op}} \le \|K - K_0\|_F < \delta$, the Weyl inequality gives

$$\lambda_{\min}(K) \ge \lambda_{\min}(K_0) - \delta > \lambda_{\min}(K_0) - \tfrac{1}{2}\lambda_{\min}(K_0) = \tfrac{1}{2}\lambda_{\min}(K_0),$$

and similarly

$$\lambda_{\max}(K) \le \lambda_{\max}(K_0) + \delta < \lambda_{\max}(K_0) + \tfrac{1}{2}\lambda_{\min}(K_0) \le \tfrac{3}{2}\lambda_{\max}(K_0) < 2\lambda_{\max}(K_0).$$

□

Lemma 5 (Finite-horizon continuity)**.** *Fix $T \in \mathbb{N}$, a compact set $\mathcal{Y}_0 \subset \mathbb{R}^N$, and a compact set $\mathcal{K} \subset \{K \succ 0\}$ in the SPD cone. Then the map*

$$(y_0, K) \mapsto F_K^T(y_0)$$

is uniformly continuous on $\mathcal{Y}_0 \times \mathcal{K}$.

Proof. The residual R is continuous, so $(K, y) \mapsto F_K(y) = y + \eta K R(y)$ is jointly continuous. Induction on T shows that every finite iterate $(K, y_0) \mapsto F_K^T(y_0)$ is jointly continuous. Its domain $\mathcal{Y}_0 \times \mathcal{K}$ is compact, so the Heine–Cantor theorem makes this map uniformly continuous there. □

C Local Endpoint Stability

We prove the local classifications in Theorem 1 and Proposition 2 by expanding the parameterized potential (17) near a candidate corner and controlling the discrete updates there. After clarifying the role of boundary faces, we derive these bounds, extend them to posterior errors, and examine the local predictions numerically at finite radii along dense and near-face directions.

A corner is called *potential-viable* when it is a strict local maximum of this potential on the closed compactified orthant. It is *decisively viable* when $d_{\min}(z) > \tau$, or when $d_{\min}(z) = \tau$ and $\chi_z < 1+\lambda$. With direct semantic supervision $\lambda > 0$, the anchor is the semantic code $z = \mathbf{1}_N$.

C.1 Closed faces and finite-logit reachability

In the native model, let $\mathcal{C} = \{(+,+),(-,+)\}$ and take $K_c = \left(\begin{smallmatrix} 1 & c \\ c & 1 \end{smallmatrix}\right)$, $0 < c < 1$, and $3\eta(1+c) < 4$. Now $R_1 = 0$ and $R_2 = 2\sigma(-y_2)$. Thus y_2 increases without a finite limit, while $y_1 - cy_2$ is invariant: every finite initialization reaches $(+,+)$. Both compactified coordinates decrease near that corner, yet the face $a_2 = 0, a_1 > 0$ is stationary. Theorem 1 therefore includes saturated-coordinate faces in its attraction criterion. Replacing c by $-c$ preserves the spectrum and sends every finite initialization to $(-,+)$.

C.2 The anchored factorization

Fix $z \in \mathcal{C}$ and use the compactified coordinates $a_i = e^{-z_i y_i} \in (0, \infty)$ of (6). For each coordinate,

$$\sigma(z_i y_i) = \frac{1}{1+a_i}, \qquad \frac{\sigma(-z_i y_i)}{\sigma(z_i y_i)} = a_i,$$

where the ratio follows from $\sigma(-t)/\sigma(t) = e^{-t}$. For $w \in \mathcal{C}$, coordinates on which w agrees with z contribute $\sigma(z_i y_i)$, whereas each defect coordinate contributes $\sigma(-z_i y_i)$. Hence

$$\prod_i \sigma(w_i y_i) = \left[\prod_i \sigma(z_i y_i)\right] \prod_{i \in D(w,z)} a_i = \frac{1}{\prod_i (1+a_i)} \prod_{i \in D(w,z)} a_i. \tag{19}$$

Writing $W_w = \pi_w^{1/\tau} \prod_i \sigma(w_i y_i)^{1/\tau}$ for the unnormalized weight, (19) gives the ratio

$$\frac{W_w}{W_z} = \left(\frac{\pi_w}{\pi_z}\right)^{1/\tau} \prod_{i \in D(w,z)} a_i^{1/\tau}, \tag{20}$$

which is (7): a monomial in a of total degree $d_{\mathrm{H}}(w,z)/\tau$, with *positive* exponents. Factoring W_z out of the log-sum in (17) therefore yields, with no approximation and for every $a \in (0,\infty)^N$,

$$\Phi_{\mathcal{C},0}(a) = 2\log \pi_z - 2\sum_i \log(1+a_i) + 2\tau \log\Big[1 + S_z(a)\Big], \qquad S_z(a) = \sum_{w \neq z} \frac{W_w}{W_z}. \tag{21}$$

The anchor factor contributes $-2\sum_i \log(1+a_i)$, while S_z collects the rival weights relative to the anchor. Writing $a = ru$ with $r = \sum_i a_i$ and $u \in \Delta^{N-1}$,

$$S_z(ru) = \sum_{w \neq z} \left(\frac{\pi_w}{\pi_z}\right)^{1/\tau} r^{d_{\mathrm{H}}(w,z)/\tau} \prod_{i \in D(w,z)} u_i^{1/\tau}. \tag{22}$$

The monomials in (22) have positive exponents, so they remain continuous when some $u_i = 0$. Together with (21), this gives a continuous potential on the closed compactified orthant near the origin, with corner value $\Phi_{\mathcal{C},0}(0) = 2\log \pi_z$. The coefficient bounds in the expansion below hold uniformly over the entire closed simplex, including its boundary faces.

Remark 1 (Where direct semantic supervision may be added). The direct semantic supervision term of (17) is $2\lambda \sum_i \log \sigma(y_i)$, written in the *semantic* margins y_i, not in the anchored margins $z_i y_i$. At

the semantic anchor $z = \mathbf{1}_N$ the two agree, $\log \sigma(y_i) = -\log(1 + a_i)$, and supervision rescales the first sum in (21) to $-2(1 + \lambda) \sum_i \log(1 + a_i)$. At any other anchor there is a coordinate with $z_i = -1$, along which $y_i \to -\infty$ as the corner is approached, so $\log \sigma(y_i) \to -\infty$ and $\Phi_{\mathcal{C},\lambda}$ has no finite value at that corner. We therefore restrict $\lambda > 0$ to $z = \mathbf{1}_N$ and set $\lambda = 0$ otherwise.

C.3 Local maximality is necessary

Lemma 6 (Viability is necessary). *Under the ascent-step condition (18), if a corner $a = 0$ is an asymptotically stable endpoint of (16), then it is a strict local maximum of $\Phi_{\mathcal{C},\lambda}$.*

Proof of Lemma 6. The monomial ratios in (20) make the posterior and residual continuous on a neighborhood of the origin in the closed compactified orthant. The update extends there as

$$T_{z,K}(a)_i = a_i \exp(-\eta z_i [KR(a)]_i), \tag{23}$$

with $T_{z,K}(0) = 0$. The factorization (21) gives the corresponding continuous extension of Φ.

Suppose that the origin is asymptotically stable for this compactified map. Then some neighborhood $\mathcal{N}$ of the origin is attracted to it. We claim that every $a \in \mathcal{N} \setminus \{0\}$ satisfies $\Phi(a) < \Phi(0)$. If $R(a) = 0$, then (23) fixes a, contradicting attraction. If $R(a) \neq 0$, the quantitative ascent inequality from Lemma 3, extended to the closed orthant by continuity, gives

$$\Phi(T_{z,K}(a)) - \Phi(a) \geq \eta \left(1 - \frac{\eta L_\Phi \lambda_{\max}(K)}{2}\right) R(a)^\top K R(a) > 0.$$

If $\Phi(a) \geq \Phi(0)$, this strict increase and the subsequent nondecrease of Φ contradict $T^t_{z,K}(a) \to 0$ and continuity of the potential. Hence $\Phi(a) < \Phi(0)$ throughout a punctured neighborhood of the origin in the closed orthant, establishing strict local maximality even along its boundary faces. □

C.4 Expansion and the three regimes

Lemma 7 (Uniform boundary expansion). *Let priors be positive, write $\alpha = d_{\min}(z)/\tau$, and take $\lambda = 0$ unless $z = \mathbf{1}_N$ is the semantic code, in which case any $\lambda \geq 0$ is allowed. Then, uniformly over $u \in \Delta^{N-1}$ including its faces,*

$$\Phi_{\mathcal{C},\lambda}(ru) - \Phi_{\mathcal{C},\lambda}(0) = -2(1+\lambda)\, r + 2\tau \sum_{w \neq z} \left(\frac{\pi_w}{\pi_z}\right)^{1/\tau} r^{d_{\mathrm{H}}(w,z)/\tau} \prod_{i \in D(w,z)} u_i^{1/\tau} + O\Big(r^{\min\{2,\, 2\alpha\}}\Big),$$

where the constant depends on N, τ, λ, $|\mathcal{C}|$ and the prior ratios $(\pi_w/\pi_z)^{1/\tau}$, but not on u or r.

Proof of Lemma 7. Put $\alpha = d_{\min}(z)/\tau$ and let $C_\pi = \sum_{w \neq z} (\pi_w/\pi_z)^{1/\tau}$. Since $u_i \leq 1$ and every exponent in (22) is at least α, for $r \leq 1$

$$0 \leq S_z(ru) \leq C_\pi\, r^\alpha \qquad \text{uniformly in } u \in \Delta^{N-1}. \tag{24}$$

Apply $\log(1 + x) = x + O(x^2)$ to (21) and $\log(1 + a_i) = a_i + O(a_i^2)$ to the first sum, using $\sum_i a_i = r$ and $\sum_i a_i^2 \leq r^2$:

$$\Phi_{\mathcal{C},\lambda}(ru) - \Phi_{\mathcal{C},\lambda}(0) = -2(1 + \lambda)\, r + 2\tau\, S_z(ru) + O\Big(r^{\min\{2,\, 2\alpha\}}\Big), \tag{25}$$

where $\lambda = 0$ unless $z = \mathbf{1}_N$ (Remark 1), and the O constant depends on N, τ, λ, $|\mathcal{C}|$ and C_π (that is, on the prior ratios), but not on u or r. The two remainder sources are the quadratic term of the first logarithm, which is $O(r^2)$, and that of the outer logarithm, which by (24) is $O(S_z^2) = O(r^{2\alpha})$. For $\alpha < 1$, the latter is the larger bound, giving the remainder $O(r^{\min\{2,2\alpha\}})$. □

C.5 Distance, temperature, and the critical coefficient

Theorem 7 (Candidate-geometry viability criterion). *Let $\lambda = 0$ and the priors be positive. If $d_{\min}(z) > \tau$, then z is potential-viable. If $d_{\min}(z) < \tau$, it is not. At $d_{\min}(z) = \tau$, the functional H_z of (8) gives, uniformly over $u \in \Delta^{N-1}$,*

$$\Phi(ru) - \Phi(0) = 2r[H_z(u) - 1] + o(r).$$

Thus $\chi_z < 1$ implies viability and $\chi_z > 1$ excludes it. At $\chi_z = 1$, this test is inconclusive.

Proof. Take $\lambda = 0$ throughout, so (25) reads $\Phi(ru) - \Phi(0) = -2r + 2\tau S_z(ru) + O(r^{\min\{2,2\alpha\}})$.

(i) $\alpha > 1$. By (24), $2\tau S_z(ru) \le 2\tau C_\pi r^\alpha = o(r)$ uniformly in u, and the remainder is $O(r^{\min\{2,2\alpha\}}) = o(r)$ since $\min\{2, 2\alpha\} > 1$. Hence there is $r_0 > 0$, depending only on $N, \tau, |\mathcal{C}|, C_\pi$, with

$$\Phi(ru) - \Phi(0) \le -r < 0 \qquad \text{for all } 0 < r \le r_0 \text{ and all } u \in \Delta^{N-1},$$

so $a = 0$ is a strict local maximum: z is potential-viable. The correction is $O(r^{\min\{2,\alpha\}})$, combining rival terms of order r^α with the quadratic logistic term. For $\alpha > 2$, the logistic term sets this order.

(ii) $\alpha < 1$. Let w be a nearest rival and choose u uniform on its defect set, $u_i = 1/d_{\min}$ for $i \in D(w, z)$ and 0 elsewhere. This is a legitimate point of the closed simplex, and (22) keeps only terms whose defect set lies inside $D(w, z)$, so $S_z(ru) \ge (\pi_w/\pi_z)^{1/\tau} d_{\min}^{-\alpha} r^\alpha =: c\, r^\alpha$ with $c > 0$. Then

$$\Phi(ru) - \Phi(0) \ge 2\tau c\, r^\alpha - 2r - O(r^{2\alpha}) > 0 \tag{26}$$

for all small $r > 0$, because $\alpha < \min\{1, 2\alpha\}$. This excludes local maximality and potential viability of z.

(iii) $\alpha = 1$. Now the nearest-rival terms of (22) are exactly linear in r and the rest are $O(r^{d_2/\tau})$ with $d_2 > \tau$ the next-nearest distance, so $2\tau S_z(ru) = 2r\, H_z(u) + o(r)$ with H_z as in (8), uniformly in u because H_z is a finite sum of monomials bounded on the simplex. The remainder in (25) is $O(r^2)$. Hence

$$\Phi(ru) - \Phi(0) = 2r\big[H_z(u) - 1\big] + o(r) \qquad \text{uniformly in } u, \tag{27}$$

The maximum $\chi_z = \max_{\Delta^{N-1}} H_z$ is attained because H_z is continuous on the compact simplex. The case $\chi_z < 1$ gives $\Phi(ru) - \Phi(0) \le 2r(\chi_z - 1) + o(r) < 0$ uniformly, so z is potential-viable, while $\chi_z > 1$ gives a direction along which the expression is positive. At $\chi_z = 1$ the leading term vanishes on the maximizing set, requiring higher-order terms or the exact expression. □

C.6 The critical functional is a concave maximization

Proposition 8 (χ_z is computable)**.** *In the critical case $d_{\min}(z) = \tau$, each summand of H_z is a weighted geometric mean of the coordinates of u whose exponents $1/\tau$ over $i \in D(w, z)$ sum to $d_{\min}(z)/\tau = 1$. Weighted geometric means with exponents summing to one are concave on the nonnegative orthant, so H_z is concave and*

$$\chi_z = \max_{u \in \Delta^{N-1}} H_z(u)$$

is a concave maximization over the simplex, solvable to global optimality.

For integer critical degree $k = \tau = d_{\min}(z) \ge 2$, the same functional is the adjacency spectral radius of a weighted k-uniform defect hypergraph, using the tensor normalization of Cooper & Dutle (2012). Give edge $D(w, z)$ weight $c_w = (\pi_w/\pi_z)^{1/k}$ and each of its tensor entries value $c_w/(k-1)!$. The tensor form is $\mathcal{A}x^k = k \sum_w c_w \prod_{i \in D(w,z)} x_i$. Substituting $x_i = u_i^{1/k}$ into its nonnegative ℓ_k-unit variational characterization gives $\chi_z = \rho(\mathcal{A})$. This spectral invariant determines the critical threshold for discrete boundary stability. The two arrangements in Example 1 give explicit values: Write $c = (\pi_w/\pi_z)^{1/\tau}$ for the common prior factor and take $\tau = 2$, $d_{\min} = 2$, so each summand is $2c\sqrt{u_i u_j}$. With either rival alone, $\chi_z = c$, since $2\sqrt{u_i u_j} \le u_i + u_j \le 1$, with equality at $u_i = u_j = 1/2$.

Disjoint defect sets $\{1, 2\}$ and $\{3, 4\}$. By AM–GM, $\sqrt{u_1 u_2} + \sqrt{u_3 u_4} \le \frac{u_1+u_2}{2} + \frac{u_3+u_4}{2} = \frac{1}{2}$, with equality at $u_1 = u_2$, $u_3 = u_4$. Hence $\chi_z = 2c \cdot \frac{1}{2} = c$.

Overlapping defect sets $\{1, 2\}$ and $\{1, 3\}$. Here $H_z = 2c\sqrt{u_1}\big(\sqrt{u_2} + \sqrt{u_3}\big)$. Setting $u_4 = 0$ and $u_2 = u_3 = (1 - u_1)/2$ by symmetry gives $H_z = 2c\sqrt{2u_1(1 - u_1)}$, maximized at $u_1 = \frac{1}{2}$, so $\chi_z = \sqrt{2}\, c$.

With $c = 0.8$, that is $\pi_w/\pi_z = c^\tau = 0.64$, the two values are 0.800 and $\sqrt{2} \cdot 0.8 \approx 1.131$: overlap alone moves χ_z across the threshold, and no numerical maximization is needed to see it.

C.7 Discrete trapping and the compactified rate

Theorem 9 (Local endpoint stability). *Fix a candidate code $z \in \mathcal{C}$. Suppose either*

$$d_{\min}(z) > \tau, \qquad \text{or} \qquad d_{\min}(z) = \tau \quad \text{and} \quad \chi_z < 1 + \lambda.$$

If $\lambda > 0$, assume $z = \mathbf{1}_N$ is the semantic code. Fix $K \succ 0$ and $\eta > 0$ satisfying (18).

Then there are positive constants $c_1, c_2, c_K, r_\star$, with c_K depending additionally on K and η, such that for every $0 < \varepsilon \le r_\star$, the compactified initialization condition

$$r_0 < \frac{c_1}{c_2}\varepsilon,$$

implies $r_t < \varepsilon$ for all $t \ge 0$ and

$$c_1 r_t \le V_z(a_t) \le c_2 r_t, V_{z,t+1} \le V_{z,t} - c_K r_t^2,$$

where $r_t = \sum_i a_{i,t}$, $a_i = e^{-z_i y_i}$, and $V_z(a) = -[\Phi(a) - \Phi(0)]$ is the potential deficit.

Consequently, z is an asymptotically stable endpoint with $r_t = O(1/t)$.

Proof. First pass to signed coordinates:

$$x_i = z_i y_i, \qquad a_i = e^{-x_i}, \qquad D_z = \mathrm{diag}(z), \qquad K_z = D_z K D_z \succ 0, \tag{28}$$

so that $x = D_z y$, $dx = D_z dy$, and the map (16) becomes

$$x^+ = x + \eta K_z g_z(a), \qquad g_z(a) := D_z R(y) = \nabla_x \Phi. \tag{29}$$

The potential deficit is $V_z(a) = -[\Phi(a) - \Phi(0)]$, and z is approached as $a \to 0$. Write $r = \sum_i a_i$ and $u = a/r \in \Delta^{N-1}$.

Exact gradient identity and uniform bounds. For each rival $w \ne z$, define the monomial

$$m_w(a) = \left(\frac{\pi_w}{\pi_z}\right)^{1/\tau} \prod_{i \in D(w,z)} a_i^{1/\tau}, \qquad S_z(a) = \sum_{w \ne z} m_w(a).$$

By direct computation from the posterior (15) and the residual definition $R_i = 2[p_i + \lambda - (1+\lambda)\sigma(y_i)]$, the signed gradient is

$$g_{z,i}(a) = \frac{2(1+\lambda)a_i}{1+a_i} - \frac{2}{1+S_z(a)} \sum_{\substack{w \ne z \\ i \in D(w,z)}} m_w(a). \tag{30}$$

This identity holds for all a, including simplex faces $a_i = 0$, because m_w has positive exponents.

Upper bound. Each $m_w(a) \le C_w r^{\deg(w)}$ with $\deg(w) = d_{\mathrm{H}}(w,z)/\tau$, and $1/(1+S_z) \le 1$. Since $a_i/(1+a_i) < a_i$, we get

$$|g_{z,i}(a)| \le 2(1+\lambda)a_i + 2 \sum_{w : i \in D(w,z)} m_w(a) \le C_4 r, \tag{31}$$

uniformly in u, where C_4 depends on N, τ, $|\mathcal{C}|$, λ, and the prior ratios.

Lower bound, strictly viable case $d_{\min} > \tau$. Set $\alpha = d_{\min}(z)/\tau > 1$. Every rival degree is at least α, so $S_z(a) = O(r^\alpha)$ and each monomial sum in (30) is $O(r^\alpha)$. The leading term of $g_{z,i}$ is $2(1+\lambda)a_i$, with correction $O(r^{\min\{2,\alpha\}}) = O(r^{1+\delta})$ where $\delta = \min\{1, \alpha - 1\} > 0$. Thus

$$g_z = 2(1+\lambda)a + O(r^{1+\delta}),$$

giving $\|g_z\|_2 \ge (1+\lambda)r/\sqrt{N}$ for sufficiently small r.

Lower bound, critically viable case $d_{\min} = \tau$, $\chi_z < 1 + \lambda$. The nearest rivals have $\deg(w) = 1$. Summing (30) over i and using $\sum_{i : i \in D(w,z)} 1 = d_{\mathrm{H}}(w,z)$:

$$\mathbf{1}^\top g_z = 2(1+\lambda) \sum_i \frac{a_i}{1+a_i} - \frac{2}{1+S_z(a)} \sum_{w \ne z} d_{\mathrm{H}}(w,z)\, m_w(a).$$

For critical nearest rivals w with $d_{\mathrm{H}}(w,z)=\tau$, define the monomial coefficient

$$h_w(u)=\left(\frac{\pi_w}{\pi_z}\right)^{1/\tau}\prod_{i\in D(w,z)}u_i^{1/\tau},$$

so that $m_w(ru)=rh_w(u)$ exactly. The critical functional is then $H_z(u)=\tau\sum_{w:\,d_{\mathrm{H}}(w,z)=\tau}h_w(u)$, as in Theorem 7. For farther rivals with $d_{\mathrm{H}}(w,z)>\tau$ the contributions are $o(r)$. Since $\sum_i a_i/(1+a_i)=r-O(r^2)$ and $1/(1+S_z)=1-O(r)$, we obtain

$$\mathbf{1}^\top g_z=2r[(1+\lambda)-H_z(u)]+o(r),$$

uniformly in u. The viability gap $\varepsilon=(1+\lambda)-\chi_z>0$ makes $H_z(u)\le\chi_z<1+\lambda$ uniformly, so $\mathbf{1}^\top g_z\ge\varepsilon r$ for small r. By Cauchy–Schwarz, $\|g_z\|_2\ge\varepsilon r/\sqrt{N}$.

Set the gradient lower-bound constant c_3 as follows:

$$c_3=\begin{cases}\dfrac{1+\lambda}{\sqrt{N}}, & d_{\min}(z)>\tau,\\[2mm] \dfrac{(1+\lambda)-\chi_z}{\sqrt{N}}, & d_{\min}(z)=\tau,\ \chi_z<1+\lambda.\end{cases}$$

From (31), the componentwise bound implies $\|g_z\|_2\le\sqrt{N}\,C_4r$. Renaming $\widetilde{C}_4=\sqrt{N}\,C_4$, we have

$$c_3r\le\|g_z\|_2\le\widetilde{C}_4r \tag{32}$$

uniformly in $u\in\Delta^{N-1}$ for sufficiently small r.

Potential gap bounds. By Lemma 7, uniformly in $u\in\Delta^{N-1}$:

$$\Phi(ru)-\Phi(0)=-2(1+\lambda)r+2\tau S_z(ru)+O\Big(r^{\min\{2,2\alpha\}}\Big),$$

where $\alpha=d_{\min}/\tau$. Since $V_z(a)=-[\Phi(a)-\Phi(0)]$:

Strictly viable ($\alpha>1$): $S_z=O(r^\alpha)$, so $V_z=2(1+\lambda)r+O(r^{\min\{2,\alpha\}})$. For small r, set $c_1=1+\lambda$ and $c_2=3(1+\lambda)$.

Critically viable ($\alpha=1$): $V_z=2r[(1+\lambda)-H_z(u)]+o(r)$. The gap $\delta=(1+\lambda)-\chi_z>0$ ensures $(1+\lambda)-H_z(u)\ge\delta$ uniformly. For small r, set $c_1=\delta$ and $c_2=2(1+\lambda)+\delta$.

Quantitative ascent and trapping. The ascent condition (18) gives the smoothness constant $L_\Phi\le N/(2\tau)+(1+\lambda)/2$. Set $\theta=\eta L_\Phi\lambda_{\max}(K)/2<1$. By the signed-coordinate map (29) and the lower smoothness inequality,

$$\Phi(x^+)-\Phi(x)\ge\eta g_z^\top K_zg_z-\frac{L_\Phi\eta^2}{2}\|K_zg_z\|^2\ge\eta(1-\theta)\lambda_{\min}(K)\|g_z\|^2.$$

Combined with $\|g_z\|\ge c_3r$, this gives

$$V_{t+1}\le V_t-\eta(1-\theta)\lambda_{\min}(K)c_3^2r_t^2. \tag{33}$$

Set $c_K=\eta(1-\theta)\lambda_{\min}(K)c_3^2>0$.

Trapping. From (31) and the one-step update,

$$\|\Delta x\|_\infty=\eta\|K_zg_z\|_\infty\le\eta\lambda_{\max}(K)\widetilde{C}_4r_t=:B_Kr_t,$$

and since $a_i^+=a_ie^{-\Delta x_i}$,

$$e^{-B_Kr_t}r_t\le r_{t+1}\le e^{B_Kr_t}r_t. \tag{34}$$

Choose $r_\star>0$ small enough that the gap bounds $c_1r\le V_z\le c_2r$, the gradient bounds (32), and $e^{2B_Kr_\star}<2$ all hold for $r\le 2r_\star$.

For any $0<\varepsilon\le r_\star$, suppose the initial condition satisfies

$$r_0<\frac{c_1}{c_2}\varepsilon,$$

so that $V_0 \le c_2 r_0 < c_1\varepsilon$. If there is a first time $t^\star$ with $r_{t^\star} \ge \varepsilon$, then $r_{t^\star-1} < \varepsilon \le r_\star$, so $r_{t^\star} < 2\varepsilon \le 2r_\star$ by (34), and the local bounds apply at $t^\star$:

$$V_{t^\star} \ge c_1 r_{t^\star} \ge c_1\varepsilon.$$

But the potential deficit V_t is nonincreasing by (33), so $V_{t^\star} \le V_0 < c_1\varepsilon$, a contradiction. Hence $r_t < \varepsilon$ for all $t \ge 0$. This proves Lyapunov stability for every $\varepsilon \le r_\star$. The Riccati argument below also gives $r_t \to 0$ throughout the trapping neighborhood, proving local asymptotic stability.

Convergence rate. From (33) and the upper gap bound,

$$V_{t+1} \le V_t - c_K r_t^2 \le V_t - \frac{c_K}{c_2^2} V_t^2. \tag{35}$$

Let $\gamma_K = c_K/c_2^2 > 0$. For small V_t (which holds once $t \ge t_0$),

$$\frac{1}{V_{t+1}} \ge \frac{1}{V_t} + \gamma_K. \tag{36}$$

Summing from t_0 to $T-1$ gives $1/V_T \ge \gamma_K(T-t_0) + 1/V_{t_0}$, so $V_T = O(1/T)$. The lower gap bound $c_1 r_T \le V_T$ then yields $r_T \le V_T/c_1 = O(1/T)$. □

C.8 Proofs of the main local classifications

Proposition 2.

Proof. With $\lambda = 0$, Theorem 7 classifies strict local maxima in every nondegenerate regime. Lemma 6 gives necessity for asymptotic stability and Theorem 9 gives sufficiency. □

Theorem 1.

Proof. Distance at least two gives stability and the $O(1/t)$ rate by Theorem 9. Otherwise a one-bit rival differs only at coordinate i. On the edge $a_i = s > 0$, $a_{k\neq i} = 0$, the exact factorization (21) has $S_z(a) = s$, so $\Phi(a) - \Phi(0) = -2\log(1+s) + 2\log(1+s) = 0$. These points are also residual zeros. They accumulate at the corner, excluding local attraction and strict local maximality. □

C.9 Direct semantic supervision

Direct semantic supervision changes the semantic corner's linear saturation term but no rival exponent. Its threshold concerns local stability at the semantic anchor, not global recovery.

Corollary 2 (Direct semantic supervision threshold)**.** *Let $+$ be the semantic code and let direct semantic supervision have relative weight $\lambda \ge 0$. Then, uniformly in u,*

$$\Phi_{+,\lambda}(ru) - \Phi_{+,\lambda}(0) = 2r\big[H_+(u) - (1+\lambda)\big] + o(r) \qquad \textit{when } d_{\min}(+) = \tau,$$

so the viability threshold becomes $\chi_+ < 1 + \lambda$. If instead $d_{\min}(+) < \tau$, the offending monomial has degree strictly below one and no finite λ makes $+$ potential-viable.

Proof of Corollary 2. By Remark 1 the anchor is $z = \mathbf{1}_N$ and the supervision term contributes $-2\lambda\sum_i \log(1+a_i) = -2\lambda r + O(r^2)$, so the only change in (25) is the linear coefficient, from -2 to $-2(1+\lambda)$. Every rival exponent $d_{\mathrm{H}}(w,+)/\tau$ is untouched. At $d_{\min}(+) = \tau$ the critical comparison in (27) becomes $H_+(u)$ against $1+\lambda$, giving the threshold $\chi_+ < 1+\lambda$. At $d_{\min}(+) < \tau$ the bound (26) holds with $2r$ replaced by $2(1+\lambda)r$, and since $\alpha < 1$ the positive r^α term still dominates for small r no matter how large λ is. Thus no finite λ makes $+$ potential-viable. □

Combining the threshold with Lemma 6 and Theorem 9 gives the semantic local classification: outside the equality case, local asymptotic stability holds exactly when $d_{\min}(+) > \tau$, or when $d_{\min}(+) = \tau$ and $\chi_+ < 1 + \lambda$. At $\chi_+ = 1 + \lambda$, the first-order expansion does not determine stability.

The direct supervision term with positive weight diverges to $-\infty$ at nonsemantic corners (Remark 1), so the set of viable codes need not be monotone in λ. At the semantic corner, increasing λ can overcome a strict critical-coefficient failure, but cannot overcome $d_{\min}(+) < \tau$.

C.10 CONTROLLED POSTERIOR-REFRESH ERRORS

We retain the parameterized posterior and potential of Appendix B.2 and a fixed positive-definite Gram matrix. An approximate posterior refresh returns code distribution $\widetilde{q}_t$ and unary marginals $\widetilde{p}_t$, inducing

$$\widetilde{R}_t = R(y_t) + e_t, \qquad e_t = 2[\widetilde{p}_t - p(y_t)], \qquad y_{t+1} = y_t + \eta K \widetilde{R}_t. \tag{37}$$

Lemma 8 (Inexact ascent and posterior error)**.** *Write $L_\Phi = \frac{1}{2}(N/\tau + 1 + \lambda)$ and $\|v\|_K^2 = v^\top K v$. For the update (37), suppose $\|e_t\|_K \le \delta \|R(y_t)\|_K$. With $\theta = \eta L_\Phi \lambda_{\max}(K)/2$,*

$$\Phi(y_{t+1}) - \Phi(y_t) \ge \eta[1 - \delta - \theta(1+\delta)^2]\|R(y_t)\|_K^2. \tag{38}$$

For $\mathrm{TV}(\widetilde{q}_t, q(y_t)) \le \varepsilon_t$, $\|e_t\|_K \le 2\sqrt{N\lambda_{\max}(K)}\varepsilon_t$.

Proof. The lower smoothness inequality, Cauchy–Schwarz in the K-norm, and $K^2 \preceq \lambda_{\max}(K)K$ give

$$\begin{aligned}\Phi(y_{t+1}) - \Phi(y_t) &\ge \eta R^\top K(R+e) - \frac{L_\Phi \eta^2}{2}\|K(R+e)\|_2^2 \\ &\ge \eta(1-\delta)\|R\|_K^2 - \frac{L_\Phi \eta^2 \lambda_{\max}(K)}{2}(1+\delta)^2\|R\|_K^2,\end{aligned}$$

which is (38).

If $\mathrm{TV}(\widetilde{q}, q) \le \varepsilon$, each unary label is an event under the code distribution, so $|\widetilde{p}_i - p_i| \le \varepsilon$. Hence

$$\|e\|_K \le \sqrt{\lambda_{\max}(K)}\|e\|_2 \le 2\sqrt{N\lambda_{\max}(K)}\,\varepsilon.$$

□

The same relative-error condition also preserves local endpoint attraction.

Corollary 3 (Endpoint stability under controlled inexact abduction)**.** *Under the assumptions of Theorem 9, replace the exact map by (37). Fix $\delta \ge 0$ such that $\beta_\delta = 1 - \delta - \eta L_\Phi \lambda_{\max}(K)(1+\delta)^2/2 > 0$, and suppose $\|e_t\|_K \le \delta\|R(y_t)\|_K$ at every step. Then the same corner is locally asymptotically stable uniformly over all such time-varying E-step errors: a sufficiently small compactified neighborhood is trapping and $r_t = O(1/t)$.*

Proof of Corollary 3. In the signed coordinates of (28), put $\epsilon_t = D_z e_t$. Since $K_z = D_z K D_z$, orthogonal sign conjugation preserves the relevant norms:

$$\|\epsilon_t\|_{K_z} \le \delta \|g_z(a_t)\|_{K_z}.$$

Lemma 8 and the gradient lower bound (32) therefore give

$$V_{t+1} \le V_t - \eta\beta_\delta \lambda_{\min}(K) c_3^2 r_t^2.$$

It remains to check that an admissible error cannot jump over the local region. Norm equivalence and (32) yield

$$\|\epsilon_t\|_2 \le \delta\sqrt{\frac{\lambda_{\max}(K)}{\lambda_{\min}(K)}}\|g_z(a_t)\|_2 = O(r_t).$$

The inexact signed update $x^+ = x + \eta K_z(g_z + \epsilon_t)$ satisfies $\|\Delta x\|_\infty \le B_{K,\delta} r_t$ for a finite constant $B_{K,\delta}$. Using this bound in (34) gives the same first-exit contradiction, with $r_\star$ reduced if needed.

Finally, substitute $c_{K,\delta} = \eta\beta_\delta\lambda_{\min}(K)c_3^2 > 0$ for c_K in (35)–(36). This proves trapping, uniform attraction, and $r_t = O(1/t)$. Relative to the exact coefficient $c_K = \eta(1-\theta)\lambda_{\min}(K)c_3^2$, the coefficient in the uniform deficit-decrease bound is multiplied by $\beta_\delta/(1-\theta)$. □

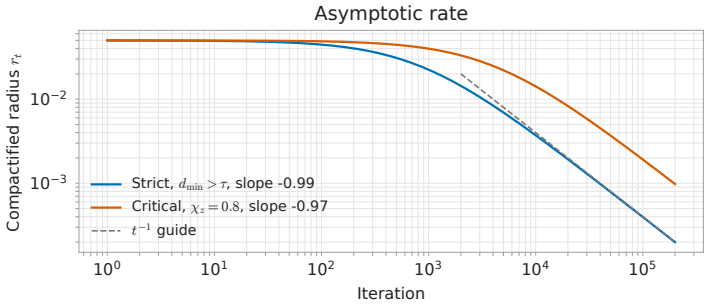


Figure 4: Compactified-radius traces for one strict and one critical stable example, with a t^{-1} guide. Log-log tail slopes are fitted over the final half of each 200,000-update trace.

C.11 Random finite candidate sets

The local criterion guarantees attraction sufficiently close to a viable endpoint. We examine contraction and escape at fixed finite radii, using seed 20260830 to sample candidate sets and initial directions. These initial states are not individually certified to lie in a trapping neighborhood.

For each $N \in \{6, 8, 10, 12\}$, each candidate count in $\{4, 8, 16\}$, and fifteen repeats, we sample distinct binary codes containing the all-positive anchor, giving 180 candidate sets. Strict tests use $\tau = d_{\min} \mp 0.5$ and $\lambda = 0$. Critical tests set $\tau = d_{\min}$ and assign one target χ_z per candidate set. The targets cycle through $(0.6, 0.8, 0.95, 1.05, 1.2, 1.4)$ across the sampled candidate sets, assigning 30 sets to each target. We numerically maximize the base-prior functional and rescale the anchor prior mass to reach the assigned target. This scalar rescaling preserves the maximizing simplex direction. All three regimes for a set use the same $K = I$ and two random SPD Grams with eigenvalues in $[0.35, 1]$, and initial radii $r_0 \in \{0.1, 0.05, 0.01\}$. For each N and τ, all trajectories use the constant step size $\eta = 1.6/(N/\tau + 1)$.

A run is classified as contracting when its sampled maximum radius stays below $1.5r_0$, its final anchor posterior exceeds 0.9, and its final radius is below 0.995 times its midpoint radius (radius is sampled every ten steps). It escapes when its radius reaches $4r_0$ or a rival obtains a log-posterior lead above 10. Other runs are unresolved. An unflagged trajectory runs for 8,000 steps. Updates stop when either escape criterion is detected. All cases run a witness, a dense, and a near-face simplex direction. The nominal witness is supported on a nearest defect for strict tests and is the numerically maximizing direction for critical tests. Before forming finite logits, every direction is floored componentwise at 10^{-10} and renormalized to sum one. For viable cases, we summarize outcomes over all three directions. For non-viable cases, we summarize the witness direction, since non-viability requires escape in some direction, not all.

At each radius, the strict viable and non-viable counts are respectively $180 \times 3 \times 3 = 1620$ and $180 \times 3 \times 1 = 540$. The critical assignments split into 90 viable and 90 non-viable sets, giving $90 \times 3 \times 3 + 90 \times 3 \times 1 = 1080$. Thus the reported comparison includes $1620 + 540 + 1080 = 3240$ trajectories per radius, or $3 \times 3240 = 9720$ overall. We also evaluate the dense and near-face directions of non-viable cases, bringing the total to 14,580 trajectories.

Table 2: Finite-radius outcomes by regime and initial radius r_0, in percent. Resolved runs meet either the contraction or escape criterion. The last three columns report contraction for viable cases and witness-direction escape for non-viable cases, conditional on resolution. The critical and combined rows pool these two outcomes.

	Resolved			Contraction / escape		
Regime (runs per radius)	0.01	0.05	0.10	0.01	0.05	0.10
Strict viable (1,620)	100	100	99.9	99.8	99.6	98.7
Strict non-viable (540)	100	97.2	92.2	100	100	100
Critical (1,080)	95.0	99.4	99.0	99.3	97.5	96.9
All regimes (3,240)	98.3	99.3	98.3	99.7	99.0	98.3

The 9,720 trajectories in this comparison comprise 180 candidate sets. Overall, 98.7% resolve within the update budget. Among resolved runs, 99.0% contract toward a viable corner or escape along

a non-viable corner's witness direction. The remaining 97 runs escape from viable corners: 10 at $r_0 = 0.01$, 33 at $r_0 = 0.05$, and 54 at $r_0 = 0.1$ (Table 2). This increase with radius illustrates why local stability must be distinguished from attraction at a prescribed finite distance. The unresolved runs are also structured: on the strict non-viable side they appear only at the larger radii, and in the critical regime only at $\chi_z \in \{0.95, 1.05\}$, the two values nearest the boundary. In Figure 4, the strict and critical 200,000-step traces have tail slopes -0.995 and -0.971.

D ISOSPECTRAL ROUTING

The routing construction places one slow direction within the candidate-active subspace. One orientation preserves the symmetric route to the semantic code, whereas the others keep a target block negative while the remaining blocks move toward positive labels. Local trapping and finite-horizon continuity extend these routes to open neighborhoods of the initial state and Gram matrix.

D.1 CANDIDATE FAMILY AND STRICT VIABILITY

Fix $m \geq 2$, set $N = 2m$, and partition the coordinates into blocks

$$\text{Block } j = \{2j-1, 2j\}, \qquad j = 1, \ldots, m.$$

Define

$$\mathcal{C}_m = \{z_0, z_1, \ldots, z_m\},$$

where $z_0 = \mathbf{1}_{2m}$ and z_j flips only block j:

$$(z_j)_i = \begin{cases} -1, & i \in \{2j-1, 2j\}, \\ +1, & \text{otherwise.} \end{cases}$$

Take uniform priors, $\tau = 1$, and $\lambda = 0$. The pairwise distances are

$$d_{\mathrm{H}}(z_0, z_j) = 2, \qquad d_{\mathrm{H}}(z_j, z_k) = 4 \quad (j \neq k).$$

Thus $d_{\min}(z) = 2 > \tau$ for every $z \in \mathcal{C}_m$. By Theorem 9, each code has a trapping neighborhood and is locally asymptotically stable for any fixed positive-definite Gram satisfying the ascent condition.

D.2 EXACT BLOCK DYNAMICS AND GENERAL RESIDUAL

Lemma 9 (Exact m-dimensional reduction). *Let $B = [c_1, \ldots, c_m]$, where $c_j = (e_{2j-1} + e_{2j})/\sqrt{2}$, and put $\widehat{K} = B^\top K B$. If $K \succ 0$ preserves $\mathrm{span}(B)$, the block-constant set $y = \sqrt{2}Bu = (u_1, u_1, \ldots, u_m, u_m)$ is invariant, with exact dynamics*

$$u^+ = u + \eta\, \widehat{K}\, r(u),$$

where $r_j(u)$ is the residual of either coordinate in block j, so $R(\sqrt{2}Bu) = \sqrt{2}Br(u)$.

On the block-constant subspace, write

$$y = (u_1, u_1, \ldots, u_m, u_m), \qquad a_j = e^{-u_j}.$$

Lemma 9 shows that every block-preserving Gram leaves this subspace invariant. The complete-code posterior is

$$q_{z_0} = \frac{1}{1 + \sum_{k=1}^m a_k^2}, \qquad q_{z_j} = \frac{a_j^2}{1 + \sum_{k=1}^m a_k^2}.$$

Indeed, within-block exchange symmetry makes $R(y)$ block-constant whenever y is, and a block-preserving K keeps $KR(y)$ in the same subspace. The normalization is $y = \sqrt{2}Bu$ and $R(y) = \sqrt{2}Br(u)$, where $r_j(u)$ is the residual of either coordinate in block j. Multiplication by $B^\top/\sqrt{2}$ therefore gives $u^+ = u + \eta B^\top K B\, r(u)$, proving Lemma 9. The restricted potential $\Phi(\sqrt{2}Bu)$ has Euclidean gradient $2r(u)$, whereas the reduced map uses the per-coordinate residual $R_j = r_j(u)$.

Lemma 10 (General block residual). *For either coordinate in block j, the residual is*

$$\frac{R_j}{2} = \frac{a_j \left(1 + \sum_{k \neq j} a_k^2 - a_j\right)}{(1 + a_j)\,(1 + \sum_{k=1}^m a_k^2)}. \tag{39}$$

Proof. For a coordinate i in block j, every code except z_j assigns $+1$ to coordinate i. Hence its binary marginal is

$$p_i = q_{z_0} + \sum_{k \neq j} q_{z_k} = \frac{1 + \sum_{k \neq j} a_k^2}{1 + \sum_{k=1}^{m} a_k^2}.$$

With $\lambda = 0$, $\tau = 1$, and $\sigma(u_j) = 1/(1 + a_j)$, substitution into $R_i/2 = p_i - \sigma(u_j)$ yields (39). □

D.3 Isospectral Gram construction

Lemma 11 (Full and active-subspace isospectrality)**.** *For every $j = 0, \ldots, m$, K_j has one eigenvalue α and $2m - 1$ eigenvalues equal to one. Its compression $B^\top K_j B = H_j$ has one eigenvalue α and $m - 1$ eigenvalues equal to one. Every K_j is identity on* $\mathrm{span}(B)^\perp$.

For block j, define the orthonormal block-constant and transverse directions

$$c_j = \frac{e_{2j-1} + e_{2j}}{\sqrt{2}}, \qquad f_j = \frac{e_{2j-1} - e_{2j}}{\sqrt{2}}.$$

Put $B = [c_1, \ldots, c_m]$ and $P_\perp = I_{2m} - BB^\top$. Choose $\alpha \in (0, 1)$ and a unit $v \in \mathbb{R}^m$ satisfying $v^\top \mathbf{1}_m = 0$, and set

$$H_0 = I_m - (1 - \alpha) v v^\top, \qquad K_0 = B H_0 B^\top + P_\perp = I_{2m} - (1 - \alpha)(Bv)(Bv)^\top.$$

The vector Bv is unit length. The matrix K_0 has eigenvalue α on Bv and eigenvalue one on its orthogonal complement, so

$$\mathrm{spec}(K_0) = \{\alpha, 1^{[2m-1]}\}, \qquad \mathrm{spec}(H_0) = \{\alpha, 1^{[m-1]}\}.$$

For rival z_j, set

$$H_j = I_m - (1 - \alpha) e_j e_j^\top, \qquad K_j = B H_j B^\top + P_\perp = I_{2m} - (1 - \alpha) c_j c_j^\top.$$

Its slow eigenspace is spanned by c_j, giving

$$\mathrm{spec}(K_j) = \mathrm{spec}(K_0), \qquad \mathrm{spec}(H_j) = \mathrm{spec}(H_0).$$

On the block-constant subspace,

$$K_j c_\ell = \begin{cases} \alpha c_j, & \ell = j, \\ c_\ell, & \ell \neq j, \end{cases}$$

For K_0, the relation $H_0 \mathbf{1}_m = \mathbf{1}_m$ preserves the common rate on the symmetric line. Every K_j, including K_0, is identity on the transverse span of $f_1, \ldots, f_m$. Thus

$$\mathrm{spec}(K_0) = \cdots = \mathrm{spec}(K_m), \qquad \mathrm{spec}(B^\top K_0 B) = \cdots = \mathrm{spec}(B^\top K_m B),$$

Lemma 11 follows: the slow direction can change block alignment while fixing both spectra.

D.4 Semantic routing

Lemma 12 (Symmetric semantic selection)**.** *For a reduced metric H satisfying $H\mathbf{1}_m = k\mathbf{1}_m$ with $k > 0$, every symmetric adverse initialization $u_{1,0} = \cdots = u_{m,0} = s_0 < 0$ converges to z_0.*

Proof. The residual on the symmetric line is a multiple of $\mathbf{1}_m$, and $H\mathbf{1}_m = k\mathbf{1}_m$ keeps this line invariant. Thus $u_{1,t} = \cdots = u_{m,t} =: s_t$ for every t. With $a = e^{-s}$, Lemma 10 reduces to

$$\frac{R(s)}{2} = \frac{a\left(1 + (m-1)a^2 - a\right)}{(1 + a)(1 + ma^2)}.$$

The quadratic $1 + (m-1)a^2 - a$ has discriminant $1 - 4(m-1) < 0$, so $R(s) > 0$ for every finite s. The sequence s_t is strictly increasing. If bounded, it would converge to a finite L, contradicting $s_{t+1} - s_t \to \eta k R(L) > 0$. Hence $s_t \to +\infty$, giving convergence to z_0. □

D.5 RIVAL ROUTING AND THE ABSORBING CONE

Fix a target z_j. Starting from the symmetric initialization and assigning one weight to block j and a common weight to all other blocks preserves the two-variable subspace

$$u_j = u, \qquad u_k = v \quad (k \neq j).$$

Put $a = e^{-u}$, $b = e^{-v}$. Lemma 10 gives

$$\frac{R_u}{2} = \frac{a\left(1 + (m-1)b^2 - a\right)}{(1+a)\left(1 + a^2 + (m-1)b^2\right)},$$
$$\frac{R_v}{2} = \frac{b\left(1 + a^2 + (m-2)b^2 - b\right)}{(1+b)\left(1 + a^2 + (m-1)b^2\right)}.$$

Lemma 13 (m-code absorbing cone)**.** *On this two-variable subspace, the region*

$$\mathcal{R}_j^{(m)} = \left\{a \geq 1 + (m-1)b^2\right\}$$

is forward invariant for every fixed pair of positive effective weights h_u, h_v. Every trajectory entering it satisfies $a_t \to \infty$ and $b_t \to 0$, and hence converges to z_j.

Proof. Inside the cone, $R_u \leq 0$. Moreover, $a \geq 1 + b^2$, so the numerator factor of R_v obeys

$$\begin{aligned} 1 + a^2 + (m-2)b^2 - b &\geq 2 + mb^2 + b^4 - b \\ &> 0, \end{aligned}$$

where $mb^2 - b \geq -1/(4m)$ and $m \geq 2$. Thus $R_v > 0$. For positive effective steps h_u, h_v,

$$a^+ = ae^{-h_u R_u} \geq a, \qquad b^+ = be^{-h_v R_v} < b. \tag{40}$$

It follows that $a^+ \geq a \geq 1 + (m-1)b^2 \geq 1 + (m-1)(b^+)^2$, proving invariance.

The sequences a_t and b_t are respectively nondecreasing and strictly decreasing. Let $w_t = a_t - 1 - (m-1)b_t^2 \geq 0$. Equation (40) makes w_t nondecreasing and $w_t > 0$ after at most one step. If $a_t \to a_\infty < \infty$ and $b_t \to 0$, this positive slack makes R_u eventually negative and bounded away from zero, contradicting finite convergence of a_t. If $a_t \to \infty$ and $b_t \to b_\infty > 0$, then

$$R_v(a_t, b_t) \longrightarrow \frac{2b_\infty}{1 + b_\infty} > 0,$$

contradicting finite convergence of b_t. If both limits are finite, both residuals must vanish. Their numerator equations would be

$$a = 1 + (m-1)b^2, \qquad b = 1 + a^2 + (m-2)b^2.$$

After substitution, the second right-hand side minus b is

$$2 + (3m-4)b^2 + (m-1)^2 b^4 - b \geq 2 + 2b^2 - b > 0,$$

so no such equilibrium exists. The only remaining limits are $a_t \to \infty$ and $b_t \to 0$. □

Lemma 14 (Anisotropic m-code selection)**.** *Fix $s_0 < 0$ and a target-block step $h_1 > 0$ satisfying*

$$h_1 \leq \frac{|s_0|}{4}.$$

Give every non-target block step $h_2 = \rho h_1$. There is a finite $\rho_m^\star(s_0, h_1)$ such that every $\rho > \rho_m^\star(s_0, h_1)$ routes the symmetric initialization to z_j.

Proof. Set $M = e^{-s_0} > 1$ and

$$\mathcal{D}_m = \left[s_0, \frac{s_0}{2}\right] \times [s_0, V_m], \qquad V_m = -\frac{1}{2}\log\frac{\sqrt{M} - 1}{m - 1}.$$

Here $V_m > s_0$, so the rectangle is nonempty. If the trajectory enters the absorbing cone, Lemma 13 gives the desired convergence. The following estimates apply until that entry. Before the trajectory

enters the cone, $a < 1 + (m-1)b^2$, so $R_u > 0$. Thus u_t cannot leave the rectangle through its lower edge. Also $|R_u| \le 2$, so the slow coordinate remains in $[s_0, s_0/2]$ for at least

$$n_\star = \left\lfloor \frac{|s_0|}{4h_1} \right\rfloor \ge 1$$

steps.

While $v \in [s_0, V_m]$, we have $a \in [\sqrt{M}, M]$, $b \le M$, and $b \ge b_{\min} := e^{-V_m} = \sqrt{(\sqrt{M}-1)/(m-1)}$. The numerator factor of R_v is at least $1 + M - b \ge 1$, and its denominator is at most $(1+M)(1+mM^2)$. Hence throughout the rectangle

$$R_v \ge \underline{c}_{2,m} := \frac{2b_{\min}}{(1+M)(1+mM^2)} > 0. \tag{41}$$

Let $c_{2,m} := \min_{\mathcal{D}_m} R_v$, so $c_{2,m} \ge \underline{c}_{2,m} > 0$. Define

$$\rho^\star_m(s_0, h_1) := \frac{V_m - s_0}{n_\star h_1 c_{2,m}}.$$

If $\rho > \rho^\star_m$, the fast coordinate reaches or crosses V_m within $n_\star$ steps, since otherwise its cumulative increase would exceed $V_m - s_0$. At that step $v \ge V_m$, so $b^2 \le (\sqrt{M}-1)/(m-1)$, while $a \ge \sqrt{M}$. Therefore $a \ge 1 + (m-1)b^2$. The trajectory enters $\mathcal{R}^{(m)}_j$, so Lemma 13 gives the result. □

D.6 Integer-aligned parameter construction

Lemma 15 (Compatible parameter choice)**.** *For every $m \ge 2$ and $s_0 < 0$, there are $\rho > 1$, $\alpha = 1/\rho$, and $\eta > 0$ such that $h_1 = \eta\alpha \le |s_0|/4$, $\rho > \rho^\star_m(s_0, h_1)$, and $\eta(2m+1) < 4$.*

Proof. Use $M, V_m, \underline{c}_{2,m}$ from the proof of Lemma 14 and define

$$\bar\rho_m = \frac{4(V_m - s_0)}{|s_0|\underline{c}_{2,m}}, \qquad \rho = 2\max\{1, \bar\rho_m\}, \qquad \alpha = \rho^{-1}.$$

Choose

$$q = \left\lceil \frac{\rho|s_0|(2m+1)}{8} \right\rceil + 1$$

and set

$$h_1 = \frac{|s_0|}{4q}, \qquad \eta = \rho h_1 = \frac{\rho|s_0|}{4q}.$$

Then $\eta\alpha = h_1 \le |s_0|/4$, and

$$\eta(2m+1) < 2 < 4.$$

Moreover, $n_\star = \lfloor |s_0|/(4h_1) \rfloor = q$, so

$$\rho^\star_m(s_0, h_1) \le \frac{V_m - s_0}{q h_1 \underline{c}_{2,m}} = \bar\rho_m < \rho.$$

This choice satisfies the slow-step, routing, and ascent requirements simultaneously. □

D.7 Robust continuation

Lemma 16 (Robust continuation of endpoint selection)**.** *Use the parameterized update (16) and let $F_K(y) = y + \eta K R(y)$. If $\lambda > 0$, take $z = \mathbf{1}_N$. Suppose that a code z satisfies either $d_{\min}(z) > \tau$, or $d_{\min}(z) = \tau$ and $\chi_z < 1 + \lambda$. Suppose also that $F^t_{K_\star}(y_\star) \to z$ for some $K_\star \succ 0$, and that the ascent condition holds strictly at $K_\star$. Then there are open neighborhoods $\mathcal{U}$ of $y_\star$ and $\mathcal{V} \subset \mathbb{S}^N_{++}$ of $K_\star$ such that*

$$(K, y_0) \in \mathcal{V} \times \mathcal{U} \quad \Longrightarrow \quad F^t_K(y_0) \to z.$$

Proof. Let $r_z(y) = \sum_i e^{-z_i y_i}$. Choose a compact SPD neighborhood $\overline{\mathcal{V}}$ of $K_\star$ small enough that, for some $0 < \kappa_- \le \kappa_+ < \infty$,

$$\kappa_- \le \lambda_{\min}(K) \le \lambda_{\max}(K) \le \kappa_+$$

and, for some $\delta > 0$,

$$\frac{\eta L_\Phi \kappa_+}{2} \le 1 - \delta$$

for every $K \in \overline{\mathcal{V}}$. Such a neighborhood exists by the strict baseline inequality and Lemma 4.

The proof of Theorem 9 uses candidate-dependent gap and gradient constants $c_1, c_2, c_3 > 0$. On $\overline{\mathcal{V}}$, its descent coefficient is uniformly bounded below by

$$\eta\delta\kappa_- c_3^2 > 0,$$

and its one-step radial coefficient is uniformly bounded above by $\eta\kappa_+\widetilde{C}_4$. Therefore one may choose a common $\bar{r} > 0$ on which the expansion, gradient, and radial-step bounds in that proof all hold. Every $K \in \overline{\mathcal{V}}$ then has a local basin containing

$$\left\{ y : r_z(y) < \frac{c_1}{c_2}\bar{r} \right\}. \tag{42}$$

Baseline convergence supplies a finite T_0 with

$$r_z\Big(F_{K_\star}^{T_0}(y_\star)\Big) < \frac{c_1}{c_2}\bar{r}.$$

By Lemma 5, the left-hand side remains below the same strict threshold for all (K, y_0) in sufficiently small open neighborhoods $\mathcal{V} \times \mathcal{U}$ of $(K_\star, y_\star)$, with $\mathcal{V} \subset \overline{\mathcal{V}}$. At time T_0, every such trajectory is in the common basin neighborhood (42), and the local theorem gives endpoint convergence to z. □

The Gram neighborhood lies in the full positive-definite matrix cone, so its perturbations may introduce off-diagonal blocks that couple the block-constant and transverse subspaces.

D.8 Proof of Theorem 3

Proof. The family in Section D.1 contains $m + 1$ strictly viable codes. Choose ρ, α, η by Lemma 15 and construct $K_0, \dots, K_m$ as in Section D.3. These Grams are positive definite, pairwise isospectral, and have $\lambda_{\max}(K_j) = 1$. The parameter choice gives $\eta(2m + 1) < 4$, the strict ascent condition for $N = 2m, \tau = 1, \lambda = 0$, so a single step size is valid for all $m + 1$ routes.

Since $H_0\mathbf{1}_m = \mathbf{1}_m$, Lemma 12 routes $y_\star = (s_0, \dots, s_0)$ to z_0. For $j \ge 1$, K_j gives block j step $h_1 = \eta\alpha$ and other blocks step $h_2 = \eta = \rho h_1$. Lemma 14 routes the same $y_\star$ to z_j.

Apply Lemma 16 to each baseline pair $(K_j, y_\star)$. It yields open neighborhoods $\mathcal{U}_j$ and $\mathcal{V}_{m,j}$ with routing to z_j. Their finite intersection

$$\mathcal{U}_m = \bigcap_{j=0}^{m} \mathcal{U}_j$$

is an open neighborhood of $y_\star$ common to every route, proving Theorem 3. □

The spectra in the theorem agree at the centers K_j. The ambient neighborhoods $\mathcal{V}_{m,j}$ include perturbations that can change the spectra and break block-subspace invariance. To preserve both spectra during perturbation, restrict them to the family

$$\mathcal{M}_\alpha = \{I_{2m} - (1 - \alpha)Bww^\top B^\top : \|w\|_2 = 1\}.$$

Each $\mathcal{V}_{m,j} \cap \mathcal{M}_\alpha$ is a relatively open neighborhood of K_j with the same routing conclusion and both spectra fixed. These restricted perturbations preserve the active subspace.

D.9 THE $m = 2$ ILLUSTRATION

For $m = 2$, the codes are

$$z_0 = (+,+,+,+), \qquad z_1 = (-,-,+,+), \qquad z_2 = (+,+,-,-).$$

On $y = (u, u, v, v)$, put $a = e^{-u}$, $b = e^{-v}$, and $Z = 1 + a^2 + b^2$. Equation (39) becomes

$$R_u = \frac{2a(1 + b^2 - a)}{(1 + a)Z}, \qquad R_v = \frac{2b(1 + a^2 - b)}{(1 + b)Z}. \tag{43}$$

For positive diagonal block metrics, the signs in this formula give the three forward-invariant regions

$$\mathcal{R}_0 = \{u > 0,\ v > 0\},$$
$$\mathcal{R}_1 = \{a \geq 1 + b^2,\ b \leq 1 + a^2\},$$
$$\mathcal{R}_2 = \{b \geq 1 + a^2,\ a \leq 1 + b^2\}.$$

They lead respectively to z_0, z_1, z_2. The rival cases are the $m = 2$ instances of Lemma 13. In $\mathcal{R}_0$, both margins increase monotonically. A finite joint limit would require $a = 1 + b^2$ and $b = 1 + a^2$, which have no common positive solution. If only one margin had a finite limit, (43) would keep its drift bounded away from zero. Both margins therefore diverge to $+\infty$, yielding the endpoint z_0.

Let $c_1 = (1, 1, 0, 0)^\top/\sqrt{2}$ and $c_2 = (0, 0, 1, 1)^\top/\sqrt{2}$. Choosing $(1, -1)^\top/\sqrt{2}$ as the slow direction of H_0, the construction specializes to

$$K_0 = I - \frac{1 - \alpha}{2}(c_1 - c_2)(c_1 - c_2)^\top,$$
$$K_1 = I - (1 - \alpha)c_1 c_1^\top, \qquad K_2 = I - (1 - \alpha)c_2 c_2^\top.$$

All three have full spectrum $\{\alpha, 1, 1, 1\}$ and active spectrum $\{\alpha, 1\}$. On the active subspace,

$$H_0 = \frac{1}{2}\begin{pmatrix} 1 + \alpha & 1 - \alpha \\ 1 - \alpha & 1 + \alpha \end{pmatrix}, \qquad H_1 = \mathrm{diag}(\alpha, 1), \qquad H_2 = \mathrm{diag}(1, \alpha).$$

The matrix H_0 has positive entries, so both margins increase in $\mathcal{R}_0$, and the preceding convergence argument also applies there. On the symmetric line $u = v$, H_0 acts as identity, while H_1 slows u and H_2 slows v. Figure 1 uses this construction with $\alpha = 1/6$ and $\eta = 0.02$.

E CAPTURE AND ITS OPTIMIZATION LIMITS

We prove Theorem 4 for arbitrary $\mathcal{C} \subseteq \{\pm 1\}^N$ with $|\mathcal{C}| \geq 2$ and pairwise Hamming distance at least two. Throughout this appendix, priors are uniform, $\tau = 1$, and $\lambda = 0$. The legal evidence in (1) gives

$$\Phi(y) = 2 \log \mathcal{E}(y), \qquad V(y) := -2 \log \mathcal{E}(y).$$

Above half evidence, the signs of the attained state form a legal code. Segment ascent prevents subsequent updates from crossing its sign boundaries, and bounds on the residual give the convergence rate. Appendix E.2 extends Theorem 4 to posterior refresh errors bounded relative to the exact residual, under uniformly positive spectral bounds on the sample Grams.

Lemma 17 (Conditional slices and legality above one half)**.** *Write $s_i = \sigma(y_i)$. For each coordinate i, let A_i and B_i be the probabilities of legal completions of the other coordinates when coordinate i is fixed to $+1$ and -1, respectively. Then*

$$A_i + B_i \leq 1, \qquad \mathcal{E} = s_i A_i + (1 - s_i)B_i \leq \max\{s_i, 1 - s_i\}. \tag{44}$$

In particular, $\mathcal{E}(y) > 1/2$ implies that every coordinate has a strict sign and that $\mathrm{sign}(y) \in \mathcal{C}$.

Proof. The two completion events are disjoint. A common completion would produce two legal codes differing only in coordinate i, violating the distance condition. This proves the first inequality, including for degenerate product distributions, and the second follows by conditioning on coordinate i. If $y_i = 0$, the bound gives $\mathcal{E} \leq 1/2$, so evidence above one half excludes every sign tie.

Orient each coordinate toward its majority bit $z_i = \mathrm{sign}(y_i)$ and write $p_i = \sigma(z_i y_i) > 1/2$. For this orientation let A_i denote the probability of a legal completion with bit z_i, and B_i that with bit $-z_i$. If

$A_i \le B_i$, then

$$\mathcal{E} = \frac{A_i + B_i}{2} + (p_i - \tfrac{1}{2})(A_i - B_i) \le \tfrac{1}{2},$$

a contradiction. Thus setting p_i to one increases the evidence. Repeat this operation coordinate by coordinate. The evidence remains above one half, and disjointness continues to hold at each intermediate product distribution. The final distribution is concentrated at z, so its evidence equals the indicator of $z \in \mathcal{C}$. This indicator must be one because rounding has kept the evidence above one half. □

Lemma 18 (Evidence along an entire update segment)**.** *Let $y_{t+1} = y_t + \eta K_t R(y_t)$ with $0 \preceq K_t \preceq \kappa_+ I$, fixed $\eta > 0$, and $\beta_0 = 1 - \eta\kappa_+(N+1)/4 > 0$. For every $u \in [0,1]$,*

$$\Phi\big(y_t + u\eta K_t R(y_t)\big) - \Phi(y_t) \ge u\eta\beta_0 \|R(y_t)\|^2_{K_t}.$$

Consequently, if $c = \mathcal{E}(y_T) > 1/2$, every subsequent update segment has evidence at least c and stays in the hard orthant of y_T.

Proof. Lemma 2 gives $L = (N+1)/2$. At one step, write $R = R(y_t)$ and $K = K_t$. The spectral upper bound gives $\|KR\|_2^2 \le \kappa_+ \|R\|_K^2$, so lower smoothness yields

$$\Phi(y_t + u\eta KR) - \Phi(y_t) \ge u\eta \left(1 - \frac{u\eta L\kappa_+}{2}\right) \|R\|_K^2 \ge u\eta\beta_0 \|R\|_K^2.$$

Thus the evidence on the segment is no smaller than at its start. By (44), no segment with evidence above one half intersects $y_i = 0$. Induction proves the claim. The inequality uses only the current spectral bound and therefore also holds for metrics chosen from the preceding history. □

These conditional slices bound the potential deficit and residual throughout the captured region, allowing convergence estimates from capture without a separate local trapping argument.

Lemma 19 (Global gap and gradient bounds after capture)**.** *Let $z = \operatorname{sign}(y) \in \mathcal{C}$ and $\mathcal{E}(y) \ge c > 1/2$. Set $a_i = e^{-z_i y_i}$, $r = \sum_i a_i$, and $d = 2c^2(2c-1)$. Then*

$$0 < a_i \le \frac{1-c}{c}, \qquad \frac{2c}{N} r \le V(y) \le 2r, \qquad \|R(y)\|_2 \ge \frac{d}{\sqrt{N}} r.$$

Moreover, $z_i R_i(y) \ge d a_i > 0$ for every coordinate.

Proof. Put $p_i = \sigma(z_i y_i)$. Equation (44) gives $p_i \ge \mathcal{E} \ge c$, proving the coordinate bound. With the conditional slices oriented toward z_i, the proof of Lemma 17 gives $A_i > B_i$. Moreover,

$$\mathcal{E} = \frac{A_i + B_i}{2} + (p_i - \tfrac{1}{2})(A_i - B_i) \le \frac{1}{2} + \frac{1}{2}(A_i - B_i),$$

so $A_i - B_i \ge 2c - 1$. Direct differentiation, together with $\mathcal{E} \le 1$, gives

$$z_i R_i(y) = \frac{2p_i(1-p_i)(A_i - B_i)}{\mathcal{E}} \ge 2c^2(2c-1)a_i = d a_i.$$

Here $p_i(1-p_i) = p_i^2 a_i \ge c^2 a_i$. Taking norms gives the gradient bound since $\|a\|_2 \ge r/\sqrt{N}$.

Since z is legal, its product probability contributes to $\mathcal{E}$, and hence

$$V \le 2 \sum_i \log(1 + a_i) \le 2r.$$

Conversely, $\mathcal{E} \le p_i$ implies $V \ge 2\log(1+a_i) \ge 2a_i/(1+a_i) \ge 2ca_i$ for every i. At least one coordinate satisfies $a_i \ge r/N$, so the bound for that coordinate yields $V \ge (2c/N)r$. □

Proof of Theorem 4. Lemma 17 identifies a unique legal hard code z at time T. Lemma 18 keeps this code fixed and gives

$$V_t - V_{t+1} \ge \eta\beta_0\kappa_- \|R(y_t)\|_2^2.$$

Apply Lemma 19 with $c = \mathcal{E}(y_T) > 1/2$, and set

$$\gamma = \frac{\eta\beta_0\kappa_- d^2}{4N} = \frac{\eta\beta_0\kappa_- c^4(2c-1)^2}{N}.$$

The gap and gradient estimates imply

$$V_{t+1} \le V_t - \frac{\eta\beta_0\kappa_- d^2}{N} r_t^2 \le V_t - \gamma V_t^2.$$

The distance condition makes $\mathcal{C}$ a proper subset of the cube. Finite logits assign positive probability to every vertex, so $0 < \mathcal{E}(y_t) < 1$ and $V_t > 0$. Therefore $1 - \gamma V_t > 0$, and the last recurrence yields $V_{t+1}^{-1} \ge V_t^{-1} + \gamma$. Summing gives the explicit bound

$$r_t \le \frac{N}{2c} \frac{1}{V_T^{-1} + \gamma(t-T)}, \qquad t \ge T. \tag{45}$$

Its constants depend on the initial evidence and the uniform metric bounds, not on the chosen continuation. Thus $r_t \to 0$, and every signed margin diverges to $+\infty$.

Conversely, approaching $z \in \mathcal{C}$ concentrates the independent prediction distribution at that legal code. Hence $\mathcal{E}(y_t) \to 1$, so the trajectory must cross the threshold (13) at a finite iteration. □

E.1 Posterior concentration without capture

Example 2 (Posterior concentration without capture). Take $\mathcal{C} = \{z^+, z^-\}$, where $z^+ = \mathbf{1}_6$, $z^- = (-1,-1,1,1,1,1)$, with native parameters and $y_0 = (1.1, 1.1, 0.01, 0.01, 0.01, 0.01)$. The hard code is z^+, and

$$\log(q_+/q_-) = 2.2 > 2, \qquad q_+ = 0.9002495\ldots,$$
$$\mathcal{E}(y_0) = [\sigma(1.1)^2 + \sigma(-1.1)^2]\sigma(0.01)^4 = 0.0398662\ldots.$$

The four common coordinates cancel from the posterior odds but reduce the legal evidence. Define

$$v_0 = (0,0,1,-1,0,0)^\top/\sqrt{2}, \qquad v_1 = (\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2\sqrt{2}}, \tfrac{1}{2\sqrt{2}}, \tfrac{1}{2\sqrt{2}}, \tfrac{1}{2\sqrt{2}})^\top, \qquad K_j = I - \tfrac{4}{5} v_j v_j^\top.$$

Both Grams have spectrum $\{1/5, 1, 1, 1, 1, 1\}$. With $\eta = 2/7$, the ascent quantity is $\eta\lambda_{\max}(K_j)(N+1) = 2 < 4$, yet K_0 selects z^+ and K_1 selects z^- from this state. The Grams share their full six-dimensional spectrum but differ in their candidate-active compressions.

Proof. The subspace $y = (s, s, r, r, r, r)$ is invariant. Its residual coordinates are $R_a(s) = 2[\sigma(2s) - \sigma(s)]$ and $R_b(r) = 2\sigma(-r)$. On this subspace K_0 acts as identity. For $s > 0$, both residuals are positive, so a finite coordinate limit would leave its increment bounded away from zero. Hence $s, r \to +\infty$. Under K_1 the exact update is

$$s^+ = s + \eta(3R_a/5 - 2\sqrt{2}R_b/5),$$
$$r^+ = r + \eta(3R_b/5 - \sqrt{2}R_a/5).$$

Seventy-digit interval propagation of the full six-dimensional map gives

$$s_{17} \in [-0.04414193676449, -0.04414193676448],$$
$$r_{17} \in [1.54375619445710, 1.54375619445711].$$

The region $s < 0 < r$ is invariant because $R_a < 0 < R_b$: s strictly decreases and r strictly increases. A finite negative limit of s would keep its negative drift away from zero. A finite limit of r would keep its positive drift away from zero. Thus $s \to -\infty$ and $r \to +\infty$, proving selection of z^-. □

Along K_0, the half-evidence test first passes numerically at step 15 ($\mathcal{E}(y_{14}) = 0.4870554\ldots$, $\mathcal{E}(y_{15}) = 0.5074577\ldots$). Switching to K_1 after capture preserves z^+, as Theorem 4 guarantees.

E.2 Capture under changing metrics and controlled errors

Theorem 4 extends to controlled posterior-refresh errors under the same uniform positive spectral bounds on the sample Grams. An error budget relative to the current exact residual preserves ascent along each update segment and prevents crossing the half-evidence barrier.

Theorem 10 (Half-evidence capture with controlled errors)**.** *Let $|\mathcal{C}| \geq 2$ and $d_{\mathrm{H}}(z, w) \geq 2$ for all distinct $z, w \in \mathcal{C}$. Use uniform priors, $\tau = 1$, and $\lambda = 0$. After any finite prefix, allow the continuation*

$$y_{t+1} = y_t + \eta K_t[R(y_t) + e_t], \qquad 0 < \kappa_- I \preceq K_t \preceq \kappa_+ I,$$

with fixed $\eta > 0$, $0 < \kappa_- \leq \kappa_+ < \infty$, and $\delta \geq 0$ satisfying

$$\|e_t\|_{K_t} \leq \delta \|R(y_t)\|_{K_t}, \qquad \beta := 1 - \delta - \frac{\eta\kappa_+(N+1)}{4}(1+\delta)^2 > 0.$$

If the finite state y_T satisfies

$$\mathcal{E}(y_T) > \tfrac{1}{2}, \tag{46}$$

then $z = \operatorname{sign}(y_T)$ is a well-defined member of $\mathcal{C}$. For all $t \geq T$, $\operatorname{sign}(y_t) = z$, and $r_t := \sum_i e^{-z_i y_{i,t}} = O((t - T + 1)^{-1})$, so every signed margin diverges to $+\infty$. The Gram matrices K_t and errors e_t may depend arbitrarily on the training history. Conversely, every trajectory approaching a code endpoint eventually satisfies (46).

Proof. Here $\Phi = 2 \log \mathcal{E}$, so its deficit is $V = -2 \log \mathcal{E}$. Lemmas 17 and 19 remain unchanged. We extend the ascent estimate to include the error along the entire update segment.

Lemma 2 gives $L = (N+1)/2$. At one step abbreviate $R = R(y_t)$, $K = K_t$, and $e = e_t$. Metric-norm Cauchy–Schwarz gives $\langle R, Ke \rangle \geq -\delta \|R\|_K^2$, while

$$\|K(R+e)\|_2^2 \leq \lambda_{\max}(K)\|R+e\|_K^2 \leq \kappa_+(1+\delta)^2\|R\|_K^2.$$

For $u \in [0, 1]$, lower smoothness yields

$$\begin{aligned}\Phi(y_t + u\eta K(R+e)) - \Phi(y_t) &\geq u\eta \left[1 - \delta - \frac{u\eta L\kappa_+}{2}(1+\delta)^2\right] \|R\|_K^2 \\ &\geq u\eta\beta\|R\|_K^2.\end{aligned}$$

Thus the evidence along the segment is no smaller than at its start. By (44), no segment with evidence above one half can intersect a hyperplane $y_i = 0$. Induction proves the claim. The argument applies pointwise to each chosen metric and error, so their dependence on the history introduces no additional condition. Applying Lemma 19 with $c = \mathcal{E}(y_T) > 1/2$ gives the same recurrence as in the exact proof, now with

$$\gamma = \frac{\eta\beta\kappa_- c^4(2c-1)^2}{N} > 0.$$

Consequently,

$$r_t \leq \frac{N}{2c} \frac{1}{V_T^{-1} + \gamma(t-T)} \qquad (t \geq T).$$

This proves uniform capture and convergence over every history-dependent metric and error sequence satisfying the stated bounds. Endpoint convergence gives $\mathcal{E}(y_t) \to 1$, proving eventual passage. □

E.3 Capture with limited control over individual predictions

Corollary 1 covers exact updates with a fixed linear representation of arbitrary rank.

Proof of Corollary 1. Set $K_t = XM_tX^\top$ and $\kappa_+ = m_+\|X\|_{\mathrm{op}}^2$. These matrices satisfy $0 \preceq K_t \preceq \kappa_+ I$, even when X is rank deficient. Lemma 18 therefore preserves the legal code $z = \operatorname{sign}(Xw_T)$ and the evidence lower bound $c = \mathcal{E}(Xw_T) > 1/2$. The proof uses $K_t^2 \preceq \kappa_+ K_t$, which holds for positive-semidefinite matrices even when some sample-space eigenvalues vanish.

Let $\mu = \min_i |(Xw_T)_i| > 0$ and $d_c = 2c^2(2c-1)$. The captured head is nonzero. By the coordinatewise bound in Lemma 19, each exact residual has the same sign as the corresponding

coordinate of Xw_T. Hence

$$\|w_T\|_2\|X^\top R_t\|_2 \geq (Xw_T)^\top R_t = \sum_i |(Xw_T)_i|\,|R_{i,t}|$$
$$\geq \mu\|R_t\|_1 \geq \mu d_c r_t,$$
$$R_t^\top K_t R_t \geq m_-\|X^\top R_t\|_2^2 \geq \frac{m_-\mu^2 d_c^2}{\|w_T\|_2^2} r_t^2.$$

Thus $X^\top R_t$ cannot vanish in the captured region. The separating head w_T supplies a quantitative lower bound for the captured labels, which need not agree with the true semantic labels.

Put $V_t = -2\log\mathcal{E}(Xw_t)$ and $\beta_0 = 1 - \eta\kappa_+(N+1)/4$. Segment ascent and $V_t \leq 2r_t$ give

$$V_{t+1} \leq V_t - \gamma_X V_t^2, \qquad \gamma_X = \frac{\eta\beta_0 m_-\mu^2 d_c^2}{4\|w_T\|_2^2} > 0.$$

Every finite iterate remains finite because $|R_i| \leq 2$, X is fixed, and $M_t \preceq m_+ I$. Thus $V_t > 0$, and the same reciprocal recurrence as in (45) yields

$$r_t \leq \frac{N}{2c}\frac{1}{V_T^{-1} + \gamma_X(t-T)}, \qquad t \geq T. \tag{47}$$

For this captured head, the constants are uniform over all admissible continuations. A trainable bias is a column of X. The representation is exactly $y = Xw$, with no fixed sample-specific offset. □

Disjoint-bag step bound. If the coordinates partition into bags I_j of size at most b, with $\mathcal{C} = \prod_j \mathcal{C}_j$, then $\Phi(y) = 2\sum_j \log\mathcal{E}_j(y_{I_j})$. The Hessian is block diagonal, so applying Lemma 2 to each bag gives

$$\|\nabla^2\Phi(y)\|_{\mathrm{op}} = \max_j \|\nabla^2(2\log\mathcal{E}_j)(y_{I_j})\|_{\mathrm{op}} \leq \frac{b+1}{2}.$$

Consequently, the corollary and (47) hold with $\eta m_+\|X\|_{\mathrm{op}}^2(b+1) < 4$ and $\beta_0 = 1 - \eta m_+\|X\|_{\mathrm{op}}^2(b+1)/4$. The threshold remains the global product $\prod_j \mathcal{E}_j > 1/2$, and the gap constants still use the full N. Only the constraints must factor by bag, whereas X and M_t may couple different bags.

E.4 Identifiability and suboptimal attracting states

Proposition 5 concerns finite attractors that prevent recovery of the unique realizable legal endpoint. We first prove the existence statement in a two-parameter model, then extend the mechanism to regular count-query graphs and fixed image features. In both constructions, the posterior conditions separately on each pair's count, while repeated inputs share predictions through the same model. The soft targets from different pairs are then combined in the parameter update. Jointly normalizing over one logical label variable per distinct input would instead leave the correct assignment as the sole candidate in these identifiable tasks and supply supervised targets.

Proof of Proposition 5. Let a shared predictor assign positive-label probabilities $p_0 = \sigma(w_0)$ and $p_1 = \sigma(w_1)$ to true classes 0 and 1. Use binary sum supervision with total weight ρ on the two mixed class pairs and weight $(1-\rho)/2$ on each homogeneous pair. The mean negative log evidence is

$$\mathcal{L}_\rho(p_0, p_1) = -(1-\rho)\log(1-p_0) - (1-\rho)\log p_1 - \rho\log\big[p_0(1-p_1) + (1-p_0)p_1\big].$$

For $0 < \rho < 1$, its unique zero-loss probability endpoint is $(0, 1)$, since the homogeneous queries force both correct labels. At $(p_0, p_1) = (s, 1-s)$, put $D = s^2 + (1-s)^2$. The probability gradient is $(g, -g)$, where

$$g = \frac{1-\rho}{1-s} - \frac{\rho(2s-1)}{D} = \frac{2s^2 - (2+\rho)s + 1}{(1-s)D}.$$

For $2\sqrt{2} - 2 < \rho < 1$, define

$$\Delta = \sqrt{\rho^2 + 4\rho - 4}, \qquad s_+ = \frac{2 + \rho + \Delta}{4} \in (1/2, 1).$$

At $(s_+, 1-s_+)$, the full two-dimensional probability Hessian has eigenvalues

$$\lambda_- = \frac{\Delta}{(1-s_+)D} > 0, \qquad \lambda_+ = \frac{1-\rho}{(1-s_+)^2} + \frac{2\rho}{D} > 0.$$

The logit Hessian at this stationary point is $[s_+(1-s_+)]^2$ times the probability Hessian, so it is also positive definite. Thus ordinary gradient descent on $\mathcal{L}_\rho$ has a locally asymptotically stable wrong fixed point for every sufficiently small positive step. Its loss is positive, whereas the correct endpoint has loss approaching zero.

To obtain a finite native instance, take $\rho = 0.9$ and 20 disjoint pairs: one $(0,0)$, one $(1,1)$, nine $(0,1)$, and nine $(1,0)$. There are $N = 40$ occurrence coordinates, with semantically oriented feature row $-e_0^\top$ for true class 0 and $e_1^\top$ for true class 1. The resulting X has rank two and $X^\top X = 20 I_2$. Each mixed pair has two legal assignments differing in both labels, so the product candidate set has 2^{18} codes and minimum Hamming distance two. The homogeneous pairs force every shared class prediction, giving $\mathcal{C}_X = \{\mathbf{1}_{40}\}$. The global potential and parameter update are

$$\Phi(Xw) = -40\mathcal{L}_{0.9}(w), \qquad w^+ = w - 40\eta\nabla_w \mathcal{L}_{0.9}(w).$$

At the wrong fixed point, $(p_0, p_1) = (0.8850781\ldots, 0.1149219\ldots)$, and

$$\mathcal{E} = (1-s_+)^4 D^{18} = 2.90834\ldots \times 10^{-6} < 1/2.$$

Its hard predictions violate both homogeneous queries. At $\eta = 0.001$, the update eigenvalues are approximately 0.997105 and 0.995931, and $\eta\|X\|_{\mathrm{op}}^2(N+1) = 0.82 < 4$. □

Regular count-query graphs. The wrong minimum extends to independent prediction coordinates for any number of images admitting the following query design. Let $c_i \in \{0,1\}$ be the true labels of n images, and let the mixed count-one queries form a k-regular bipartite graph G, with $k > 0$, whose edges join opposite classes. The homogeneous queries form a matching covering all images, with each pair belonging to one class. Write E_G for the mixed edge set, $m = |E_G| = nk/2$, and v_i for the predicted probability of the correct label on image i. Weight the mixed-query mean loss by ρ and the homogeneous-query mean by $1-\rho$. The resulting mean loss is

$$\mathcal{L}_\rho(v) = -\frac{2(1-\rho)}{n}\sum_{i=1}^{n}\log v_i - \frac{\rho}{m}\sum_{\{i,j\}\in E_G}\log\big[v_i v_j + (1-v_i)(1-v_j)\big]. \tag{48}$$

Every term is nonnegative, and the homogeneous contribution makes $v = \mathbf{1}_n$ the unique zero-loss probability endpoint.

Proposition 11 (Suboptimal attraction on regular query graphs)**.** *For the count-query design above, suppose $2\sqrt{2} - 2 < \rho < 1$ and set $v_* = (2 - \rho - \sqrt{\rho^2 + 4\rho - 4})/4$. Then $v_*\mathbf{1}_n$ is a strictly suboptimal local minimum of $\mathcal{L}_\rho$ in all n prediction coordinates, and every hard label is incorrect. For any full-row-rank feature matrix $F \in \mathbb{R}^{n\times d}$, a finite head w_* realizes this center, and $w_* + \ker F$ is a locally attracting stationary set under ordinary parameter gradient descent with sufficiently small positive step size. Nearby heads converge to its member with the same initial null-space component.*

Proof. Write $\Delta = \sqrt{\rho^2 + 4\rho - 4}$ and $D = v_*^2 + (1-v_*)^2$. The center satisfies $v_* = 1 - s_+ \in (0, 1/2)$, with s_+ as in the prototype. At $v = v_*\mathbf{1}_n$, every coordinate gradient equals

$$\frac{2}{n}\left[-\frac{1-\rho}{v_*} - \rho\frac{2v_* - 1}{D}\right] = 0.$$

Let A_G be the adjacency matrix and $L_G = kI_n - A_G$ the graph Laplacian. An adjacent off-diagonal Hessian entry is $-\rho/(mD^2)$, and each diagonal entry is

$$\frac{2(1-\rho)}{nv_*^2} + \frac{k\rho(2v_* - 1)^2}{mD^2}.$$

Using $k/m = 2/n$ and the stationary equation gives the full Hessian

$$H_v := \nabla_v^2 \mathcal{L}_\rho(v_*\mathbf{1}_n) = \frac{2\Delta}{nv_*D} I_n + \frac{\rho}{mD^2} L_G \succ 0. \tag{49}$$

Because L_G is positive semidefinite and annihilates the constant vector, $\lambda_{\min}(H_v) = 2\Delta/(nv_*D)$, including disconnected graphs. For semantic logits $u_i = \log(v_i/(1-v_i))$, stationarity gives

$$H_u = [v_*(1-v_*)]^2 H_v \succ 0.$$

This proves strict local minimality in all n prediction directions for the homogeneous reversal branch.

Real image features and redundant parameters. Let $F \in \mathbb{R}^{n\times d}$ be any full-row-rank image-feature matrix, including a bias if used, with raw logits $\ell = Fw$. With $S = \operatorname{diag}(2c_i - 1)$, the semantic logits are $u = S\ell$. The wrong center has $\ell_* = S\mathbf{1}_n \log(v_*/(1-v_*))$, which a finite head w_* realizes by full row rank. Its raw-logit Hessian is $H_\ell = SH_uS \succ 0$, and its parameter Hessian is $F^\top H_\ell F$. For a thin singular-value decomposition $F = U\Sigma V^\top$, its restriction to the full effective parameter space $\operatorname{row}(F)$ is

$$H_{\text{eff}} = \Sigma U^\top H_\ell U \Sigma \succ 0.$$

The remaining $d - n$ directions form $\ker F$ and leave every training prediction unchanged. Ordinary parameter gradients lie in $\operatorname{row}(F)$, so each trajectory preserves its initial null-space component. For $0 < \alpha < 2/\lambda_{\max}(H_{\text{eff}})$, the effective update Jacobian $I - \alpha H_{\text{eff}}$ has spectral radius below one. Hence the affine stationary set $w_* + \ker F$ is locally attracting: nearby heads converge to the member with their own initial null-space component, and their predictions converge to the wrong center. □

Native occurrence representation and step size. For rational ρ, repeat queries an integer number of times so that their normalized frequencies equal the weights in (48). If there are B repeated queries, let X contain their semantically oriented occurrence feature rows and let $\mathcal{C}$ be the product of their legal candidate sets. Every image has total normalized exposure $2/n$, giving

$$\Phi(Xw) = -2B\mathcal{L}_\rho(w), \qquad X^\top X = \frac{2B}{n}F^\top F, \qquad \eta = \frac{\alpha}{2B}.$$

Thus ordinary descent with step α on the mean loss is exactly the native parameter update with step η. Distinct complete candidates differ in at least two occurrence labels. The homogeneous matching uniquely fixes every shared image label, and full row rank realizes the correct signs, so $\mathcal{C}_X = \{\mathbf{1}_{2B}\}$. The occurrence Gram is singular even though F has full row rank on distinct images. For the dense experimental step $\alpha = n/(1.5\|F\|_{\text{op}}^2)$, the two-image bag bound above gives

$$\eta\|X\|_{\text{op}}^2(2+1) = \frac{3\alpha\|F\|_{\text{op}}^2}{n} = 2 < 4.$$

The wrong center violates every homogeneous count and lies below the half-evidence barrier.

F Experimental Protocols and Supplemental Results

The experiments distinguish selection among legal explanations from capture of an attained explanation. We first give the exact CIFAR-10 count model, then examine its selected codes and subsequent metric interventions. We then examine capture with a restricted head and suboptimal attraction under identifiable counts. Finite-radius checks of local stability accompany the corresponding theory in Appendix C.11. Every binary experiment uses exact posterior marginals as detached targets and refreshes them at each update.

F.1 CIFAR-10 count supervision and the exact binary model

We use the binary bird/cat setting of learning from label proportions (Baručić & Kybic, 2022), with a fixed-feature specialization of its complete configuration posterior and detached marginal-target update. The encoder is frozen, the posterior is refreshed at every full-batch head step, and no probability floor or additional supervised loss is used. The original method instead trains the encoder end to end with Adam and refreshes targets once per epoch. Freezing the representation and refreshing at every step make the count-supervised update coincide with the binary dynamics analyzed here.

F.1.1 Disjoint counts and global evidence

Each sampled training set contains 24 official CIFAR-10 training images, partitioned into six bags of four with cat-positive counts $k = (1, 2, 3, 1, 2, 3)$. Let $A \in \{0,1\}^{6\times 24}$ be the bag-membership

matrix, $A_{bi} = 1$ exactly when image i belongs to bag b. With $h_i \in \{0, 1\}$ and complete code $z = 2h - \mathbf{1}$, the candidates are

$$\mathcal{C} = \{2h - \mathbf{1} : Ah = k\} = \prod_{b=1}^{6} \mathcal{C}_b, \qquad |\mathcal{C}| = \left[\binom{4}{1}\binom{4}{2}\binom{4}{3}\right]^2 = 9216.$$

Every change between two legal assignments exchanges a positive and a negative bit in some bag, so $d_{\min}(\mathcal{C}) = 2$. The one-cat and three-cat counts exclude a global label complement. Within-bag exchanges span the zero-sum space, so $S_{\mathcal{C}} = \ker A$ has dimension $24 - 6 = 18$.

The experimental logits and candidates use the fixed cat-positive convention. To obtain the semantic coordinates of Section 2, let $c^\star$ be the true training-label code and $D = \mathrm{diag}(c^\star)$. For raw logits ℓ, set $y = D\ell$, $\mathcal{C}_{\mathrm{sem}} = D\mathcal{C}_{\mathrm{raw}}$, and $K_{\mathrm{sem}} = DK_{\mathrm{raw}}D$. Candidate likelihoods and global evidence are unchanged, and $R_{\mathrm{sem}}(D\ell) = DR_{\mathrm{raw}}(\ell)$. Distances and positive definiteness are preserved, and an active basis P becomes DP, preserving its compressed spectrum as well. This conjugation is used only for analysis. Training and metric construction use raw coordinates without per-image labels.

For raw cat-positive logits y, let $\mathcal{E}_b$ sum the product Bernoulli probabilities of the $\binom{4}{k_b}$ legal assignments in bag b. Disjointness gives

$$\mathcal{E}_{\mathrm{all}}(y) = \prod_{b=1}^{6} \mathcal{E}_b(y_b), \qquad \Phi(y) = 2\sum_{b=1}^{6} \log \mathcal{E}_b(y_b), \qquad R_i(y) = 2\{p_i(y) - \sigma(y_i)\},$$

where p_i is the exact count-posterior marginal for image i. The local posterior marginal equals the marginal of the global complete-code posterior. Capture uses $\sum_b \log \mathcal{E}_b > -\log 2$, so it certifies the entire training code even when off-diagonal Gram entries couple predictions in different bags. An independent Bernoulli-count dynamic program verifies the marginals and evidence values computed by enumerating legal assignments within each bag.

F.1.2 A SHARED HEAD WITH A FULL-RANK SAMPLE METRIC

Features are the 512 outputs before the final layer of torchvision ResNet-18 with supervised ImageNet-1K weights `IMAGENET1K_V1`. The encoder uses evaluation mode, including its pretrained normalization layers. The official deterministic transform resizes the shorter image edge to 256 by bilinear interpolation, center-crops to 224×224, and applies ImageNet channel normalization, without augmentation. Extraction uses PyTorch 2.4.1 and torchvision 0.19.1 in float32, whereas all head updates and posterior calculations use float64. For extracted row f_i, the training design row is $x_i = (f_i/\|f_i\|_2, 1)/\sqrt{2} \in \mathbb{R}^{513}$. All 24 singular values of X are retained, with a required ratio $\sigma_{24}/\sigma_1 > 10^{-8}$. All nine training sets satisfy this condition, with observed ratios between 0.0475 and 0.0507.

Every entry of w_0, including the bias, is independently uniform on $[-1/\sqrt{513}, 1/\sqrt{513}]$. Four heads per training set share the same six metrics, and each head is identical across its metric conditions. For any prescribed $K \succ 0$, the parameter metric

$$M = X^+K(X^+)^\top + (I - X^+X), \qquad XMX^\top = K$$

is positive definite, with identity action on the null space of X. Twice the *sum* of binary cross-entropies against detached p_i has parameter gradient $-X^\top R$. It therefore yields exactly

$$w_{t+1} = w_t + \eta MX^\top R_t, \qquad y_{t+1} = y_t + \eta KR_t, \qquad w_t = w_0 + X^+(y_t - y_0). \tag{50}$$

The effective step $\eta = 0.032$ refers to this residual convention. A mean binary cross-entropy without the factor two would require parameter step $2N\eta$. Independent integration of the parameter updates reproduces the sample-coordinate trajectories with maximum logit discrepancy 3.78×10^{-13}.

F.1.3 FOUR GRAM ORIENTATIONS WITH COMMON FULL AND ACTIVE SPECTRA

Let $P \in \mathbb{R}^{24\times 18}$ contain orthonormal within-bag Helmert contrasts and $T \in \mathbb{R}^{24\times 6}$ contain the normalized bag indicators, with each nonzero entry of T equal to $1/2$. Thus $\mathrm{im}\, P = \ker A$ and $[P\ T]$ is orthogonal. The same bases define every active-spectrum comparison within a training set. Writing $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{18})$,

$$\lambda_j = 4\,16^{-(j-1)/17}, \qquad K_Q = PQ\Lambda Q^\top P^\top + TT^\top,$$

Table 3: Selection and capture across development and confirmation runs. Each head gives six comparisons among four isospectral conditions. Pairs share data and initialization. The 144 confirmation pairs come from 24 initial heads on six training sets.

Quantity	Development	Confirmation
All baseline trajectories	72	144
Isospectral trajectories	48	96
Strictly resolved / captured isospectral trajectories	48/48	96/96
Both-resolved, distinct noncomplement pairs	72/72	144/144
Training sets with such a pair	3/3	6/6
Captured isospectral continuations, changed / executed	0/144	0/288
All captured continuations, changed / executed	0/180	0/360
Raw-reference runs without capture	12	24

Table 4: CIFAR-10 confirmation, 24 trajectories per row. Accuracies are median [minimum, maximum] percentages with cat-positive labels. Reference conditions have separate spectra.

Metric	Resolved	Captured	Train accuracy
Feature	24/24	24/24	50.00 [33.33,100]
Feature-reverse	24/24	24/24	58.33 [50.00,83.33]
Random 1	24/24	24/24	58.33 [41.67,83.33]
Random 2	24/24	24/24	58.33 [41.67,83.33]
Isotropic	24/24	24/24	58.33 [41.67,100]
Raw scaled	0/24	0/24	85.42 [79.17,100]

fixes both the full spectrum $\operatorname{spec}(\Lambda) \cup \{1\}^6$ and the spectrum on $S_{\mathcal{C}}$. The feature condition uses the descending eigenbasis of $P^\top X X^\top P$ for Q. Feature-reverse reverses the assignment of the same λ_j to that eigenbasis. Random 1 and random 2 use two independent Gaussian-QR bases, drawn once per training set with seed $2026090800 +$ set. Gram construction uses input features, counts, and fixed randomness, without a target code, per-image truth, or adaptation to an individual head. An independent calculation verifies agreement of both spectra to floating-point roundoff. The reference metrics are $K_{\text{iso}} = \operatorname{tr}(K_Q)I/24$ and $K_{\text{raw}} = 4XX^\top/\lambda_{\max}(XX^\top)$. Neither reference belongs to the fixed-spectrum comparison. The raw reference corresponds to scaled ordinary parameter gradient descent. The sample identity corresponds to a different preconditioner in parameter space.

F.2 Selected explanations and semantic accuracy

A fixed permutation within each official training class, seed 2026090601, allocates nine disjoint 24-image training sets. Three sets are used for development, followed by six confirmation sets evaluated with the same fixed protocol. Head seeds are $2026090700 + 100\,\text{set} + \text{head}$, with head $\in \{0, 1, 2, 3\}$. Per-image truth supplies the disclosed counts and subsequent accuracy measurements. It does not enter the head loss or metric construction. The cohorts have disjoint training image IDs and heads, while sharing one pretrained encoder. Baselines take 8,000 updates. Figure 2 illustrates feature and feature-reverse from the first confirmation training set and initial head.

A state is well fitted when its hard counts are legal, the mean bag evidence is at least 0.95, and every image's modal probability is at least 0.90. A strictly resolved final endpoint additionally has the same complete hard code at all 101 states from update 7,900 through 8,000.

The four isospectral Grams select distinct, strictly resolved complete codes from each confirmation head (Table 3). Pairwise Hamming distances have minimum 2, median 10, and maximum 14 out of 24 bits. Across the 96 final states, mean bag evidence ranges from 0.9922 to 0.9961, and the smallest modal probability among the 24 images ranges from 0.9929 to 0.9975.

Of the 96 captured isospectral codes, 93 contain at least one incorrect per-image label. Count fitting and capture therefore leave the semantic ambiguity among these explanations unresolved. Table 4 includes both references and reports training-label accuracy using cat as the fixed positive class, without a label swap. The raw reference attains higher median training-label accuracy than the captured conditions, yet none of its 24 runs meets the fitting or capture criteria within the fixed horizon. These outcomes distinguish recovery of the true labels from fitting the counts and preserving a selected legal explanation.

F.3 CIFAR-10 CAPTURE UNDER SUBSEQUENT METRIC CHANGES

The four isospectral metrics and isotropic reference obey $0.25I \preceq K \preceq 4I$, and $\eta\kappa_+(N+1) = 0.032 \cdot 4 \cdot 25 = 3.2 < 4$. The raw reference has its own measured positive lower eigenvalue and the same upper bound. At the first global capture time T_{cap}, we fix the current sign code z and run three separate 4,000-step continuations. The first switches feature and feature-reverse, or random 1 and random 2. The isotropic reference switches to feature. For the other two, choose i minimizing $z_i y_i$ and set

$$u = \frac{z_i e_i + R/\|R\|_2}{\|z_i e_i + R/\|R\|_2\|_2}, \qquad K = 4I - 3.75uu^\top.$$

This Gram minimizes the instantaneous drift of the weakest signed margin relative to the captured code within the stated one-slow-eigenvalue family. When the numerator vanishes an orthogonal direction gives the same maximum Rayleigh value zero. One branch fixes u at capture and the other recomputes it at every step relative to the same captured code. Both retain the uniform spectral bounds, with a full spectrum that differs from the prefix's spectrum. They use the current state and captured code, without future endpoints or per-image ground-truth labels.

The 96 isospectral captures yield 288 executed continuations, and the 24 isotropic captures yield 72 more. None of the 360 changes a captured label or decreases global log evidence beyond 10^{-10}. The raw reference never reaches the capture threshold, so no continuation starts from its runs. Theorem 4 explains the preservation under both fixed and adaptive interventions: every subsequent metric sequence satisfying its uniform conditions must retain the captured code.

F.4 CAPTURE WITH A RANK-DEFICIENT SHARED HEAD

To test capture under a rank-deficient sample Gram, we train 33 parameters on 96 images, arranged in 24 four-image bags with counts $(1, 2, 3)$ repeated eight times. The same frozen ResNet-18 features are row-normalized, centered on the 10,000-image training pool, and projected onto its leading 32 covariance eigenvectors. Dividing by the corresponding standard deviations whitens these coordinates. A bias is appended and all coordinates are scaled by $1/\sqrt{33}$. This preprocessing uses unlabeled training inputs, including confirmation inputs. Per-image labels enter bag construction and evaluation, not projection, head updates or metric design.

We use two development sets and four confirmation sets, all disjoint and excluding the original 216 full-row-rank training images. The protocol is fixed before evaluating the confirmation sets. Each training set uses four heads with all coordinates sampled independently from $\mathrm{Uniform}(-1/\sqrt{33}, 1/\sqrt{33})$. The three parameter metrics are I, a diagonal with 33 geometrically spaced entries from 0.5 to 2, and its reverse. They share the head initialization and the fixed step $\eta = 3.2/[2\|X\|_{\mathrm{op}}^2 \cdot 5]$. The disjoint-bag refinement in Appendix E.3 applies: with four images per bag and $m_+ = 2$, the step satisfies $\eta m_+ \|X\|_{\mathrm{op}}^2 (4+1) = 3.2 < 4$. We update the 33 shared parameters directly according to (14). Each design has rank 33 and sample nullity 63. These parameter metrics need not induce identical sample spectra. All training sets share the fitted projection. An independent count-posterior calculation reproduces every parameter update with maximum discrepancy 1.43×10^{-14} and verifies the captured-head energy lower bound.

Every initial run receives 20,000 updates. The fit rule is legal hard bag counts, mean bag evidence at least 0.95, and minimum image modal probability at least 0.90. Resolution also requires the same code in the last 101 states. Capture occurs when the product of all 24 bag evidences first exceeds $1/2$.

At capture, three 4,000-step continuations use identity, a fixed diagonal, or an adaptive adverse diagonal. The fixed continuation uses the reversed diagonal after an identity or forward-diagonal initial run, and the forward diagonal after a reversed-diagonal initial run. For $g = X^\top R$ and the weakest captured signed margin $z_i(Xw)_i$, the adverse rule sets entry j to 0.5 when $z_i X_{ij} g_j \geq 0$, and to 2 otherwise. It minimizes the coordinate's instantaneous drift within the box using the captured code and current state. The step bound is unchanged.

Table 5 shows that all 48 confirmation runs fit and capture, and all 144 confirmation continuations preserve the captured code. Development contributes 24 captures and 72 unchanged continuations. The original runs also preserve the code after their first crossing. Every captured code in both phases contains at least one semantic-label error.

Table 5: Capture and training accuracy with a restricted shared head. Fit, resolution and global capture counts coincide. Times are descriptive medians. Accuracies use fixed semantic labels. Each development condition has 8 runs and each confirmation condition has 16.

Phase	Parameter metric	Fit and capture	Median T_{cap}	Train acc.
Development	Identity	8/8	2280	91.67%
Development	Forward diagonal	8/8	4682	80.21%
Development	Reverse diagonal	8/8	1569.5	92.71%
Confirmation	Identity	16/16	3163.5	83.33%
Confirmation	Forward diagonal	16/16	3796.5	65.63%
Confirmation	Reverse diagonal	16/16	2355	87.50%

By separating the captured labels within the restricted representation, the head at the first evidence crossing supplies the witness in Corollary 1 for convergence under admissible parameter updates. The semantic errors in these labels persist throughout the tested continuations.

F.5 Suboptimal attraction under identifiable counts

Proposition 11 predicts finite wrong attractors despite uniquely determined labels. We examine attraction near these states and recovery from standard random initializations of the same shared classifier.

Images, features, and identifiable queries. Each sampled training set contains 128 bird and 128 cat images from the official CIFAR-10 training split. Four disjoint training sets support ordinary-head experiments, with one development set followed by three confirmation sets. An ImageNet-pretrained ResNet-18 supplies 512 raw frozen features, and a bias gives a 256×513 design F. Unlike the normalized features in the disjoint-bag experiment, these features undergo no additional normalization, projection, whitening, or fitted adaptation. The input preprocessing is Resize 256, CenterCrop 224, and ImageNet channel normalization, without augmentation.

Every training set has 1,024 mixed pairs forming a connected degree-eight bipartite graph, with one image from each class per edge. We add 128 disjoint homogeneous pairs, covering every image once. The count-constraint matrix has rank 256, proving uniqueness of the satisfying shared hard labels. The feature matrix also has rank 256, and a separate least-squares witness realizes the true labels with unit margin. All four supervised controls, one per frozen-feature training set, reach 100% training concept accuracy on these same image features.

Objective and native updates. We minimize the weighted mean negative log evidence in (48), with ρ the total mixed-query weight rather than the fraction of distinct queries. Each query uses its complete legal-pair posterior, refreshed before every update and held fixed for cross-entropy differentiation. Repeated images share predictions through F, while their query occurrences remain separate in the posterior. Ordinary parameter gradient descent uses no momentum or weight decay. The step is $\alpha = 256/(1.5\|F\|_{\mathrm{op}}^2)$, shared across ρ. Appendix E.4 gives its exact native-update scaling and sufficient two-image bag bound. SVD supplies the scalar step size and the effective coordinates used for curvature diagnostics.

Analytic centers and full effective curvature. For each of the three confirmation training sets and $\rho \in \{0.85, 0.9\}$, set the correct-label probability to $v_* = 1 - s_+$ and realize the resulting raw logits by the minimum-norm head. These diagnostic initializations use the true labels. Every image has wrong-label probability 0.8 or 0.8850781, with corresponding losses 0.81064448 and 0.63739642. Their hard predictions satisfy every mixed query and violate every homogeneous query, giving weighted hard reasoning accuracy ρ and zero concept accuracy. The full logit Hessian is computed independently from explicit legal-pair probability products and compared with the logit-coordinate form of (49). The two calculations agree within 5.9×10^{-17}. Its minimum eigenvalue is 0.000514706 at $\rho = 0.85$ and 0.000565358 at $\rho = 0.9$. Table 6 reports the full 256-dimensional effective parameter Hessian. The other 257 head directions constitute $\ker F$.

Local attraction after perturbation. Each center receives four nonzero perturbations: the weakest and strongest effective-Hessian eigenvectors, an isotropic Gaussian parameter vector, and independent Gaussian logit displacements realized by a minimum-norm head change. Eigenvector signs place their largest-magnitude parameter component above zero. All perturbations are scaled to RMS training-logit displacement 0.01. The unperturbed center and four perturbed heads each receive

Table 6: Curvature and local update stability at the six analytic wrong centers. Here H_{eff} is the Hessian restricted to $\text{row}(F)$, and r_J is the spectral radius of $I - \alpha H_{\text{eff}}$ at the original step. The largest absolute parameter-gradient component across all centers is 1.78×10^{-13}.

Set	ρ	$\lambda_{\min}(H_{\text{eff}})$	$\lambda_{\max}(H_{\text{eff}})$	r_J
1	0.85	0.00297946	165.5725	0.999996177
1	0.90	0.00233763	105.4787	0.999997000
2	0.85	0.00294513	169.1343	0.999996301
2	0.90	0.00231732	107.7610	0.999997090
3	0.85	0.00334100	164.3945	0.999995688
3	0.90	0.00266123	104.7974	0.999996565

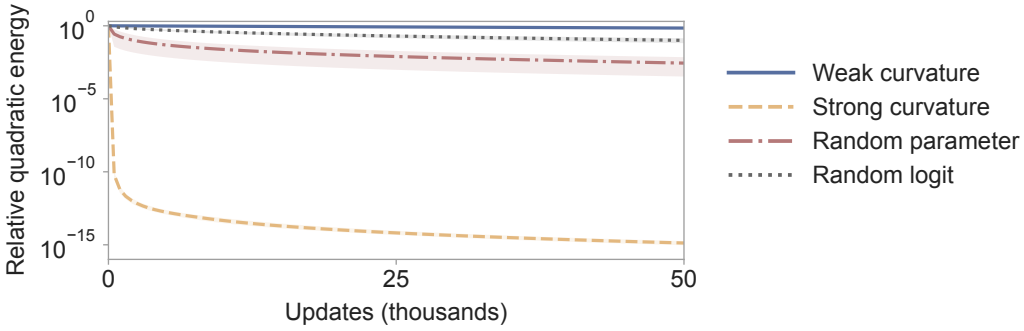


Figure 5: Local quadratic energy along the perturbed paths in Figure 3(b). The energy $\delta\ell^\top H_\ell \delta\ell/2$ uses each center's full logit Hessian and is normalized by its initial value. Lines and bands give medians and full ranges across three training sets and two query weights. The weak-curvature direction contracts slowly, whereas the strong-curvature direction yields much faster energy decay.

Table 7: Recovery from random initialization under identifiable counts. Recovery means all training concept labels are correct at the fixed horizon.

Phase	Runs	Recovered
Development	48	48
Confirmation	144	144

50,000 updates at the original step, giving 30 trajectories across the six centers. Distances and energies are evaluated every 500 updates.

All 24 nonzero perturbations contract in logit distance, probability distance, effective parameter distance, and local quadratic energy, and all retain zero training concept accuracy. The final logit-distance ratios are 0.8061–0.8647 for weak curvature, 3.89×10^{-8}–4.66×10^{-8} for strong curvature, 0.0218–0.0844 for random parameter directions, and 0.2604–0.3759 for random logit directions. Weak-direction quadratic energies retain 0.6499–0.7478 of their initial values, illustrating the slow modes predicted by the near-unit Jacobian eigenvalues. Figure 5 shows the energy decay across the four perturbation types. The strongest-direction loss excess reaches floating-point roundoff, while its quadratic energy remains a resolved distance diagnostic. Null-space drift is at most 1.80×10^{-13}, and unperturbed-center RMS logit drift stays below 7.0×10^{-14}.

Recovery from random initialization. For each training set, 12 standard random linear heads are shared across $\rho \in \{0.5, 0.7, 0.85, 0.9\}$. Weights and bias are sampled uniformly from $[-1/\sqrt{512}, 1/\sqrt{512}]$, and all trajectories receive 30,000 full-batch updates. Table 7 reports recovery for both cohorts.

All 192 random starts recover the correct labels. Successful recovery from these random starts thus coexists with local attraction to incorrect predictions under the same identifiable supervision.