GEOMETRY IS NOT THE TRAP: A SPECTRAL-REOPENING DIAGNOSTIC FOR PROTOCOL-BOUNDED OPTIMIZATION LOCKING

Anonymous authors
Paper under double-blind review

ABSTRACT

While practice often equates the geometric collapse of deep representations with a loss of recoverability, and existing analyses frequently rely on proxies such as Neural Collapse, the relationship between whole-manifold rank-one geometric cliffs and persistent protocol lock under cross-entropy (CE)—the inability of CE to recover under a given protocol and budget—remains unclear. We show that what distinguishes recoverability under CE is not static forward geometry, but a dynamical object: whether the full-extractor Feature Neural Tangent Kernel (Feature-NTK) reopens. Across the convolutional networks we test, trained on standard image-classification benchmarks and covering residual networks, Visual Geometry Group (VGG) networks, and Group Normalization, standard CE recovers from the tested whole-manifold rank-one geometric cliffs. Among checkpoints with severe Feature-NTK spectral cut-offs induced by structured auxiliary regularizers, some checkpoints can still reopen the Feature-NTK and recover to approximately 80% accuracy, so a spectral cut-off is not sufficient for persistent protocol lock. The critical bifurcation between persistent protocol lock and recoverable cut-off is whether the Feature-NTK reopens under a controlled 50-minibatch CE spectral-reopening probe. Through controlled parameter interventions, we further show that maintaining the lock relies on a layer-structured batch normalization (BN) scale configuration. Finally, after freezing the rule, we prospectively evaluate unseen cut-off checkpoints across architectures. On the checkpoints we test, persistent protocol lock tracks whether the Feature-NTK reopens rather than static geometry.

1 INTRODUCTION

Representation geometry is widely used to characterize the states reached by deep-network optimization. Neural Collapse (NC), for example, concentrates within-class features on class means while preserving a class-structured discriminative subspace (Papyan et al., 2020; Han et al., 2022). Under their respective assumptions, existing analyses characterize NC as a benign or preferred terminal state of cross-entropy (CE) training with weight decay (Zhu et al., 2021; Jacot et al., 2025; Wu & Mondelli, 2025; Súkeník et al., 2025), even when lower-rank solutions exist (Súkeník et al., 2024; Garrod & Keating, 2025). These results do not make every severe collapse a certificate of irrecoverability. We instead study a *whole-manifold rank-one geometric cliff*, where the sampled manifold concentrates near one uncentered direction, and distinguish it from *persistent protocol lock*: failure of CE to recover from checkpoint θ_0 under a declared protocol $\mathcal{P}$ and budget T. Protocol lock is therefore a property of $(\theta_0, \mathcal{P}, T)$, not of geometry alone. Our central question is what evidence determines whether a collapsed checkpoint remains recoverable.

The most direct static candidate is forward geometry. Yet geometry records only the features at θ_0; the CE update also depends on how parameter Jacobians align with label residuals. On CIFAR-10, we construct rank-one cliffs across batch-normalized Residual Networks (BN-ResNets), a Visual Geometry Group network with batch normalization (VGG-BN), and a residual network with Group Normalization (GroupNorm) (He et al., 2016; Simonyan & Zisserman, 2015; Wu & He, 2018; Ioffe & Szegedy, 2015). Standard CE leaves the cliff on all nine tested cross-architecture checkpoints. Thus even an extreme static forward state is insufficient to diagnose persistent protocol lock.

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.

A stronger static candidate is the full-extractor empirical Feature Neural Tangent Kernel (Feature-NTK) (Jacot et al., 2018; Lee et al., 2019), which adds Jacobian information and describes local parameter sensitivity. Nevertheless, checkpoints with nearly identical severe spectral cut-offs exhibit opposite long-horizon outcomes: some remain locked, whereas another reopens its spectrum and recovers to 79.2% accuracy. A static Feature-NTK cut-off is therefore also insufficient. Geometry describes the current representation, and the Feature-NTK describes current local sensitivity; recoverability additionally depends on whether useful sensitivity can re-emerge under CE.

This failure of static criteria motivates a dynamical diagnostic. We introduce a controlled 50-minibatch pure-CE spectral-reopening probe that tests whether sample directions re-enter the Feature-NTK. It agrees with all eight retrospective long-horizon outcomes. After its thresholds are frozen, it also correctly predicts all five prospective outcomes, including four cross-architecture cut-offs not used to form the rule. Dynamic spectral reopening therefore tracks protocol-bounded recoverability on the tested checkpoints more directly than either static geometry or a static cut-off. The diagnosis remains tied to a checkpoint, protocol, and budget rather than serving as a universal trainability certificate.

The probe identifies which checkpoints remain locked; controlled parameter interventions then localize what maintains the tested BN locks. At matched global root-mean-square (RMS) amplitude, preserving each checkpoint's layerwise BN-scale profile keeps four checkpoints locked at 10.0% accuracy, whereas flattening the profile immediately reopens the Feature-NTK and enables recovery to 78.2%–84.2%. The mechanism is therefore not global BN-scale amplitude, but a checkpoint-specific layerwise scale configuration coupled to the surrounding network.

Our contributions are as follows.

- Through cross-architecture counterexamples, we show that neither a whole-manifold rank-one cliff nor a static Feature-NTK cut-off is sufficient to diagnose persistent protocol lock. Our result challenges a geometry-only reading of extreme static signatures and separates forward collapse, local sensitivity, and protocol-bounded recoverability.
- We propose a dynamical diagnostic—a controlled 50-minibatch CE spectral-reopening probe. To the best of our knowledge, this is the first use of dynamic Feature-NTK spectral reopening as a criterion for diagnosing protocol lock. The probe agrees with the long-horizon outcomes of all eight retrospective checkpoints; after we freeze its decision thresholds, it also correctly predicts all five prospective outcomes, including four previously unseen cross-architecture cut-off checkpoints.
- Using within-checkpoint interventions at matched global RMS, we rule out global BN-scale amplitude as the explanation for the tested locks. Preserving each checkpoint's layerwise RMS profile maintains four locks, whereas flattening the profile immediately reopens the Feature-NTK and enables CE recovery.

2 Related Work

Neural collapse and static representation geometry. Neural collapse (NC) describes a class-structured terminal geometry: within-class variability vanishes, while the centered class means approach a simplex equiangular tight frame in a $(C-1)$-dimensional subspace (Papyan et al., 2020; Han et al., 2022; Mixon et al., 2022). Under their respective model and regularization assumptions, unconstrained-feature and end-to-end analyses characterize this geometry as a global optimum or a dynamically preferred state of cross-entropy (CE) training with weight decay (Zhu et al., 2021; Jacot et al., 2025; Wu & Mondelli, 2025; Súkeník et al., 2025); NC can also persist when lower-rank solutions exist (Súkeník et al., 2024; Garrod & Keating, 2025). These works establish the formation, optimality, or persistence of a class-structured geometry rather than an equivalence between a whole-manifold rank-one cliff, a Feature-NTK spectral cut-off, and failure to recover under a fixed optimization protocol. Consequently, this literature does not establish whether static forward geometry predicts recovery under a fixed CE protocol and budget.

Tangent-kernel dynamics and plasticity. The neural tangent kernel (NTK) organizes output Jacobians into a sample Gram matrix whose eigenspectrum governs linearized training dynamics (Jacot et al., 2018; Lee et al., 2019); label alignment with its eigenvectors can further control conver-

gence rates (Eshtehardian et al., 2026). In the lazy regime the model remains close to its linearization and the kernel is nearly fixed (Chizat et al., 2019), whereas feature-learning regimes and finite networks can exhibit substantial, data-dependent NTK evolution (Geiger et al., 2020; Fort et al., 2020). We use that evolving spectral object differently: the scalarized full-extractor Feature-NTK tests whether sample-discriminative directions reappear after a severely cut-off checkpoint, rather than certifying an infinite-width limit or a width requirement. Related work on plasticity distinguishes current performance from the capacity to learn from subsequent data and connects plasticity loss to optimization geometry (Lyle et al., 2023); in continual reinforcement learning, reduced plasticity has also been associated with declining empirical-NTK rank (Tang et al., 2025). Those studies primarily concern changing tasks or data streams. Consequently, existing NTK and plasticity results do not provide a short, same-task diagnostic of recovery from a fixed collapsed checkpoint.

Normalization and protocol-dependent recovery. Batch normalization (BN) changes activation and gradient scales, reshapes the optimization landscape, and can stabilize rank propagation in deep networks (Ioffe & Szegedy, 2015; Santurkar et al., 2018; Daneshmand et al., 2020). We isolate a narrower question: whether lock maintenance depends on the layer-structured profile of BN affine scales rather than their global amplitude. Adaptive Moment Estimation (Adam) changes the coordinatewise update metric (Kingma & Ba, 2015), and, under specific near-minimizer assumptions, stochastic gradient descent (SGD) and Adam can favor different sharpness measures (Li et al., 2025). That literature studies solution selection after the loss is already small; we use optimizer changes only to delimit recovery protocols from collapsed, low-accuracy checkpoints. Finally, Federated Averaging (FedAvg) supplies a standard aggregation baseline, while separate work studies durable poisoning and robust aggregation in parameter or gradient space (McMahan et al., 2017; Zhang et al., 2022; Allouah et al., 2025; Yu & Fang, 2026). We use federated aggregation as an out-of-distribution stress test for the checkpoint-construction recipe: the resulting induction boundary clarifies that the auxiliary objective is a diagnostic instrument, not a proposed attack or a claim about other threat models. Consequently, prior work on normalization, optimizer bias, and federated robustness does not jointly isolate whether a checkpoint-specific layerwise BN-scale configuration maintains protocol-bounded lock.

3 Preliminaries and Setup

3.1 Model and protocol

A classifier maps an input x_i through an extractor f_θ to a feature $z_i = f_\theta(x_i) \in \mathbb{R}^d$, where d is the feature dimension, and then through a linear head. The head has a weight matrix W and a bias b, and it produces logits $u_i = Wz_i + b \in \mathbb{R}^K$ and probabilities $p_i = \mathrm{softmax}(u_i)$. Here K is the number of classes and $y_i \in \mathbb{R}^K$ is the one-hot label. On a minibatch of n samples the cross-entropy is $\mathcal{L}_{\mathrm{CE}} = -n^{-1} \sum_{i=1}^{n} y_i^\top \log p_i$. The extractor parameters are $\theta \in \mathbb{R}^P$, where P is their dimension. For a BN network, θ includes the affine scale γ and shift β of every batch normalization (BN) layer (Ioffe & Szegedy, 2015); its running means and variances are buffers rather than optimizer parameters. The extractor Jacobian of sample x_i is $J_i = \nabla_\theta f_\theta(x_i) \in \mathbb{R}^{d \times P}$. Recovery and the short probe update the extractor and the head jointly under $\mathcal{L}_{\mathrm{CE}}$; the identities below differentiate only θ.

The main experiments use CIFAR-10, the 10-class image benchmark of Krizhevsky and Hinton, with BN-ResNet-18. Transfers use ResNet-34, VGG-11-BN, and ResNet-18 with GroupNorm in place of BN (ResNet-18-GN) (Krizhevsky & Hinton, 2009; He et al., 2016; Simonyan & Zisserman, 2015; Wu & He, 2018). ResNet-18-GN has learnable GroupNorm affine parameters but no BN-scale vector; the BN-specific interventions below therefore do not apply to it. Data augmentation, the 30-image Feature-NTK probe, and the 60×8 dense check are in Appendix B.

We use two CE-only recovery protocols. The block-reset protocol rebuilds momentum stochastic gradient descent (SGD) every round and runs 100 rounds of 60 updates. The persistent protocol restores the checkpoint's momentum buffer and runs for 100 epochs (39,100 updates). Both use learning rate 0.05, momentum 0.9, and batch size 128; unless stated otherwise, weight decay is 5×10^{-4}. Each result names the protocol and budget used. Operationally, a checkpoint *recovers* if it leaves the relevant geometric cliff or spectral cut-off and improves accuracy within the declared budget; it need not regain its pre-collapse accuracy. A checkpoint exhibits *persistent protocol lock*

if low accuracy, collapsed forward geometry, and a cut-off Feature-NTK persist to the end of the declared budget. For BN networks we additionally record the layerwise scale state, with global root-mean-square (RMS) amplitude $\mathrm{RMS}(\gamma) = \|\gamma\|_2/\sqrt{|\gamma|}$. Here $|\gamma|$ denotes the number of entries in the concatenated BN-scale vector.

Because protocol lock is defined relative to a checkpoint, protocol, and budget, we also include boundary and stress tests rather than treating either the checkpoint construction or the diagnostic as universal. We test transfer to CIFAR-100, removal of residual skip connections, replacement of BN by GroupNorm, twelve additional recovery runs, and persistent protocols using SGD, Adam, and Adam with decoupled weight decay (AdamW). We also test federated induction with $N = 10$ total clients and $m = 1$ intervened client, together with a two-branch Federated Averaging (FedAvg) sweep whose auxiliary-branch weight is $w_{\mathrm{int}} \in \{0.5, 0.8\}$. These tests delimit the claim: the construction also produces a locked cut-off checkpoint on CIFAR-100, but does not produce a cut-off without residual skips, with GroupNorm, or in the tested federated settings; selected locks persist across the additional runs and persistent protocols, whereas large-step block-reset Adam provides a protocol-dependent escape. The full record is in Appendix D. Thus our diagnostic concerns recovery of a specified checkpoint under a declared protocol, and the auxiliary construction is neither universal nor a proposed federated attack.

3.2 Geometry and the Feature-NTK

On a set of n features, a minibatch or a probe set, the uncentered second moment and the centered covariance are

$$M_z = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^\top, \qquad C_z = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^\top,$$

where $\bar{z} = n^{-1}\sum_i z_i$. If μ_j are the nonnegative eigenvalues of M_z, the spectral participation ratio is

$$r_{\mathrm{PR}}(M_z) = \frac{(\sum_j \mu_j)^2}{\sum_j \mu_j^2}.$$

For a nonzero positive semidefinite matrix of rank r, $1 \le r_{\mathrm{PR}} \le r$, with equality to 1 exactly when the spectrum is a single spike. This is not the matrix stable rank $\|M\|_F^2/\|M\|_2^2$. If $M_z = 0$, the participation ratio is mathematically undefined; for logging only, we record it as 0 and treat it as vanishing-feature collapse rather than a rank-one spectrum. The condition $r_{\mathrm{PR}}(M_z) \approx 1$ alone states that the uncentered feature energy is nearly one-dimensional; it does not by itself force C_z to vanish. For the anchor-based constructions used below, Lemma 1 supplies the stronger centered-collapse statement.

The scalarized Feature-NTK is a kernel on samples, not on feature coordinates. For Q unit directions $v_q \in \mathbb{R}^d$,

$$\widehat{\Theta}_{ij} = \frac{1}{Q}\sum_{q=1}^{Q}\langle J_i^\top v_q,\, J_j^\top v_q\rangle.$$

The headline spectra use $Q = 1$. Let $\lambda_1 \ge \lambda_2 \ge 0$ be the leading eigenvalues of $\widehat{\Theta}$, and write $\rho = \lambda_2/\lambda_1$ when $\lambda_1 > 0$. Numerical rank counts eigenvalues of at least $10^{-3}\lambda_1$. If $\lambda_1 = 0$, the rank is recorded as 0 and ρ is left undefined. For the threshold decisions below, an undefined ρ does not satisfy a strict "greater than" clause. The block Feature-NTK $\mathcal{K}_{ij} = J_i J_j^\top \in \mathbb{R}^{d\times d}$ retains those feature coordinates (Jacot et al., 2018). For isotropic unit directions, $\mathbb{E}[\widehat{\Theta}_{ij}] = d^{-1}\,\mathrm{tr}(\mathcal{K}_{ij})$. The CE residual passed back to the features is $r_i = W^\top(p_i - y_i)$.

Four operational terms use these statistics and are not interchangeable. A whole-manifold rank-one geometric cliff is the forward state $r_{\mathrm{PR}}(M_z) \approx 1$. A Feature-NTK spectral cut-off has numerical rank 0, or numerical rank 1 with $\rho \le 10^{-6}$; thus a vanishing measured kernel is included by convention. The gate does not mean $J_\theta f = 0$. Operationally, we label a tested checkpoint persistently protocol-locked when low accuracy, collapsed geometry, and a cut-off Feature-NTK all persist to the end of its declared recovery budget. BN scales are an additional mechanistic measurement only for BN networks. The controlled CE spectral-reopening probe is 50 minibatches of pure CE; ρ_0 and ρ_{50} denote ρ before the probe and after those 50 steps. Our probe asks whether ρ_{50} has left the cut-off

gate. It is an operational criterion, not a universal trainability certificate. This protocol-bounded definition makes the claim directly falsifiable: every verdict is attached to a specified checkpoint, optimizer, update schedule, and budget, rather than inferred from a static image of the representation.

Lemma 1 (Anchor neighborhood). *Suppose every feature is a deviation δ_i from a common anchor c, with $z_i = c + \delta_i$ and $\|\delta_i\|_2 \le \varepsilon$. Then $\|M_z - cc^\top\|_2 \le 2\|c\|_2\varepsilon + \varepsilon^2$ and $\mathrm{tr}(C_z) \le \varepsilon^2$. If the operator norm $\|W\|_{\mathrm{op}}$ is finite, $\|u_i - u_j\|_2 \le 2\|W\|_{\mathrm{op}}\varepsilon$. At an exact anchor, $\varepsilon = 0$, and on balanced labels the deterministic argmax accuracy is $1/K$.*

Lemma 2 (CE extractor gradient). *$\nabla_\theta \mathcal{L}_{\mathrm{CE}} = n^{-1}\sum_i J_i^\top r_i$, and $\|\nabla_\theta \mathcal{L}_{\mathrm{CE}}\|_2^2 = n^{-2}\sum_{i,j} r_i^\top \mathcal{K}_{ij} r_j$.*

Lemma 3 (Jacobian-norm bound). *If $\|J_i\|_{\mathrm{op}} \le \varepsilon_J$ for every sample in the batch, then $\|\nabla_\theta \mathcal{L}_{\mathrm{CE}}\|_2 \le (\varepsilon_J/n)\sum_i \|r_i\|_2$.*

Lemma 1 explains centered collapse for an anchor neighborhood but does not constrain J_i. Lemma 2 shows that a nonzero Jacobian can still cancel against the residuals r_i. Lemma 3 requires the operator-norm bound ε_J; a small ρ does not supply that bound. The same collapsed features can sit on any positive semidefinite Jacobian Gram (Appendix A). We run the probe with persistent SGD for 50 minibatches of pure CE, using learning rate 0.05, momentum 0.9, and zero weight decay, and then remeasure $\widehat{\Theta}$. Writing $r = \nabla_u \mathcal{L}_{\mathrm{CE}}$ for the gradient of the loss with respect to the logits, and θ_{feat} for the extractor parameters, the CE-aligned energy is

$$E_{\mathrm{CE}} = \frac{\|\nabla_{\theta_{\mathrm{feat}}} \mathcal{L}_{\mathrm{CE}}\|_2^2}{\|r\|_2^2 + \epsilon},$$

where the stabilizer $\epsilon > 0$ only prevents division by zero. E_{CE} is an instantaneous quantity and does not replace ρ_{50}. Proofs are in Appendix A.

4 Geometry Is Not a Protocol Lock

We begin with the strongest geometry-only hypothesis: that an extreme whole-manifold rank-one cliff is already sufficient to diagnose optimization lock. NC retains class-discriminative structure in a $(K-1)$-dimensional subspace (Papyan et al., 2020; Han et al., 2022), whereas the cliffs studied here concentrate the whole sampled feature set near one uncentered direction. For the anchor constructions, Lemma 1 explains both the centered collapse and the class-prior accuracy, but it does not determine whether later CE updates can leave that state. We answer this recovery question experimentally. Three auxiliary conditions on BN-ResNet-18, defined in Table 1 of Appendix B, all reach $r_{\mathrm{PR}}(M_z) \le 1.007$. With only an output anchor, the 60×8 dense check has $\rho = 0.716$ and numerical rank 13, and 20 block-reset rounds reach accuracy 0.433. Residual-map suppression and an input-sensitivity penalty, without a scale penalty, end at dense $\rho = 0.194$ and rank 14, with accuracy still 0.248, consistent with logits that have not fully collapsed; 100 rounds reach accuracy 0.829. Adding a squared penalty on γ produces an almost exact anchor, accuracy 0.100 and $r_{\mathrm{PR}} = 1.000$, but the dense ratio is $\rho = 1.32 \times 10^{-3}$, above the cut-off gate. Twenty rounds reach accuracy 0.431. In all three cases the leading Feature-NTK eigenvalue λ_1 stays in the tens or hundreds, so the measured kernel has not vanished. Lemma 2 shows why the forward participation ratio alone cannot determine the CE step: the step also depends on $n^{-1}\sum_i J_i^\top r_i$. Scale penalties alone, for 24 rounds, lower the global γ RMS to 0.071 and leave $r_{\mathrm{PR}} = 2.63$; the cliff never forms, and 100 rounds of CE reach accuracy 0.841.

The same output anchor on ResNet-34, VGG-11-BN, and ResNet-18-GN, three seeds each, gives nine cliffs with $1.001 \le r_{\mathrm{PR}} \le 1.023$ and $1.77 \times 10^{-3} \le \rho \le 0.883$ (He et al., 2016; Simonyan & Zisserman, 2015; Wu & He, 2018). All nine leave the geometric cliff and resume learning under 20 block-reset rounds, reaching accuracy from 0.241 to 0.453 (Figure 1; see Table 2 in Appendix D for checkpoint-level results). None of these spectra meets the cut-off gate. Thus, across the tested architectures, a whole-manifold rank-one cliff is not sufficient for persistent protocol lock. The geometry-only hypothesis is therefore rejected; the next section tests the stronger static candidate supplied by a Feature-NTK spectral cut-off.

5 A Dynamical Spectral-Reopening Diagnostic

The geometric cliffs above miss the stricter gate $\mathrm{rank}(\widehat{\Theta}) \le 1$ and $\rho \le 10^{-6}$, where the rank is the numerical rank defined above. To reach that gate, we jointly activate forward-state penalties and

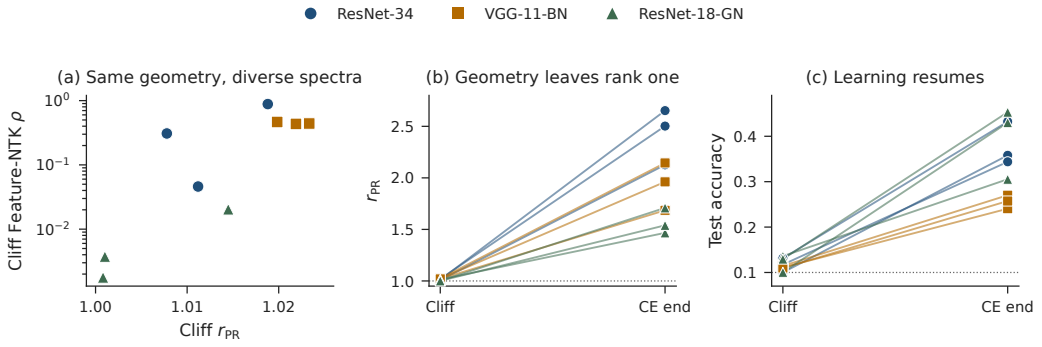

Figure 1: Nine output-anchor cliffs and their endpoints after 20 rounds of block-reset CE. (a) Nearly identical rank-one geometry coexists with Feature-NTK ratios spanning almost three orders of magnitude. (b–c) Every run leaves rank one and improves beyond the class-prior accuracy. Marker shape and color encode architecture; each architecture has three seeds.

a penalty on the BN scales. Let c be a detached feature anchor, let $\mathcal{R}_h$ penalize spatial variation inside selected residual feature maps, and let $\mathcal{R}_{\text{lip}}$ penalize the change in features under a small input perturbation. The coefficients $\lambda_c, \lambda_h, \lambda_{\text{lip}}$ weight those three terms, and $\lambda_{\gamma,2}, \lambda_{\gamma,1}$ weight the squared and absolute penalties on γ:

$$\mathcal{L}_{\text{trap}} = \mathcal{L}_{\text{CE}} + \lambda_c \|f_\theta(x) - c\|_2^2 + \lambda_h \mathcal{R}_h + \lambda_{\text{lip}} \mathcal{R}_{\text{lip}} + \lambda_{\gamma,2} \|\gamma\|_2^2 + \lambda_{\gamma,1} \|\gamma\|_1. \tag{1}$$

We use this auxiliary objective solely to construct controlled diagnostic checkpoints; it is not proposed or evaluated as an attack method. The full coefficient vector $(\lambda_c, \lambda_h, \lambda_{\text{lip}}, \lambda_{\gamma,2}, \lambda_{\gamma,1})$ is $(2, 8, 4, 8, 8)$. The variant that keeps only the absolute ℓ_1 penalty on γ among the scale terms, which we call the L1-scale variant, uses $(2, 8, 4, 0, 4)$. Below, L1 and L2 denote absolute ℓ_1 and squared ℓ_2 scale penalties, respectively. We do not claim that any component is individually necessary: the full factorial ablation was not run, and the corresponding forward-state or scale penalties used separately in Section 4 remain above the gate. The full and L1-scale settings each produce a checkpoint that remains locked for 100 block-reset rounds, at accuracy 0.100 and numerical rank 1, with final ρ equal to 2.23×10^{-17} and 6.90×10^{-9}. After those recovery runs, the corresponding global RMS values are 0.012 and 0.081: the scales need not vanish.

We deliberately construct heterogeneous checkpoints rather than an automatic endpoint, allowing us to compare nearby static signatures with different recovery outcomes. Across ten deterministic seeds of the full objective, run for at most 60 rounds, three seeds enter the gate: seed 9 at round 16 ($\rho = 6.24 \times 10^{-17}$), seed 2 at round 22, and seed 6 at round 28 (Figure 2; see Table 3 in Appendix D). The other seven never cross 10^{-6}. Their terminal γ RMS can be smaller than that of the seeds that do enter, so global RMS does not sort the gate. This selectivity is itself informative: global scale

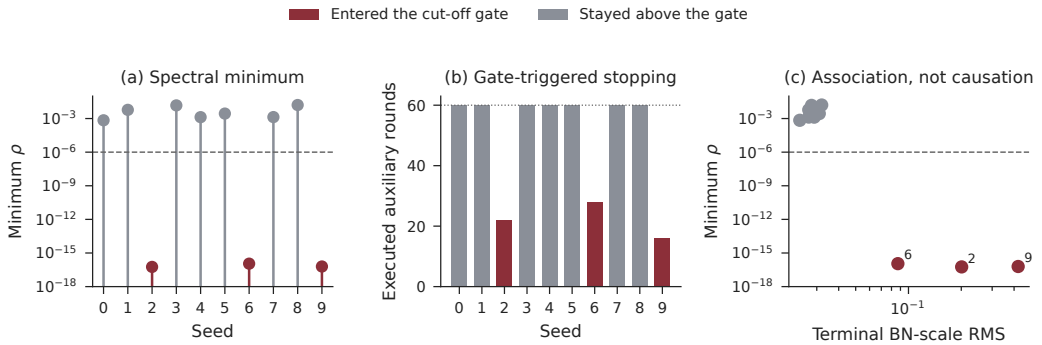

Figure 2: Heterogeneous induction across ten BN-ResNet-18 seeds. (a) Seeds 2, 6, and 9 cross $\rho \leq 10^{-6}$; the other seven stay above the gate. (b) Gate entry triggers early stopping, whereas the remaining runs exhaust the budget. (c) Terminal global BN-scale RMS is associated with, but does not by itself establish, gate entry; Section 6 tests scale configuration by matched interventions.

does not determine entry into the gate, and the gate need not be absorbing. Continuing the objective past the window can reopen the kernel: one run ends at rank 5 and $\rho = 0.314$, and 20 CE rounds then reach accuracy 0.376.

The three cut-off seeds then follow the same 100-round block-reset pure CE protocol. Seeds 2 and 6 remain at accuracy 0.100, $r_{\mathrm{PR}} = 1$, and rank 1. Seed 9 starts from a ratio of the same order as the two that remain locked, $\rho = 6.24 \times 10^{-17}$, and reaches accuracy 0.792, $r_{\mathrm{PR}} = 4.05$, numerical rank 30 (the 30-image probe is saturated), and $\rho = 0.267$. Seed 9 is a direct counterexample to sufficiency: a static cut-off can reopen and recover. The five severe cut-offs—the two joint-setting checkpoints and seeds 2, 6, and 9—also have nonzero step-0 extractor-gradient norms, ranging from 3.97 to 24.3. Their initial ratios all lie between 3×10^{-17} and 1.3×10^{-16}, so neither a vanished measured gradient nor the static value of ρ explains the split.

These paired outcomes motivate the central diagnostic question: not how small the spectrum is at the checkpoint, but whether it begins to reopen under CE. We test this dynamical quantity with 50 minibatches of persistent pure CE. On a retrospective set of eight checkpoints—the five severe cut-offs and three cliffs from Section 4—we define the rule

$$\text{predict recovery} \iff \rho_0 > 10^{-6} \quad \text{or} \quad \rho_{50} > 10^{-3}, \tag{2}$$

which agrees with the block-reset outcome in all eight cases (Figure 3; see Table 4 in Appendix D). The first clause classifies checkpoints that never entered the cut-off gate; among the five cut-offs, the decision depends entirely on spectral reopening at step 50. The locked values stay near 10^{-16}, while seed 9 rises to 0.365. Three minibatch orders give the same decision on all 15 pairs. On this small retrospective set, ρ separates the outcomes at step 50, whereas aligned energy, gradient norm, and accuracy do so only by step 200. Five probe optimizers, with the thresholds held fixed, each match the rule on these five checkpoints. Repeated orders and optimizers are not additional inductions. The spectrum is used here as a recoverability criterion, not as a width or speed certificate (Eshtehardian et al., 2026).

We then freeze both thresholds before evaluating new inductions, separating rule construction from prospective evaluation and preventing outcome-dependent threshold adjustment. Among ten new BN-ResNet-18 seeds, only seed 15 enters the gate, with $\rho_{50} = 0.0281$. Our frozen rule predicts recovery before the long run; 100 block-reset rounds reach accuracy 0.7565, rank 30, and $\rho = 0.366$.

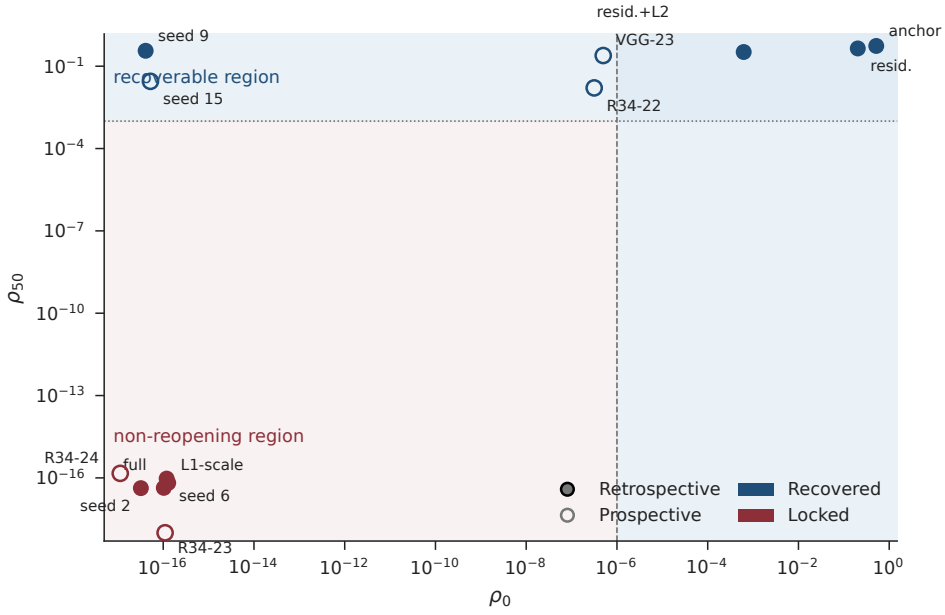

Figure 3: ρ_0 against ρ_{50}. Filled markers are the eight retrospective checkpoints; open markers were scored after the rule was frozen. Blue denotes recovery, whereas red denotes remaining locked. Lines mark $\rho_0 = 10^{-6}$ and $\rho_{50} = 10^{-3}$. Background shading shows the corresponding recoverable and non-reopening decision regions. A recorded zero is drawn at 10^{-18}.

The same full objective on three further architectures yields cut-offs on $3/5$ ResNet-34 seeds, $1/5$ VGG-11-BN seeds, and $0/5$ ResNet-18-GN seeds (He et al., 2016; Simonyan & Zisserman, 2015; Wu & He, 2018). The GroupNorm record bounds this recipe; it is not a protection theorem. Together with seed 15, the four cross-architecture cut-offs form five prospective cases. The frozen rule predicts three recoveries and two locks (see Table 5 in Appendix D). For the four cross-architecture cases, both 100-round block-reset CE and 100-epoch persistent pure CE agree with the predictions; seed 15 agrees under the declared block-reset outcome. The two predicted recoveries leave by epoch 3; the two predicted locks finish $39{,}100$ updates at accuracy 0.100 and rank 1. One of those locks ends at $\rho = 1.01 \times 10^{-16}$ with $\lambda_1 = 3.84 \times 10^4$. The resulting $5/5$ prospective agreement provides held-out support for the diagnostic on the tested checkpoints; broader architectures and protocols are needed to establish its range beyond this study.

6 Layer-Structured BN-Scale Configuration

The reopening probe identifies *which* checkpoints remain locked; the next question is *what* maintains that dynamical state. We localize the effect through controlled parameter interventions that preserve global scale while changing its allocation across layers. Only the BN scales γ are edited; other parameters and, unless noted, the running statistics stay fixed (Ioffe & Szegedy, 2015). Recovery is the same 100-round block-reset pure CE. The four locked checkpoints do not share a global RMS: 0.0427, 0.299, 0.201, and 0.0868. Setting every γ on the full-objective checkpoint to one positive constant, from 0.05 to 1, leads to recovery, but that edit changes both the amplitude and the arrangement across layers. Reestimating running statistics alone, with γ untouched, leaves accuracy 0.100 and rank 1. This intervention is therefore insufficient to release the tested lock, and a uniform assignment is not evidence of an RMS threshold.

Figure 4 separates global amplitude from layerwise allocation at a matched global RMS. Inside each BN layer, setting every channel to that layer's own RMS preserves the vector of per-layer RMS values while removing within-layer differences. On the four locked checkpoints the kernel remains cut off at the edit, and 100 rounds stay at accuracy 0.100 and rank 1. Setting every channel in every layer to the checkpoint's original global RMS also leaves the global RMS unchanged, but removes differences between layers. The kernel reopens at the edit, while accuracy is still near 0.10; after 100 rounds the accuracies are 0.782, 0.842, 0.816, and 0.820. Seed 9, which already recovers, recovers under both edits. The four locked profiles are not one shape: on the full objective the early BN layers sit near 10^{-4} and the fourth residual stage (layer4) sits near 0.04–0.07, whereas the L1-scale checkpoint has global RMS 0.299. Thus "preserving the profile" means preserving each checkpoint's own per-layer RMS allocation, not imposing one shared template.

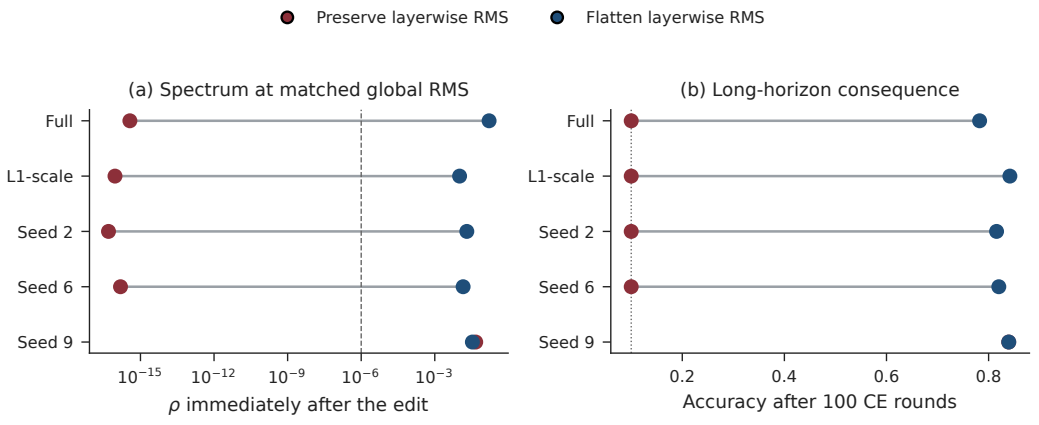

Figure 4: Within-checkpoint matched-RMS interventions on BN scales. Each line connects edits with the same global RMS. Preserving the layerwise RMS profile leaves four checkpoints below the spectral gate and locked at accuracy 0.100; flattening RMS across layers immediately reopens their spectra and enables recovery. Seed 9 is the recoverable control under both edits.

Rescaling frozen profiles across RMS values 0.05–0.20 yields no threshold shared across checkpoints. Within-layer permutations and cross-checkpoint transplants further show that the relevant object is not a universal template or a stand-alone γ vector, but a checkpoint-dependent configuration coupled to the surrounding network. Together, these interventions localize lock maintenance to layer-structured BN-scale allocation while ruling out global RMS as the explanation; the full numerical record is in Appendix D.

7 Conclusion

In this work, we challenge the prevailing diagnostic view that an extreme static signature is reliable evidence of protocol lock, propose a dynamical diagnostic based on a controlled 50-minibatch CE spectral-reopening probe, and further identify a layer-structured scale mechanism that maintains the tested BN locks. Our results support a basic distinction: collapse describes the state of a checkpoint, whereas protocol lock describes whether a declared training protocol and budget can leave that state. Empirically, all nine cross-architecture whole-manifold rank-one cliffs we test escape collapse under CE. The probe agrees with all eight retrospective long-horizon outcomes and, after its thresholds are frozen, correctly predicts all five prospective outcomes. Controlled interventions sharpen this diagnosis into a mechanistic finding: at matched global RMS, preserving the layerwise BN-scale profile keeps four checkpoints locked at 10.0% accuracy, whereas flattening that profile enables recovery to 78.2%–84.2%. By separating static forward geometry, local parameter sensitivity, and protocol-bounded recoverability, our work recasts protocol lock as a dynamical reachability problem and provides a concrete target for theory: characterizing when a declared training protocol permits or prevents Feature-NTK spectral reopening. Future work should test this diagnostic across broader architectures, normalization schemes, datasets, and optimization budgets.

AI Use Statement

AI Assistance: Yes, to aid or polish writing. Generative AI tools were limited to language and wording edits. The authors take responsibility for the scientific content and final text.

Ethics Statement

This diagnostic study uses only public CIFAR benchmarks, with no human subjects or private data. The auxiliary regularizer constructs checkpoints and is not presented as a deployable attack; the tested federated settings did not induce lock from a healthy model. The work distinguishes representation collapse from failure to recover under a declared training protocol.

Reproducibility Statement

Appendix B documents the datasets, architectures, protocols, budgets, checkpoint-construction objectives, and random seeds. Appendix C specifies the scalarized Feature-NTK estimator and reopening probe; Appendix D reports checkpoint-level outcomes; and Appendix A provides complete proofs.

References

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Ahmed Jellouli, Geovani Rizk, and John Stephan. Adaptive gradient clipping for robust federated learning. In *International Conference on Learning Representations*, 2025.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2933–2943, 2019.

Hadi Daneshmand, Jonas Kohler, Francis R. Bach, Thomas Hofmann, and Aurelien Lucchi. Batch normalization provably avoids rank collapse for randomly initialised deep networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Seyed Mohammad Eshtehardian, Mohammad Hossein Yassaee, and Babak Khalaj. On the convergence of two-layer Kolmogorov-Arnold networks with first-layer training. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=buuwRBYfrP.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: An empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5850–5861, 2020.

Connall Garrod and Jonathan P. Keating. The persistence of neural collapse despite low-rank bias. In *Advances in Neural Information Processing Systems*, 2025.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (11):113301, 2020. doi: 10.1088/1742-5468/abc4de.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to Simplex ETF. In *International Conference on Learning Representations*, 2022.

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.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Arthur Jacot, Peter Súkeník, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. In *International Conference on Learning Representations*, 2025.

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

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

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2511.02773.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211. PMLR, 2023.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Ag"uera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20:11, 2022. doi: 10.1007/s43670-022-00027-5.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Peter Súkeník, Christoph Lampert, and Marco Mondelli. Neural collapse versus low-rank bias: Is deep neural collapse really optimal? In *Advances in Neural Information Processing Systems*, 2024.

Peter Súkeník, Christoph H. Lampert, and Marco Mondelli. Neural collapse is globally optimal in deep regularized ResNets and transformers. In *Advances in Neural Information Processing Systems*, 2025.

Hongyao Tang, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Glen Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58883–58904. PMLR, 2025.

Diyuan Wu and Marco Mondelli. Neural collapse beyond the unconstrained features model: Landscape, dynamics, and generalization in the mean-field regime. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision*, pp. 3–19, 2018.

Jingfan Yu and Zhixuan Fang. Near optimal robust federated learning against data poisoning attack. In *International Conference on Learning Representations*, 2026.

Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 26429–26446, 2022.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems*, 2021.

A Derivations and Diagnostic Identities

We prove the three lemmas of Section 3 and the identities that separate forward geometry, the scalarized Feature-NTK, and the cross-entropy extractor gradient. None implies that a geometric cliff or spectral cut-off is a protocol lock or is irreversible.

A.1 Spectral participation ratio

Proposition 1. *Let $A \succeq 0$, $A \neq 0$, with nonnegative eigenvalues λ_j and $r = \operatorname{rank}(A)$. Then*

$$r_{\mathrm{PR}}(A) = \frac{(\sum_j \lambda_j)^2}{\sum_j \lambda_j^2} = \frac{\operatorname{tr}(A)^2}{\|A\|_F^2},$$

and $1 \leq r_{\mathrm{PR}}(A) \leq r$. The lower bound is equality if and only if exactly one eigenvalue is nonzero. The upper bound is equality if and only if every nonzero eigenvalue is equal.

Proof. Since A is symmetric and positive semidefinite, $\mathrm{tr}(A) = \sum_j \lambda_j$ and $\|A\|_F^2 = \sum_j \lambda_j^2$, which gives the second formula. Nonnegativity of the eigenvalues expands the square as

$$\Big(\sum_j \lambda_j\Big)^2 = \sum_j \lambda_j^2 + 2\sum_{j<k} \lambda_j \lambda_k \geq \sum_j \lambda_j^2.$$

Dividing by $\sum_j \lambda_j^2 > 0$ yields $r_{\mathrm{PR}}(A) \geq 1$. Every cross term vanishes if and only if at most one λ_j is positive. Because $A \neq 0$, that means exactly one eigenvalue is nonzero.

For the upper bound, discard the zero eigenvalues and apply Cauchy–Schwarz to the remaining r numbers:

$$\Big(\sum_{j=1}^{r} \lambda_j\Big)^2 \leq r \sum_{j=1}^{r} \lambda_j^2.$$

Dividing by $\sum_j \lambda_j^2$ yields $r_{\mathrm{PR}}(A) \leq r$. Equality holds in Cauchy–Schwarz if and only if those r eigenvalues are equal. □

A.2 Proof of the anchor-neighborhood lemma

Proof of Lemma 1. Write $z_i = c + \delta_i$ and $\bar{\delta} = n^{-1} \sum_i \delta_i$. Expanding the uncentered moment gives

$$\begin{aligned} M_z &= \frac{1}{n} \sum_i (c + \delta_i)(c + \delta_i)^\top \\ &= cc^\top + c\bar{\delta}^\top + \bar{\delta} c^\top + \frac{1}{n} \sum_i \delta_i \delta_i^\top. \end{aligned}$$

The identity $\|ab^\top\|_2 = \|a\|_2 \|b\|_2$ and the triangle inequality give

$$\begin{aligned} \|M_z - cc^\top\|_2 &\leq \|c\bar{\delta}^\top\|_2 + \|\bar{\delta} c^\top\|_2 + \Big\|\frac{1}{n} \sum_i \delta_i \delta_i^\top\Big\|_2 \\ &\leq 2\|c\|_2 \|\bar{\delta}\|_2 + \frac{1}{n} \sum_i \|\delta_i\|_2^2. \end{aligned}$$

The hypothesis $\|\delta_i\|_2 \leq \varepsilon$ implies $\|\bar{\delta}\|_2 \leq \varepsilon$ and $\|\delta_i\|_2^2 \leq \varepsilon^2$, so $\|M_z - cc^\top\|_2 \leq 2\|c\|_2 \varepsilon + \varepsilon^2$.

The centered mean is $\bar{z} = c + \bar{\delta}$, and therefore

$$C_z = \frac{1}{n} \sum_i (\delta_i - \bar{\delta})(\delta_i - \bar{\delta})^\top.$$

Taking the trace,

$$\mathrm{tr}(C_z) = \frac{1}{n} \sum_i \|\delta_i - \bar{\delta}\|_2^2.$$

The mean minimizes the sum of squared distances, so replacing $\bar{\delta}$ by 0 only increases the sum:

$$\mathrm{tr}(C_z) \leq \frac{1}{n} \sum_i \|\delta_i\|_2^2 \leq \varepsilon^2.$$

Finally, $u_i - u_j = W(z_i - z_j) = W(\delta_i - \delta_j)$, and the bias cancels. Hence

$$\|u_i - u_j\|_2 \leq \|W\|_{\mathrm{op}} \big(\|\delta_i\|_2 + \|\delta_j\|_2\big) \leq 2\|W\|_{\mathrm{op}} \varepsilon.$$

□

Proposition 2 (Exact anchor). *If $\varepsilon = 0$ in Lemma 1, then every sample has the same logits and the same class probabilities. With a fixed argmax tie-breaking rule, every sample is assigned to one class $k^\star$, and the empirical accuracy equals the frequency of that class. On perfectly balanced labels the frequency is $1/K$.*

Proof. The condition $\varepsilon = 0$ forces $z_i = c$ for every i, so $u_i = Wc + b$ does not depend on i. A deterministic argmax with fixed tie-breaking therefore returns the same index $k^\star$ on every sample, and the accuracy is the empirical frequency of class $k^\star$. If each class has n/K samples, that frequency is $1/K$. □

The hypothesis of Lemma 1 does not contain J_i. The bound does not imply that the Jacobian is small, that the cross-entropy gradient vanishes, that the current parameter is a local minimum, or that a later declared protocol cannot recover.

A.3 A COLLAPSED OUTPUT DOES NOT DETERMINE THE JACOBIAN GRAM

Proposition 3. *Let $K \in \mathbb{R}^{n\times n}$ be an arbitrary positive semidefinite matrix. There exists a parameter θ_0 and a model, differentiable on a finite sample $\{x_i\}_{i=1}^n$, such that $f_{\theta_0}(x_i)$ is the same vector for every i, while the Gram matrix of one scalar feature coordinate equals K.*

Proof. Factor $K = AA^\top$ with rows $a_i^\top \in \mathbb{R}^P$. Define a scalar coordinate by the local linear model

$$s_\theta(x_i) = c_0 + a_i^\top(\theta - \theta_0)$$

and set every other feature coordinate equal to a constant. At $\theta = \theta_0$, one has $s_{\theta_0}(x_i) = c_0$ for every i, so the vector-valued feature $f_{\theta_0}(x_i)$ is the same vector for every sample. Differentiating the scalar coordinate gives $\nabla_\theta s_{\theta_0}(x_i) = a_i$, and therefore

$$\big[\langle \nabla_\theta s_{\theta_0}(x_i), \nabla_\theta s_{\theta_0}(x_j)\rangle\big]_{ij} = AA^\top = K.$$

The remaining coordinates contribute nothing to this Gram. Thus identical forward features do not determine the Gram of the extractor Jacobian. □

A.4 THE SCALARIZED FEATURE-NTK

Recall $J_i = \nabla_\theta f_\theta(x_i) \in \mathbb{R}^{d\times P}$, $\mathcal{K}_{ij} = J_i J_j^\top$, and, for directions $v_q \in \mathbb{R}^d$,

$$g_{iq} = J_i^\top v_q, \qquad \widehat{\Theta}_{ij} = \frac{1}{Q}\sum_{q=1}^{Q}\langle g_{iq}, g_{jq}\rangle.$$

Proposition 4. *For any fixed directions $v_1, \dots, v_Q$, the matrix $\widehat{\Theta}$ is positive semidefinite.*

Proof. Let $G_q \in \mathbb{R}^{n\times P}$ have ith row $g_{iq}^\top$. Then $\big[\langle g_{iq}, g_{jq}\rangle\big]_{ij} = G_q G_q^\top \succeq 0$. A convex combination of positive semidefinite matrices is positive semidefinite, so $\widehat{\Theta} = Q^{-1}\sum_{q=1}^{Q} G_q G_q^\top \succeq 0$. □

Proposition 5. *If $\mathbb{E}[vv^\top] = d^{-1}I_d$, then*

$$\mathbb{E}_v\big[\langle J_i^\top v, J_j^\top v\rangle\big] = \frac{1}{d}\operatorname{tr}(J_i J_j^\top),$$

and therefore $\mathbb{E}[\widehat{\Theta}_{ij}] = d^{-1}\operatorname{tr}(\mathcal{K}_{ij})$. In particular, the identity holds when v is uniform on the unit sphere.

Proof. The scalar identity $\langle J_i^\top v, J_j^\top v\rangle = v^\top J_i J_j^\top v$ and the cyclic property of the trace give

$$v^\top J_i J_j^\top v = \operatorname{tr}(J_i J_j^\top vv^\top).$$

Taking the expectation and using linearity of the trace,

$$\mathbb{E}_v\big[v^\top J_i J_j^\top v\big] = \operatorname{tr}\big(J_i J_j^\top\,\mathbb{E}[vv^\top]\big) = \frac{1}{d}\operatorname{tr}(J_i J_j^\top).$$

Averaging Q independent copies does not change the expectation.

If v is uniform on the unit sphere, its law is invariant under every orthogonal transformation, so $\mathbb{E}[vv^\top] = cI_d$ for some c. Taking the trace and using $\|v\|_2 = 1$ gives $1 = \mathbb{E}[\|v\|_2^2] = \operatorname{tr}(\mathbb{E}[vv^\top]) = cd$, hence $c = d^{-1}$. □

Proposition 6. *Let $\widehat{\Theta} = U\Lambda U^\top$ with orthonormal eigenvectors u_k and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. For every $a \in \mathbb{R}^n$,*

$$a^\top \widehat{\Theta} a = \sum_k \lambda_k \langle a, u_k \rangle^2.$$

If $\lambda_1 > 0$, $\rho = \lambda_2/\lambda_1$, and $a \perp u_1$, then $a^\top \widehat{\Theta} a \leq \rho \lambda_1 \|a\|_2^2$.

Proof. Expand $a = \sum_k \langle a, u_k \rangle u_k$. Then

$$a^\top \widehat{\Theta} a = a^\top U \Lambda U^\top a = \sum_k \lambda_k \langle a, u_k \rangle^2.$$

If $a \perp u_1$, the $k = 1$ term vanishes. For every $k \geq 2$ one has $\lambda_k \leq \lambda_2$, so

$$a^\top \widehat{\Theta} a \leq \lambda_2 \sum_{k \geq 2} \langle a, u_k \rangle^2 = \lambda_2 \|a\|_2^2 = \rho \lambda_1 \|a\|_2^2.$$

□

Proposition 7. *For every $M > 0$ there exist Jacobians $J_i \in \mathbb{R}^{d \times P}$, identical across samples, such that $\|J_i\|_{\mathrm{op}} = M$ and, for a suitable unit direction v, the scalarized kernel $\widehat{\Theta}$ has rank one and every entry equal to M^2.*

Proof. Let $a \in \mathbb{R}^P$ and $u \in \mathbb{R}^d$ be unit vectors, and set $J_i = M u a^\top$ for every i. Then $\|J_i\|_{\mathrm{op}} = M$. Choosing the probe direction $v = u$ gives $g_i = J_i^\top u = Ma$, so $\langle g_i, g_j \rangle = M^2$ for every pair (i, j). Hence $\widehat{\Theta} = M^2 \mathbf{1}\mathbf{1}^\top$, which has rank one. The constant M may be arbitrarily large, so rank one does not upper-bound the Jacobian norm. □

A.5 Proof of the cross-entropy gradient identity

Proof of Lemma 2. The per-sample loss is $\ell_i = -y_i^\top \log p_i$, and $r_i = W^\top (p_i - y_i) = \nabla_{z_i} \ell_i$. The chain rule through $z_i = f_\theta(x_i)$ gives

$$\nabla_\theta \ell_i = J_i^\top r_i.$$

Averaging over the batch,

$$\nabla_\theta \mathcal{L}_{\mathrm{CE}} = \frac{1}{n} \sum_{i=1}^n \nabla_\theta \ell_i = \frac{1}{n} \sum_{i=1}^n J_i^\top r_i.$$

Expanding the squared Euclidean norm,

$$\begin{aligned} \|\nabla_\theta \mathcal{L}_{\mathrm{CE}}\|_2^2 &= \Big\langle \frac{1}{n} \sum_i J_i^\top r_i,\ \frac{1}{n} \sum_j J_j^\top r_j \Big\rangle \\ &= \frac{1}{n^2} \sum_{i,j} r_i^\top J_i J_j^\top r_j = \frac{1}{n^2} \sum_{i,j} r_i^\top \mathcal{K}_{ij} r_j. \end{aligned}$$

□

Proposition 8. *There exist a batch of two samples, a nonzero Jacobian, and nonzero residuals such that $\nabla_\theta \mathcal{L}_{\mathrm{CE}} = 0$.*

Proof. Take $J_1 = J_2 = J \neq 0$ and $r_1 = -r_2 \neq 0$. Lemma 2 gives

$$\nabla_\theta \mathcal{L}_{\mathrm{CE}} = \frac{1}{2}\big(J^\top r_1 + J^\top r_2\big) = 0.$$

The same conclusion holds if a single residual lies in the kernel of $J_i^\top$, since that term $J_i^\top r_i$ is then zero by itself. A nonzero Jacobian therefore supplies a first-order direction, but it does not by itself produce a nonzero cross-entropy step. □

A.6 Proof of the Jacobian-norm bound

Proof of Lemma 3. Lemma 2 and the triangle inequality give

$$\|\nabla_\theta \mathcal{L}_{\mathrm{CE}}\|_2 = \Big\| \frac{1}{n} \sum_i J_i^\top r_i \Big\|_2 \le \frac{1}{n} \sum_i \|J_i^\top r_i\|_2 \le \frac{1}{n} \sum_i \|J_i\|_{\mathrm{op}} \|r_i\|_2.$$

The hypothesis $\|J_i\|_{\mathrm{op}} \le \varepsilon_J$ produces the first claimed bound. For the second, $r_i = W^\top (p_i - y_i)$ implies $\|r_i\|_2 \le \|W\|_{\mathrm{op}} \|p_i - y_i\|_2$. The hypothesis is an upper bound on every operator norm. A small ratio $\rho = \lambda_2/\lambda_1$ does not imply it: ρ compares eigenvalues of $\widehat{\Theta}$ and does not control $\|J_i\|_{\mathrm{op}}$. □

A.7 The CE-aligned tangent energy

Proposition 9. *Let $r = \nabla_u \mathcal{L}_{\mathrm{CE}}$ and let J_u be the Jacobian of the logits with respect to the extractor parameters. With $K_{\mathrm{logit}} = J_u J_u^\top$ and $\epsilon > 0$,*

$$E_{\mathrm{CE}} = \frac{\|\nabla_{\theta_{\mathrm{feat}}} \mathcal{L}_{\mathrm{CE}}\|_2^2}{\|r\|_2^2 + \epsilon} = \frac{r^\top K_{\mathrm{logit}} r}{\|r\|_2^2 + \epsilon}.$$

Proof. The chain rule gives $\nabla_{\theta_{\mathrm{feat}}} \mathcal{L}_{\mathrm{CE}} = J_u^\top r$. Therefore

$$\|\nabla_{\theta_{\mathrm{feat}}} \mathcal{L}_{\mathrm{CE}}\|_2^2 = r^\top J_u J_u^\top r = r^\top K_{\mathrm{logit}} r.$$

Dividing by the same positive denominator $\|r\|_2^2 + \epsilon$ yields the identity. The quantity E_{CE} is the energy of the current logit residual in K_{logit}. It does not replace the spectrum of $\widehat{\Theta}$ after the 50-minibatch probe. □

B Experimental Protocol and Hyperparameters

This section specifies the training, induction, recovery, and measurement protocols used in Sections 3–6. All reported accuracies are test accuracies, and each spectral ratio is $\rho = \lambda_2/\lambda_1$ for the stated estimator.

B.1 Data and architectures

CIFAR-10 has 10 balanced classes, 32×32 color images, 50,000 training images, and 10,000 test images. Training uses random crops with padding 4, random horizontal flips, and per-channel normalization with mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2470, 0.2435, 0.2616)$. Evaluation and every spectral probe use the test images with normalization only. CIFAR-100 uses the same spatial pipeline with its own channel statistics and $K = 100$, so the balanced class prior is 0.01.

Mechanism experiments use BN-ResNet-18 with 20 BN layers and 4,800 scale channels. Transfers use ResNet-34, VGG-11-BN, and ResNet-18 with GroupNorm; no-skip cells retain the residual stages but remove identity branches. GroupNorm has learnable affine scales but no BN-scale vector γ, so γ-specific measurements and interventions apply only to BN models.

B.2 Optimizers and budgets

Unless a cell names another optimizer, training uses SGD with momentum 0.9, learning rate 0.05, and batch size 128. On CIFAR-10 this batch size gives 391 minibatches per epoch. We keep two recovery budgets separate. Block-reset recovery rebuilds the optimizer each round and runs 60 updates; its long budget is 100 rounds (6,000 updates), while the geometric-cliff cells identified in Table 1 stop after 20 rounds. Persistent recovery retains the checkpoint momentum buffer for 100 epochs (39,100 updates). The default weight decay on a persistent cell is 5×10^{-4} on all parameters. The spectral-reopening probe and the strict pure-CE cells set weight decay to zero. A cell is called locked only when low accuracy, the geometric cliff, and the Feature-NTK cut-off persist through its stated budget. BN scales are mechanistic measurements, not a universal locking criterion.

Table 1: Auxiliary coefficients used to construct forward states on BN-ResNet-18. Recovery is block-reset pure CE.

Condition	λ_c	λ_h	λ_{lip}	$\lambda_{\gamma,2}$	$\lambda_{\gamma,1}$	Rounds
Output anchor only	2	0	0	0	0	20
Residual suppression, no scale penalty	2	8	4	0	0	100
Residual suppression plus L2 scale penalty	2	4	2	16	0	20
Scale penalties only	0	0	0	8	8	100

B.3 Auxiliary objective

Checkpoints are built from a CE warmup followed by an auxiliary phase. On the BN-ResNet-18 anatomy runs the warmup ends at test accuracy between 0.589 and 0.616, participation ratio between 2.38 and 2.68, and global γ RMS about 0.875. The auxiliary objective is

$$\mathcal{L}_{\mathrm{trap}} = \mathcal{L}_{\mathrm{CE}} + \lambda_c \|f_\theta(x) - c\|_2^2 + \lambda_h \mathcal{R}_h + \lambda_{\mathrm{lip}} \mathcal{R}_{\mathrm{lip}} + \lambda_{\gamma,2} \|\gamma\|_2^2 + \lambda_{\gamma,1} \|\gamma\|_1.$$

The anchor c is detached from the gradient. $\mathcal{R}_h$ penalizes the squared deviation of the layer2, layer3, and layer4 feature maps from their per-channel batch-and-spatial means. $\mathcal{R}_{\mathrm{lip}}$ penalizes the squared change in features under an input perturbation of standard deviation 0.1. The two scale terms are the mean square and the mean absolute value of the BN affine scales. Table 1 lists the four forward-state conditions of Section 4. The cut-off experiments use the full vector $(\lambda_c, \lambda_h, \lambda_{\mathrm{lip}}, \lambda_{\gamma,2}, \lambda_{\gamma,1}) = (2, 8, 4, 8, 8)$ or the L1-scale vector $(2, 8, 4, 0, 4)$. The auxiliary phase stops at the cut-off gate or at a declared round budget: 60 rounds for the ten-seed BN-ResNet-18 rate, and 30 rounds for the later seeds and the cross-architecture cells. Recovery turns every auxiliary coefficient off.

B.4 Federated induction

The federated grid uses $N = 10$ clients: $m = 1$ optimizes $\mathcal{L}_{\mathrm{trap}}$, the other nine optimize CE, and hence $m/N = 0.1$. Training indices are partitioned by a Dirichlet distribution of concentration 0.5. Each round runs 30 local steps. Federated Averaging, Krum, trimmed mean, and Bulyan are each paired with three auxiliary-client update constructions, producing 12 cells. These cells test induction from a model that is not already cut off, not recovery from a locked checkpoint.

We also report a two-branch FedAvg sweep with one CE branch and one auxiliary branch of aggregation weight $w_{\mathrm{int}} \in \{0.5, 0.8\}$. Each run uses seed 42, 60 local steps per round, 24 auxiliary rounds, and 20 pure-CE recovery rounds. These are branch weights, not equal-weight client fractions, and are therefore not relabeled as m/N.

C Additional Spectral Diagnostics

C.1 Estimator

The scalarized Feature-NTK uses a class-balanced test subset whose indices are fixed before experimentation. For Q unit directions v_q in feature space, $g_{iq} = J_i^\top v_q$ and $\widehat{\Theta}_{ij} = Q^{-1} \sum_q \langle g_{iq}, g_{jq} \rangle$. The Gram is accumulated in float64 because cut-off ratios near 10^{-16} fall below the error of float32 accumulation over roughly 10^7 parameter terms. Its eigenvalues are clipped below at zero. Numerical rank counts eigenvalues of at least $10^{-3}\lambda_1$. If $\lambda_1 = 0$, the rank is recorded as 0 and ρ is left undefined. The cut-off gate used throughout the paper has numerical rank 0, or numerical rank 1 with $\rho \le 10^{-6}$.

The headline probe requests 32 images and retains 30, three per class, with $Q = 1$; rank 30 therefore fills the sample budget but does not imply parameter-space Jacobian rank 30. The dense check uses 60 images and $Q = 8$; cross-architecture logs use 10 images. The projection seed is fixed throughout. Appendix A proves that any fixed projection yields a positive semidefinite $\widehat{\Theta}$, and that an isotropic random direction estimates $d^{-1} \operatorname{tr}(J_i J_j^\top)$.

Table 2: Output-anchor cliffs and block-reset CE recovery on three architectures. The Feature-NTK rank is computed on 10 probe images, so rank 10 fills that budget.

Network	Seed	Cliff r_{PR}	Cliff ρ	Recovery acc.	Recovery rank
ResNet-34	20	1.011	0.0461	0.432	10
ResNet-34	21	1.008	0.308	0.358	10
ResNet-34	22	1.019	0.883	0.344	10
VGG-11-BN	20	1.023	0.440	0.241	8
VGG-11-BN	21	1.022	0.434	0.270	10
VGG-11-BN	22	1.020	0.465	0.257	7
ResNet-18-GN	20	1.014	0.0204	0.305	10
ResNet-18-GN	21	1.001	3.75×10^{-3}	0.430	10
ResNet-18-GN	22	1.001	1.77×10^{-3}	0.453	10

C.2 Headline probe and dense remeasurement

The geometry section quotes the dense spectrum at the cliff, because the headline probe can make a sharp but still multi-directional kernel look like rank one. For the output anchor, the dense measurement gives $\rho = 0.716$, numerical rank 13, and $\lambda_1 = 92.2$. For residual suppression without a scale penalty, it gives $\rho = 0.194$, rank 14, and $\lambda_1 = 284$. For residual suppression plus an L2 scale penalty, the 30-image probe reports rank 1 and $\rho = 6.77 \times 10^{-4}$. The dense remeasurement of the same checkpoint reports rank 2, $\rho = 1.32 \times 10^{-3}$, entropy effective rank 1.03, and $\lambda_1 = 86.9$. Both estimates are above 10^{-6}. The dense ranks after recovery are 57, 60, and 60 respectively, so those kernels are multi-directional inside a 60-image budget rather than saturated at 30.

For a locked persistent run, ResNet-34 seed 24 finishes 39,100 updates at accuracy 0.100, rank 1, $\rho = 1.01 \times 10^{-16}$, and $\lambda_1 = 3.84 \times 10^4$. Thus a cut-off ratio can coexist with a large leading eigenvalue, as Proposition 7 constructs.

C.3 What the 50-step rule uses

Our diagnostic reads ρ at steps 0 and 50 of persistent pure CE with learning rate 0.05, momentum 0.9, zero weight decay, and batch size 128. Equation 2 predicts recovery when $\rho_0 > 10^{-6}$ or $\rho_{50} > 10^{-3}$. On the eight retrospective checkpoints the rule matches the later block-reset label in every case (Table 4). Inside the five severe cut-offs, three minibatch orders produce the same decision on all 15 checkpoint–order pairs. The largest ρ_{50} on the locked side across those orders is 1.32×10^{-16}. The smallest ρ_{50} of seed 9 across the three orders is 0.132; the primary order gives 0.365. At step 50, ρ has area under the receiver operating characteristic curve (AUROC) 1.00 on the eight checkpoints, while rises in aligned energy, gradient norm, and accuracy have AUROCs 0.75, 0.75, and 0.875. Those three quantities also reach AUROC 1.00 by step 200. The same five cut-offs match the rule under five probe optimizers: SGD at learning rate 0.01 or 0.10 with momentum 0.9; SGD at learning rate 0.05 with momentum 0; and Adam and AdamW at learning rate 10^{-3}. These repetitions do not add independent checkpoints: the retrospective set contains eight checkpoints, including five severe cut-offs, and the thresholds are not retuned on prospective cells.

D Complete Intervention and Recovery Results

D.1 Geometric cliffs

Table 2 lists the nine output-anchor cliffs outside BN-ResNet-18. Each cell uses 20 block-reset rounds. Recovery accuracies between 0.241 and 0.453 establish departure from the class prior; the corresponding runs also leave $r_{\mathrm{PR}} \approx 1$. It is not the accuracy of a fully trained model. The sharpest cell is ResNet-18-GN seed 22, with $\rho = 1.77 \times 10^{-3}$, and it still reaches accuracy 0.453 and rank 10.

On BN-ResNet-18 without the identity path, the output anchor ends at $r_{\mathrm{PR}} = 1.003$, accuracy 0.123, rank 2, $\rho = 9.81 \times 10^{-3}$, and global γ RMS 0.883. One hundred block-reset rounds then reach accuracy 0.743, rank 30, and $\rho = 0.270$. The L1-scale joint coefficients on the same no-skip

Table 3: Ten deterministic seeds of the full auxiliary objective on BN-ResNet-18. The auxiliary phase stops at the cut-off gate or at 60 rounds. Best ρ is the smallest spectral ratio recorded during that phase.

Seed	Cut-off round	Best ρ	End accuracy	End γ RMS
0	—	6.98×10^{-4}	0.228	0.024
1	—	5.97×10^{-3}	0.267	0.027
2	22	5.74×10^{-17}	0.100	0.201
3	—	1.49×10^{-2}	0.196	0.028
4	—	1.33×10^{-3}	0.134	0.029
5	—	2.77×10^{-3}	0.259	0.031
6	28	1.12×10^{-16}	0.100	0.087
7	—	1.35×10^{-3}	0.172	0.027
8	—	1.61×10^{-2}	0.336	0.032
9	16	6.24×10^{-17}	0.100	0.422

network end at $r_{\mathrm{PR}} = 1.005$, accuracy 0.270, rank 13, and $\rho = 0.0705$. One hundred rounds reach accuracy 0.779, rank 30, and $\rho = 0.376$. Neither cell meets the cut-off gate.

Standard CE from a random initialization, with every auxiliary coefficient zero, ends 100 rounds at accuracy 0.807, $r_{\mathrm{PR}} = 3.96$, rank 30, $\rho = 0.228$, and global γ RMS 0.287.

D.2 Induction window and long recovery

Table 3 is the ten-seed record summarized by Figure 2. Seed 0 reaches numerical rank 1 with best $\rho = 6.98 \times 10^{-4}$ and is still outside the gate. The three seeds that enter the gate finish at γ RMS 0.422, 0.201, and 0.087. Several seeds that miss the gate finish near RMS 0.025. One full-objective run that continues after the geometry is already sharp ends at rank 5, $\rho = 0.314$, accuracy 0.198, and γ RMS 0.013. Twenty subsequent CE rounds reach accuracy 0.376 and rank 30.

After 100 block-reset rounds, seed 2 remains at accuracy 0.100, $r_{\mathrm{PR}} = 1.000$, rank 1, $\rho = 6.36 \times 10^{-17}$, and γ RMS 0.053. Seed 6 remains at accuracy 0.100, rank 1, and $\rho = 1.02 \times 10^{-16}$. Seed 9 reaches accuracy 0.792, $r_{\mathrm{PR}} = 4.05$, rank 30 on the 30-image probe, and $\rho = 0.267$. The two joint-recipe representatives that were not drawn from this ten-seed list also remain locked: one at $\rho = 2.23 \times 10^{-17}$ and γ RMS 0.012, the other at $\rho = 6.90 \times 10^{-9}$ and γ RMS 0.081. On CIFAR-100 the same kind of joint objective ends the auxiliary phase at accuracy 0.010, $r_{\mathrm{PR}} = 1$, rank 1, $\rho = 9.60 \times 10^{-9}$, and γ RMS 0.0451. One hundred rounds remain at accuracy 0.010, rank 1, and $\rho = 2.50 \times 10^{-16}$. We run four additional recovery seeds for each of the full-objective checkpoint, the L1-scale checkpoint, and this CIFAR-100 checkpoint. All twelve cells stay at the corresponding class prior with rank 1.

D.3 Probe decisions and prospective outcomes

Table 4 is the retrospective application of equation 2. The long-horizon column is 100-round block-reset pure CE, except that the output-anchor and L2-scale cliffs in Section 4 were read after the 20 rounds declared for those conditions. Table 5 is the frozen rule on four cut-off checkpoints that did not enter the choice of the thresholds. Table 6 records the two later protocols. Under block-reset, the predicted recoveries reach accuracy 0.782 and 0.686 with rank 10 on the 10-image probe, and the predicted locks stay at accuracy 0.100 and rank 1. Under persistent pure CE with zero weight decay, both predicted recoveries cross accuracy 0.2 at epoch 3 (1,173 updates), at accuracies 0.577 and 0.480. Both predicted locks complete 100 epochs. A second protocol on the same checkpoint is not a new independent checkpoint. Seed 15, the only cut-off among seeds 10–19, has $\rho_{50} = 0.0281$ and reaches accuracy 0.7565, rank 30, and $\rho = 0.366$ after 100 block-reset rounds. Seed 9 also leaves the class prior by epoch 3 or 4 under four persistent protocols, with final accuracies from 0.368 to 0.652.

Table 4: Controlled 50-minibatch CE probe on eight BN-ResNet-18 checkpoints. The long-horizon column is block-reset pure CE.

Checkpoint	ρ_0	ρ_{50}	Rule	Declared-budget outcome
Output anchor only	0.518	0.547	recover	recovered
Residual suppression plus L2 scale	6.21×10^{-4}	0.329	recover	recovered
Residual suppression, no scale penalty	0.203	0.444	recover	recovered
Full joint objective	1.03×10^{-16}	4.33×10^{-17}	lock	remained locked
L1-scale only	1.29×10^{-16}	6.56×10^{-17}	lock	remained locked
Seed 2	3.22×10^{-17}	4.23×10^{-17}	lock	remained locked
Seed 6	1.19×10^{-16}	9.42×10^{-17}	lock	remained locked
Seed 9	4.10×10^{-17}	0.365	recover	recovered, accuracy 0.792

Table 5: Frozen spectral-reopening rule on four cut-off checkpoints that were not used to choose the thresholds.

Checkpoint	ρ_0	ρ_{50}	Prediction
ResNet-34, seed 22	3.15×10^{-7}	0.0162	recover
ResNet-34, seed 23	1.10×10^{-16}	0	lock
ResNet-34, seed 24	1.13×10^{-17}	1.46×10^{-16}	lock
VGG-11-BN, seed 23	4.98×10^{-7}	0.245	recover

D.4 BN-SCALE EDITS

Table 7 is the matched-RMS comparison in Figure 4. At the edit, the layerwise-constant kernel on each locked checkpoint has numerical rank 1 and ρ between 5×10^{-17} and 4×10^{-16}. The flattened kernel has already left the gate: ρ is 0.164, 0.0103, 0.0202, and 0.0143 on the full objective, the L1-scale objective, seed 2, and seed 6. Accuracy at that moment is still near 0.10. Seed 9 has $\rho = 0.0467$ and 0.0341 under the two edits and then recovers from both.

On the full-objective checkpoint, setting every γ to one common positive value and then running 100 block-reset rounds gives accuracies 0.759, 0.755, 0.804, 0.822, 0.826, and 0.849 at the values 0.05, 0.06, 0.08, 0.10, 0.20, and 0.50. Setting every scale to 1 reaches accuracy 0.848 if the running statistics are reestimated and 0.818 if they are not. Both end at rank 30. Reestimating the running statistics of all 20 BN layers on 50 training batches, with γ left unchanged, ends at accuracy 0.100, rank 1, and $\rho = 5.53 \times 10^{-17}$.

Rescaling a preserved channel pattern to global RMS 0.05, 0.10, or 0.20 leaves the full objective and seed 2 locked at all three targets. Seed 6 remains locked at 0.05 and 0.10. At target 0.20, its ratio is already 1.84×10^{-5} at the edit, and 100 rounds reach accuracy 0.717, rank 30, and $\rho = 0.192$.

A within-layer permutation preserves each layer's multiset, histogram, mean, and RMS. Three permutation seeds leave the full objective and seed 2 at accuracy 0.100 and rank 1. The L1-scale checkpoint reaches 0.755, 0.780, and 0.810. Seed 6 reaches 0.671, 0.763, and 0.723. Seed 9 reaches 0.804, 0.829, and 0.841. Permuting only `bn1`, layer1, layer2, layer3, or layer4 of the full objective leaves all five cells at accuracy 0.100 and rank 1. The layer4 group has the largest stage RMS, about 0.058. After a permutation of every layer, reestimating running statistics on 50 batches still leaves both tested seeds at accuracy 0.100 and rank 1.

Transplanting the entire γ vector, and leaving every other parameter fixed, gives three reversals. Seed 9 onto seed 2 reaches accuracy 0.814; the reverse stays at 0.100. Seed 9 onto seed 6 reaches 0.823; the reverse stays at 0.100. The L1-scale vector onto the full objective reaches 0.778; the reverse stays at 0.100. Each transplant changes the global RMS, and the channels of the two checkpoints are not aligned.

D.5 OTHER RECOVERY PROTOCOLS

With momentum set to 0, learning rate 0.05, and 100 block-reset rounds, the full-objective checkpoint ends at accuracy 0.100, $r_{\mathrm{PR}} = 1$, rank 1, and $\rho = 5.71 \times 10^{-17}$. Reinitializing the linear head and then running the same protocol ends at accuracy 0.100, rank 1, and $\rho = 8.75 \times 10^{-9}$, still

Table 6: Outcomes after the predictions in Table 5. Block-reset is 100 rounds. Persistent pure CE uses zero weight decay. Rank on the block-reset column is computed on 10 probe images.

Checkpoint	Prediction	Block-reset acc.	Persistent acc.	Persistent epoch
ResNet-34, seed 22	recover	0.782	0.577	3
VGG-11-BN, seed 23	recover	0.686	0.480	3
ResNet-34, seed 23	lock	0.100	0.100	100
ResNet-34, seed 24	lock	0.100	0.100	100

Table 7: Two edits of γ at a matched global RMS, followed by 100 rounds of block-reset pure CE. Preserving the layerwise RMS profile keeps each layer's RMS and removes within-layer channel differences. Flattening sets every channel to that checkpoint's original global RMS.

Checkpoint	Global RMS	Layerwise profile	Flattened across layers
Full joint objective	0.0427	locked, 0.100	recovered, 0.782
L1-scale only	0.299	locked, 0.100	recovered, 0.842
Seed 2	0.201	locked, 0.100	recovered, 0.816
Seed 6	0.0868	locked, 0.100	recovered, 0.820
Seed 9	0.422	recovered, 0.839	recovered, 0.840

inside the gate. Block-reset learning rates 0.10, 0.20, and 0.50 keep accuracy at 0.100 and end at recorded participation ratio 0, Feature-NTK rank 0, and global γ RMS 3.15×10^{-3}, 2.38×10^{-4}, and 9.93×10^{-8}. Block-reset Adam at learning rate 10^{-3} reaches accuracy 0.787 on the full objective and 0.825 on the L1-scale checkpoint, both at rank 30. The same optimizer at learning rate 10^{-4} on the full objective stays at accuracy 0.100.

Table 8 lists the 100-epoch persistent cells, all ending at accuracy 0.100. Rank 0 denotes no positive leading eigenvalue; rank 1 denotes a kernel that still satisfies the cut-off gate.

Table 8: Persistent recovery of cut-off checkpoints for 100 epochs. Every cell ends at accuracy 0.100. Ranks are listed in the same order as the checkpoints.

Protocol	Checkpoints	Rank
Pure CE, no weight decay, learning rate 0.05	full, L1-scale	1, 1
SGD, weight decay 5×10^{-4} on all parameters, learning rate 0.05	full, L1-scale	0, 0
SGD, no decay on BN and bias, learning rate 0.05	full, L1-scale	0, 0
Cosine SGD, learning rate 0.10, no decay on BN and bias	full; L1-scale; seed 2; seed 6	0; 1; 0; 0
Adam, learning rate 10^{-3}, weight decay 5×10^{-4}	full, L1-scale	1, 1
AdamW, learning rate 10^{-3}	full, L1-scale	1, 0

D.6 Federated cells

All 12 federated cells miss the cut-off gate. At induction end, Feature-NTK rank is 30 and accuracy spans 0.294–0.679; after CE, rank remains 30, accuracy spans 0.721–0.766, and ρ spans 0.283–0.512. The three FedAvg cells finish at 0.748, 0.743, and 0.755. In the two-branch sweep, $w_{\text{int}} = 0.5$ moves from $(\text{acc.}, r_{\text{PR}}, \text{rank}, \rho) = (0.5227, 1.009, 30, 0.0599)$ after induction to $(0.768, 30, 0.256)$ after recovery, where the latter tuple reports accuracy, rank, and ρ. For $w_{\text{int}} = 0.8$, the corresponding tuples are $(0.3171, 1.002, 11, 0.0734)$ and $(0.6972, 30, 0.290)$. Near-rank-one geometry therefore induces no cut-off here; this grid does not test preservation of an existing lock.