# REPRESENTATIONAL FRUSTRATION: LOCAL INCOMPATIBILITY IN NEURAL NETWORKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Examples that are nearby in a ReLU network's hidden-state space can nevertheless activate different sets of units. We call this local mismatch representational frustration and ask how it is organized across geometric scales and shaped by learning. We measure activation-support dissimilarity among pairs selected at increasing distance thresholds in hidden-state space. Since frustration has no intrinsic quality ordering, we investigate it through a temporary learning-rate increase, motivated by evidence that large steps can reorganize learning dynamics. We then restore the baseline and compare the resulting trajectory with a matched control. In CIFAR-10 MLPs, support incompatibility among control-nearby pairs remains displaced through the observed training horizon, after the learning rate returns to baseline and the large accuracy disruption subsides. Comparing local changes with the all-pair response reveals a reorganization across geometric scales. Under cross-entropy, the closest pairs initially change less than the global pair population but retain excess incompatibility afterward; under mean-squared error, positive local excess is already present at the measured pulse checkpoints. The retained local excess also depends on training conditions: it is negative in the full-batch runs, and different hidden layers can retain opposite signs. These observations show a persistent, scale-dependent organization of representational frustration that predictive accuracy alone does not describe. What these persistent differences retain from learning, and whether they play a role in what the network remembers, remain open questions.

## 1 INTRODUCTION

When two examples are nearby in a learned representation, how similar are the activation patterns used to represent them? In a ReLU layer, each example produces a hidden-state vector and an activation support: the set of units whose outputs are positive. Nearby hidden states need not share the same active units, even though support is determined by the hidden state itself. We call this local mismatch **representational frustration** and ask how learning organizes it across examples, geometric scales, and layers.

Small positive activations can distinguish the supports of nearby states, while states with identical supports can differ substantially in activation magnitude. We examine this organization by measuring support dissimilarity among pairs selected at progressively larger distance thresholds in hidden-state space. Prior work compares representations across networks (Kornblith et al., 2019), studies the encoding properties of activation patterns (He et al., 2021), and examines how paths in Euclidean input space map into activation space (Lewandowski et al., 2025). Studies of representational drift also show that internal states can evolve while task performance remains stable (Pashakhanloo & Koulakov, 2023). We focus on the local relation between hidden-state proximity and activation support, and how its organization changes during learning. A frustration value alone does not establish whether this organization is useful for the task. We therefore investigate its meaning through comparative responses to a controlled change in training.

Learning rate offers a motivated probe: prior work identifies qualitatively different dynamics at large learning rates (Lewkowycz et al., 2020) and shows that large-step SGD can change the features learned (Andriushchenko et al., 2023). We temporarily increase the learning rate and then restore its baseline, asking how local geometry–support organization responds and what remains afterward.



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.

We compare with a matched constant-rate control, evaluating both runs on the same pairs selected using the control representation at each checkpoint. Varying the objective, optimization settings, timing, and depth lets us ask which features of the response recur and which depend on how and where the perturbation is encountered.

In the reference experiment, local support incompatibility remains positively displaced through epoch 200, even after the large accuracy disruption subsides; test loss remains elevated. Relative to the all-pair change, the closest pairs initially show a smaller response under cross-entropy but retain a larger response afterward. Cross-entropy and mean-squared error retain positive fine-scale excess through different measured trajectories. The sign of retained excess also varies across batch settings and can differ between hidden layers of the same network. Together, these findings show that the relation between hidden-state proximity and activation support can retain a structured trace of training history that accuracy alone does not describe. What, beyond the occurrence of a perturbation, is retained in this local organization? Could it preserve distinctions among nearby examples that matter for what the network remembers?

## 2 Measuring Representational Frustration

In ReLU networks, activation support gives a concrete way to investigate whether nearby representations share similar active units. We measure this relation across geometric scales, then compare its response to a temporary change in training.

At a fixed layer, let $h_i^r(t)$ be the post-ReLU hidden state of example $i$ in run $r$ at checkpoint $t$, and let $A_i^r(t) = \{a : h_{ia}^r(t) > 0\}$ be its active support, where $a$ indexes the layer's units. To compare which units are active, we use Jaccard dissimilarity:

$$u_{ij}^r(t) = 1 - \frac{|A_i^r(t) \cap A_j^r(t)|}{|A_i^r(t) \cup A_j^r(t)|}. \tag{1}$$

Here $|A|$ denotes the number of units in $A$. The dissimilarity is the fraction of units active for at least one example that are active for only one of the two; jointly inactive units do not enter the comparison. It ranges from zero for identical supports to one for disjoint supports with a nonempty union. We set $u_{ij}^r(t) = 0$ for two empty supports.[1]

**Why distance and support can disagree.** For $h_i = (1, \epsilon, 0)$ and $h_j = (1, 0, \epsilon)$, the Euclidean distance is $\sqrt{2}\epsilon$, while $u_{ij} = 2/3$ for every $\epsilon > 0$. Thus, nearby states can have substantially different supports even when the differing activations are small. Support dissimilarity describes which units differ; their functional importance also depends on activation magnitudes and downstream computation.

### 2.1 Frustration Across Geometric Scales

Let $\mathcal{P}$ be a fixed sample of unordered, distinct-example pairs, shared across runs and checkpoints. [2]

We rank pairs by Euclidean distance between their raw post-ReLU hidden states. Write $\mathcal{P}_{\le q}^r(t)$ for the pairs at or below the $q$th empirical distance percentile in run $r$ at checkpoint $t$. Here $q$ is expressed in percentage points: $q = 1$ selects approximately the closest $1\%$ of sampled pairs. These sets are cumulative, and all ties at the threshold are included.

The frustration profile is the mean support dissimilarity among the selected pairs:

$$F_{\le q}^r(t) = \frac{1}{|\mathcal{P}_{\le q}^r(t)|} \sum_{(i,j) \in \mathcal{P}_{\le q}^r(t)} u_{ij}^r(t), \qquad 0 < q \le 100. \tag{2}$$

At small $q$, the profile describes the closest sampled pairs; at $q = 100$, it gives the all-pair mean over $\mathcal{P}$. Larger values indicate greater support disagreement in the selected pairs, without implying poorer task performance.

[1] The implementation excludes a pair from both compared runs if either support union is empty. No such exclusions occur in the reported measurements, so the results agree with the definition above (Appendix B).

[2] Example identities are aligned before sampling; sampling details are given in Section 3 and Appendix B.

### 2.2 Comparing Frustration on a Common Geometry

Equation 2 uses each run's own geometry. Comparing runs this way can change both support dissimilarities and the identities of the selected pairs. We therefore select pairs using the control geometry at each checkpoint.

Let $\mathrm{p}$ denote the perturbed run and $\mathrm{c}$ its control. We evaluate both runs on the same control-selected pairs:

$$\widetilde{F}^{r}_{\le q}(t) = \frac{1}{|\mathcal{P}^{\mathrm{c}}_{\le q}(t)|} \sum_{(i,j)\in\mathcal{P}^{\mathrm{c}}_{\le q}(t)} u^{r}_{ij}(t), \qquad r \in \{\mathrm{p}, \mathrm{c}\}. \tag{3}$$

The pool $\mathcal{P}$ is fixed, while its selected subsets are recomputed from the control at each checkpoint. Pair identities therefore match between runs at each time, but need not remain fixed over time or be nearby in the perturbed representation.

### 2.3 Local and Global Changes in Frustration

Write $\Delta\widetilde{F}_{\le q}(t) = \widetilde{F}^{\mathrm{p}}_{\le q}(t) - \widetilde{F}^{\mathrm{c}}_{\le q}(t)$ for the raw pulse–control difference. To compare the local response with the overall change in the pair pool, we define the local excess response:

$$S_{\le q}(t) = \Delta\widetilde{F}_{\le q}(t) - \Delta\widetilde{F}_{\le 100}(t). \tag{4}$$

Positive $S_{\le q}$ indicates a larger signed response among the selected local pairs than overall; negative $S_{\le q}$ indicates a smaller one. A negative excess can therefore coexist with a positive raw pulse–control displacement. By construction, $S_{\le 100}(t) = 0$: at full scale, the local set equals the full pair pool. [3]

## 3 Probing Frustration with Learning-Rate Perturbations

We use a temporary learning-rate increase to investigate how frustration responds to a change in training dynamics and what remains after the baseline rate is restored. Each pulse run is compared with a control that retains the baseline learning rate throughout training.

### 3.1 Reference Experiment

The reference experiment uses a ReLU MLP with one hidden layer of width $512$, trained on a fixed, class-balanced subset of $10{,}000$ CIFAR-10 training images (Krizhevsky et al., 2009). We focus on this setting because it combines near-perfect training accuracy with substantially lower test accuracy (Zhang et al., 2016), making it useful for investigating how a fitted representation retains distinctions among training examples. ReLU MLPs make activation support explicit and allow the geometry–support relation to be examined at each hidden layer under a common measurement procedure. The reference uses cross-entropy on logits and SGD with batch size $128$ and momentum $0.9$, without dropout, normalization layers, or weight decay. A triangular learning-rate pulse begins at epoch $111$, reaches $10^{-2}$ at epoch $120$, and returns to $10^{-3}$ at epoch $130$; both runs continue through epoch $200$.

Pulse and control use the same initialization, training examples, and nominal training settings apart from the learning-rate schedule, but are trained as separate runs. The reference plots show training seed $42$. Conditions are evaluated with seeds $42$ and $43$. [4]

### 3.2 Measurements and Comparisons

We track the raw frustration displacement $\Delta\widetilde{F}_{\le q}$ and local excess response $S_{\le q}$ defined in Section 2, together with train and test accuracy and loss. Frustration is measured on the $10{,}000$ training

[3] The response is not adjusted by a pre-pulse value; pre-pulse discrepancies between independently trained runs are reported in Appendix D.

[4] Appendix A gives the full configurations and schedule; Appendix D reports the pre-pulse discrepancies between separately trained runs.

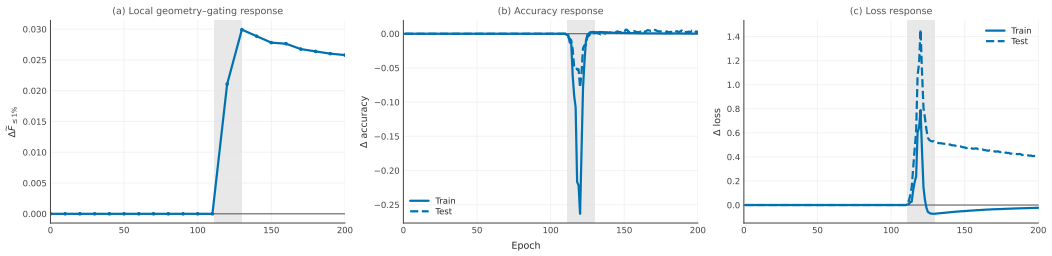


Figure 1: **Local frustration remains changed after the pulse.** The large accuracy gaps subside, while frustration and test loss remain displaced through epoch 200. Panels show pulse-minus-control differences in (a) frustration among the closest $1\%$ of control-selected training pairs, (b) accuracy, and (c) cross-entropy. Reference CE experiment, seed 42, hidden layer 1. Shading marks the pulse; structural measurements are ten epochs apart. Accuracy gaps are plotted as fractions.

examples using a fixed pool of $250{,}000$ unique unordered pairs at epochs $0, 10, \ldots, 200$. At each checkpoint, both runs are evaluated on the same control-selected pairs, as defined in Section 2.

We compare objectives to examine how the response depends on the loss, batch size and momentum to examine dependence on optimization settings, and pulse timing and depth to ask when and where the response is retained. Each condition has a control with the same objective and training configuration, apart from the learning-rate schedule. The baseline and peak learning rates and the epoch-based pulse schedule are shared across conditions, with the schedule shifted for the timing comparisons. This does not match optimizer update counts across batch sizes or time since restoration across the epoch-200 timing comparisons. The comparisons characterize how the response varies across these conditions; they do not isolate a single mechanism behind those differences.

## 4 Persistence and Geometric Organization of Frustration

Can the local relation between examples retain a trace of how a network was trained? The learning-rate pulse makes this question concrete: after the baseline rate is restored, we can ask what remains of the change in frustration. We examine this possibility in the reference cross-entropy experiment, first over time and then across geometric scales.

### 4.1 Frustration After the Perturbation

Frustration, accuracy, and loss respond differently to the pulse (Fig. 1; reference seed 42). Among the closest $1\%$ of control-selected pairs, the raw pulse–control displacement $\Delta\widetilde{F}_{\leq 1\%}$ is $0.02992$ at epoch 130 and $0.02579$ at epoch 200. It is positive at every measured ten-epoch checkpoint in this interval, although the endpoint displacement is smaller. The displacement therefore persists across the observed 70-epoch post-restoration interval.

The accuracy gaps, meanwhile, largely subside by baseline restoration. At the peak-rate checkpoint, epoch 120, the pulse–control gaps are $-26.33$ percentage points in training accuracy and $-7.86$ points in test accuracy. These gaps largely close by baseline restoration at epoch 130 and remain small through epoch 200: their absolute values are at most $0.25$ percentage points for training accuracy and $0.60$ for test accuracy over the recorded epochs in this interval.

Loss also remains different between the runs: at epoch 200, test cross-entropy is higher in the pulse run than in the control ($2.716$ versus $2.312$), while training loss is lower ($0.015$ versus $0.038$).

The runs therefore return to similar accuracies while retaining differences in local frustration and loss.

### 4.2 Frustration Across Geometric Scales

The preceding comparison fixes $q$ at $1\%$. Varying $q$ reveals whether the retained displacement is similar across geometric scales or differs as more distant pairs are included (Fig. 2a,c). Comparing

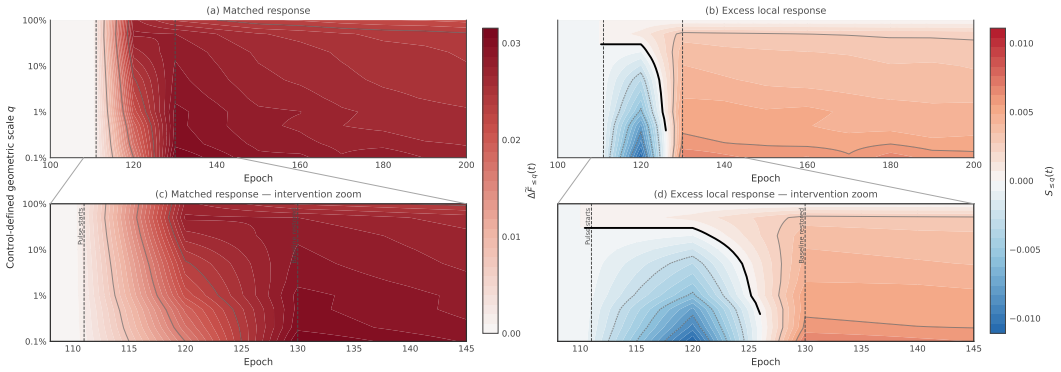


Figure 2: **The frustration response varies across geometric scales.** Local displacement differs from the all-pair response. Left: frustration displacement. Right: its excess over the all-pair response. Both use cumulative control-defined distance selections; the bottom row enlarges the pulse interval. Same run as Fig. 1. Colors interpolate ten-epoch measurements in time and log-percentile; the black contour marks zero excess, and dashed lines mark pulse onset and baseline restoration. Excess at $100\%$ is zero by definition.

each selection with the all pair response shows whether its change is larger or smaller than the overall shift (Fig. 2b,d).

During the pulse, the measured raw response becomes positive across the sampled scales, with smaller displacements at the finest scales than at broader thresholds (Fig. 2a,c). The fine-scale raw displacement remains positive after restoration, as shown by the $1\%$ trace in Fig. 1.

A positive raw displacement alone does not tell us whether nearby pairs have changed more or less than pairs overall. In this reference experiment, the fine-scale displacement is smaller than the all-pair response at the peak-rate checkpoint, but larger by epoch 200, although it is positive at both checkpoints (Fig. 2b,d). At the all-pair endpoint, excess is zero by construction.

These profiles describe cumulative distance selections recomputed at each checkpoint, so the result concerns the organization across scales rather than the persistence of particular pair identities.

The retained displacement is therefore not uniform across geometric scales. A single local value or an all-pair average would leave this organization unresolved.

## 5 Frustration Across Training Conditions

The reference experiment leaves open how the observed organization depends on the conditions of training. We compare objectives, batch sizes, momentum settings, pulse timings, and hidden layers, using a matched control for each training configuration. These comparisons distinguish changes in the measured trajectory, the retained local excess, and its organization across layers.[5]

### 5.1 Objectives organize frustration differently across scale

Cross-entropy (CE) and mean-squared error (MSE) reach positive retained local excess through different changes in accuracy and in the organization of frustration across scales (Fig. 3). Under MSE, the maximum test-accuracy deficit during the pulse is smaller: $6.58$ and $5.79$ percentage points for seeds 42 and 43, compared with $7.86$ and $6.85$ for CE. By epoch 200, MSE also has a positive test-accuracy difference from its control in both seeds, while CE remains closer to its control. At epoch 120, fine-scale excess is negative under CE and positive under MSE in both seeds: local displacement is below the all-pair response for CE and above it for MSE. The learning-rate schedules match, but the different objectives do not imply matched gradient or update magnitudes.

[5]Responses are shown without pre-pulse adjustment; Appendix D quantifies differences present before the intervention.

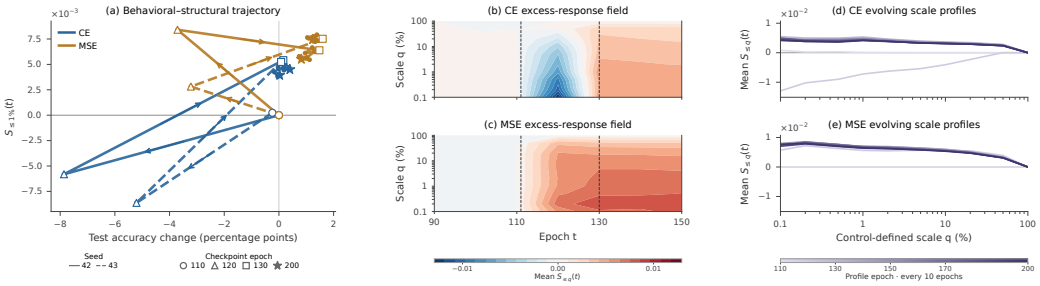


Figure 3: **CE and MSE reorganize frustration differently across geometric scales.** (a) Individual-seed trajectories of test-accuracy difference and local excess at $q = 1\%$. (b,c) Excess across time and geometric scale; (d,e) profiles from epochs 110–200, darker for later checkpoints. Panels (b–e) average seeds 42 and 43. Structural measurements are ten epochs apart; smooth fields interpolate between measurements. Excess at $q = 100\%$ is zero by definition.

The profiles across geometric scales show that this contrast extends beyond the $1\%$ selection (Fig. 3b,c). Under CE, fine-scale displacement is initially smaller than the all-pair response before exceeding it later in training. MSE develops positive fine-scale excess without this measured phase of relative suppression. The contrast suggests different forms of reorganization across scales, even when the retained excess has the same sign. The ten-epoch structural sampling does not resolve whether shorter intervening phases occur.

Both objectives retain positive fine-scale excess at epoch 200 (Fig. 3d,e). At $q = 1\%$, CE retains $0.00450$ and $0.00394$ for seeds 42 and 43, while MSE retains $0.00552$ and $0.00745$. The smaller accuracy disruption under MSE therefore does not preclude a retained local excess. The contrast suggests that the severity of the accuracy disruption and the retained organization of frustration are distinct aspects of the response.

### 5.2 Optimizer Settings Change the Trajectory and Retained Sign

Batch size changes the retained sign of local excess, while the tested momentum settings change the transient response and retain positive excess at epoch 200 (Fig. 4a,b,e,f). At batch size 128, fine-scale excess is negative during the pulse and positive at epoch 200. At batch size 1024, fine-scale excess is positive during the measured pulse response and remains positive afterward. Full-batch runs retain negative local excess, with the epoch-200 profile below zero across the sampled local scales. The retained sign contrast holds separately in both observed seeds.

The fixed epoch-based schedule gives 79, 10, and 1 optimizer updates per epoch at batch sizes 128, 1024, and 10000, respectively. This comparison changes both minibatch sampling and update count; it does not isolate stochasticity.

Across the tested momentum settings, the transient response differs while the retained fine-scale excess stays positive at epoch 200 (Fig. 4b,f). With momentum 0 and 0.5, fine-scale excess is positive during the pulse and remains positive afterward. With momentum 0.9, fine-scale excess is negative during the pulse before becoming positive. Positive retained excess thus appears both without momentum and without the negative transient seen in the reference setting. The ordering between momentum 0 and 0.5 also differs across seeds: at $q = 1\%$, their retained excesses are $0.00597$ versus $0.00582$ for seed 42, and $0.00648$ versus $0.00701$ for seed 43.

### 5.3 Retention Varies with Pulse Timing and Hidden Layer

Retention differs across pulse timings and hidden layers: the early-pulse mean is larger across much of the measured scale range at epoch 200, while the two-hidden-layer network retains opposite signs in its two layers (Fig. 4c,d,g,h).

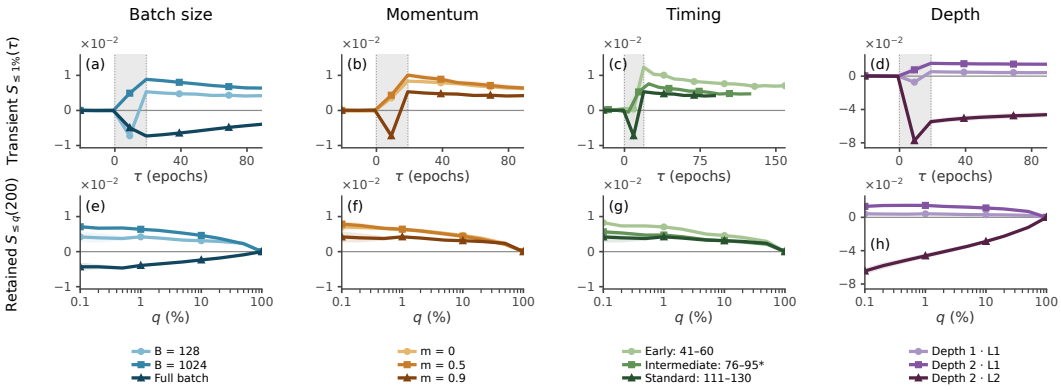


Figure 4: **Retained local excess depends on training conditions and hidden layer.** Its sign differs between minibatch and full-batch runs, and between layers of the two-hidden-layer network. Top: excess at $q = 1\%$ against time since pulse onset. Bottom: epoch-200 scale profiles. Columns vary batch size, momentum, timing, and depth. Lines average seeds 42 and 43; shading spans their values, not confidence intervals. Depth panels use a separate vertical scale.

**Timing.** Aligning the schedules by pulse onset compares the response when the perturbation is introduced at different stages of training (Fig. 4c). The standard pulse shows the reference pattern of negative transient excess followed by positive retained excess. The early-pulse transient differs across seeds: at epoch 50, its fine-scale excess is $-0.00377$ for seed 42 and $+0.00456$ for seed 43, becoming positive in both by epoch 60. For intermediate timing, seed 43 has negative fine-scale excess at epoch 80 and positive excess at epoch 90. Earlier pulses therefore do not consistently remove the negative transient in the measured runs.

At epoch 200, all three schedules show positive local excess in the available runs, with the early-pulse mean largest across much of the sampled scale range (Fig. 4g). The epoch-200 profiles compare equal training ages but unequal recovery durations: 140, 105, and 70 epochs after baseline restoration for early, intermediate, and standard timing. The endpoint comparison therefore combines pulse timing with time since restoration; it does not establish a monotone effect of timing alone.

**Depth.** In the two-hidden-layer network, the layers retain opposite signs of local excess (Fig. 4d,h). Its first hidden layer retains positive local excess, with a larger fine-scale endpoint response than the first layer of the one-hidden-layer reference. The second hidden layer has negative local excess during the pulse and retains that sign through epoch 200. Both seeds show this retained sign contrast: at $q = 1\%$, depth-two L1 retains $0.01555$ and $0.01303$, while L2 retains $-0.04919$ and $-0.04327$, for seeds 42 and 43, respectively. These layer-specific excess responses use each layer's own control geometry and all-pair reference. A single model-level sign would therefore miss the difference between the layers' local responses relative to their respective all-pair shifts.

Across these comparisons, the trajectory, retained sign, and layer profile of local excess depend on the training configuration. Together, they suggest that the retained response reflects how the local relation is reorganized during training, rather than a uniform degree of representational disruption.

## 6 Discussion

### 6.1 Frustration Retains a Trace of Training History

The relation between hidden-state proximity and activation support can retain a trace of training history. In the reference experiment, local frustration remains displaced after the learning rate returns to baseline and the large accuracy disruption subsides; loss also remains different. Similar accuracies can therefore coexist with different local organizations of activation support. We interpret this per-

sistence as a trace of how training organizes what nearby examples share and where their activation supports remain distinct.

This perspective places frustration alongside work showing that performance and internal organization can evolve on different courses. Representational drift provides one example (Pashakhanloo & Koulakov, 2023); the continued evolution of separable linear classifiers after classification error vanishes provides another (Soudry et al., 2018). Frustration makes this inquiry local: it follows how examples that are nearby in hidden-state space differ in the units they activate, and how that relation changes across geometric scales.

The timing comparisons suggest that the retained relation may also depend on when training is perturbed. Critical-learning-period studies establish a related form of historical dependence under temporary changes in training experience (Achille et al., 2017). In our case, earlier pulses leave more training time before the epoch-200 comparison, so the timing results combine the effects of training stage and time since learning-rate restoration.

### 6.2 Transient Responses and Retained Reorganization

The coexistence of a transient accuracy disruption and a retained change in local organization suggests that the immediate response and the subsequent representational adjustment need not have the same course. Under CE, fine-scale displacement is initially smaller than the all-pair response and larger later; MSE develops positive local excess without the measured negative phase. MSE also shows a smaller accuracy deficit and a positive late test-accuracy difference, whereas CE ends closer to its control. We read these contrasts as different courses of representational adjustment: the larger transient accuracy disruption under CE does not imply that its retained organization is simply a more pronounced version of the MSE response. Retained local excess therefore has no simple ordering as a measure of behavioral improvement or deterioration. Opposite retained excess signs in two layers further suggest that a network need not undergo a uniform local reorganization. The resulting picture is of a response whose form depends on both geometric scale and location within the network, rather than one that can be summarized by the magnitude of the initial disturbance.

Large-step optimization offers one route by which this organization could change. Large-step SGD can favor sparse features in studied settings (Andriushchenko et al., 2023), while analyses of simplified two-layer networks connect large learning rates to the learning of threshold-like neurons (Ahn et al., 2023). These results motivate examining which features become active and how that activity is shared across examples. Frustration adds a relational view: how the supports of nearby examples differ, and how that relation changes across geometric scales. Support occupancy, activations crossing zero, and changes in the control-defined selection of nearby pairs could each contribute to the measured response. The all-pair comparison does not separate these contributions or establish which learning mechanism operates here.

### 6.3 Local Incompatibility Within a Distributed Representation

We view local support incompatibility as a possible expression of how a shared representation retains distinctions among examples. A group of examples can share active units while remaining differentiated in other parts of their supports. Geometric proximity and support sharing therefore capture different aspects of this organization. Frustration makes their relation accessible: it describes how much examples differ in active support at the geometric scales being examined.

Work on activation codes shows that patterns of active units can distinguish instances while also supporting categorization (He et al., 2021). Toy models of superposition offer a complementary perspective on how distinct features can share representational dimensions (Elhage et al., 2022). These studies motivate examining sharing and differentiation together. Frustration addresses a different relation: how support differences are organized among examples selected by their proximity in hidden-state space. A related distinction also arises in mathematical neuroscience, where analyses of receptive-field codes examine how distances between activity patterns reflect relationships among stimuli (Curto et al., 2013).

Learning can bring examples together without erasing every distinction between them. We propose that the organization of these retained distinctions is worth studying in its own right; frustration provides a local, scale-dependent view of it.

### 6.4 Toward a Structural Account of Memory

We hypothesize that some of the support differences among nearby examples participate in retaining task-relevant distinctions. The central possibility is that what a network remembers is expressed partly in how it keeps examples distinct within the structure they share. The label-corruption contrasts, typical of characterizing memorization regimes Zhang et al. (2016), offer a complementary motivation (Appendix E). Work on long-tailed data gives a reason that maintaining such distinctions could matter: retaining information about rare examples can support generalization (Feldman, 2020; Feldman & Zhang, 2020). The present experiments identify a retained relation; they do not yet identify what information its contributing support differences preserve or whether the network uses it.

Testing this account requires connecting the aggregate profile to particular examples and then to functional changes in their predictions. A first step is to identify pairs and examples contributing to the retained displacement and compare them with behavioral measures of retention. Example-level memorization estimates and forgetting histories provide such comparisons (Feldman & Zhang, 2020; Toneva et al., 2018). Neuron-level interventions offer a way to test whether the implicated support differences are functionally used (Maini et al., 2023). Selective prediction changes under matched interventions would support a functional role for the implicated units; attributing that role specifically to their support differences would require controlling the accompanying changes in activation magnitude. If the profile can be changed substantially without selective behavioral consequences, it may instead describe aspects of training history that the present task does not use.

Representational frustration offers an empirical starting point for a broader question: how learning builds shared representations while preserving distinctions that may matter for what a network remembers.

## 7 Conclusion

We introduced representational frustration to examine how geometrically nearby examples differ in activation support. In the reference CIFAR-10 ReLU MLP experiment, a brief learning-rate pulse leaves this relation displaced through epoch 200, after the baseline learning rate is restored and the large accuracy disruption subsides; test loss remains different. Across conditions, the response varies with geometric scale, objective, optimization settings, timing, and layer. Together, these findings reveal a structured trace of training history in the relation between representational proximity and support sharing.

**Scope and limitations.** The evidence is limited to these models and the observed horizon. Most conditions have two seeds, and independently trained pulse/control runs can differ before the pulse, limiting attribution of small effects (Appendix D). Epoch-matched comparisons do not isolate batch noise or training stage because update counts and post-pulse durations differ. The geometric pair sets overlap and evolve; the profiles neither track fixed pairs nor establish transport across scales or layers. Ten-epoch sampling leaves faster transients unresolved. Extension beyond ReLU requires an appropriate definition of computational support. Strict support counts arbitrarily small positive activations, and the experiments do not establish that retained support differences carry task information.

Our working hypothesis is that some of these local distinctions participate in what a network retains. The broader opportunity is to understand how learning builds shared representations while keeping examples distinguishable where it matters.

### AI use statement

The research idea, hypotheses, and experimental program originated with the authors. We used generative AI tools in author-directed discussions to obtain methodological feedback, clarify the presentation of our conceptual framing and hypotheses, examine mathematical definitions and derivations, and explore interpretations and alternative explanations of results. The authors guided these discussions and selected, revised, or rejected the suggestions. We also used these tools for coding assistance, literature search and summarization, manuscript drafting and editing, assembling and

organizing the appendices from separate pieces of author-provided text, experimental details, and results, and plotting figures.

### Reproducibility Statement

We describe the experimental and measurement procedures in the main text and appendices to support reproduction of the reported results. Section 2 defines the observables, and Section 3 describes the experimental design. Appendix A specifies the data preparation, model and training configurations, and intervention schedules. Appendix B details the measurement and sampling procedures, while Appendix C describes temporal resolution, aggregation across seeds, and figure conventions. Appendix D reports pre-intervention discrepancies relevant to interpreting the comparisons. Appendix E defines the cycle-based measure used in the supplementary label-corruption experiments.

## References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep neural networks. *arXiv preprint arXiv:1711.08856*, 2017.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with large step sizes learns sparse features. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 903–925. PMLR, 2023. URL https://proceedings.mlr.press/v202/andriushchenko23b.html.

Carina Curto, Vladimir Itskov, Katherine Morrison, Zachary Roth, and Judy L Walker. Combinatorial neural codes from a mathematical coding theory perspective. *Neural computation*, 25(7): 1891–1925, 2013.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, pp. 954–959, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in neural information processing systems*, 33:2881–2891, 2020.

Fengxiang He, Shiye Lei, Jianmin Ji, and Dacheng Tao. Neural networks behave as hash encoders: An empirical study. *arXiv preprint arXiv:2101.05490*, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Michal Lewandowski, Hamid Eghbalzadeh, Bernhard Heinzl, Raphael Pisoni, and Bernhard A Moser. On space folds of relu neural networks. *arXiv preprint arXiv:2502.09954*, 2025.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Pratyush Maini, Michael C Mozer, Hanie Sedghi, Zachary C Lipton, J Zico Kolter, and Chiyuan Zhang. Can neural network memorization be localized? *arXiv preprint arXiv:2307.09542*, 2023.

Farhad Pashakhanloo and Alexei Koulakov. Stochastic gradient descent-induced drift of representation in a two-layer neural network. In *International conference on machine learning*, pp. 27401–27419. PMLR, 2023.

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.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

# A Experimental Design and Training Details

## A.1 Data, Architecture, and Objectives

All learning-rate perturbation comparisons use a fixed, class-balanced subset of $10{,}000$ CIFAR-10 training images, with $1{,}000$ examples per class, selected using data seed 42. The same subset supplies the examples used to measure frustration; test accuracy is evaluated on the standard test split. Labels are uncorrupted. Images are converted to tensors and normalized channelwise with means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2470, 0.2435, 0.2616)$, without random crops or flips.

Each image is flattened to $3{,}072$ coordinates. The MLP has width $512$ in each hidden layer, ReLU activations, enabled biases, and ten output logits. The reference network has one hidden layer; the depth comparison adds a second hidden layer of the same width. There are no dropout or normalization layers.

Cross-entropy is applied to logits. For mean-squared error, the objective is

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{2N} \sum_{i=1}^{N} \| z_i - e_{y_i} \|_2^2, \tag{5}$$

where $z_i$ is the vector of logits, $e_{y_i}$ is the one-hot target, and $N$ is the minibatch size. This objective sums over output coordinates and averages over examples, without a softmax transformation. Consequently, using the same learning-rate schedule for CE and MSE does not equate their gradient or update magnitudes.

## A.2 Learning-Rate Pulse and Experimental Variations

All runs in these comparisons train for 200 epochs with SGD, zero weight decay, and baseline learning rate $\eta_0 = 0.001$, without multiplicative decay. The reference uses CE, batch size 128, momentum $0.9$, and a pulse starting at epoch 111. Each comparison changes one factor from this reference; Table 1 summarizes the tested settings. Training seeds are 42 and 43.

Table 1: Settings varied from the reference experiment. These are one-factor comparisons, not a full factorial grid. Each pulse run has a control with the same objective, architecture, optimizer settings, and seed.

| Factor | Tested settings |
|---|---|
| Objective | CE, MSE |
| Batch size | 128, 1,024, 10,000 |
| Momentum | 0, 0.5, 0.9 |
| Hidden layers | One or two, each of width 512 |
| Pulse start epoch | 41, 76, 111 |

The learning rate is set before training the indicated epoch. If $s$ is the first pulse epoch, the schedule is

$$\eta(t) = \begin{cases} 0.001 + 0.0009(t - s + 1), & s \le t \le s + 9, \\ 0.001 + 0.0009(s + 19 - t), & s + 10 \le t \le s + 19, \\ 0.001, & \text{otherwise.} \end{cases} \tag{6}$$

The first pulse epoch uses $0.0019$, the peak at $s+9$ uses $0.01$, and epoch $s+19$ restores the baseline. The early, intermediate, and standard schedules have $s = 41$, 76, and 111, respectively. Controls remain at $0.001$ throughout.

The final partial minibatch is retained. Batch sizes 128, $1{,}024$, and $10{,}000$ therefore give 79, 10, and 1 optimizer updates per epoch, respectively; the last setting uses the entire training subset in each update. Matching the epoch-based pulse schedule does not match update counts, so the batch comparison does not isolate stochastic noise alone.

# B Properties and Implementation of the Frustration Measurement

## B.1 Fixed Pair Pool and Evolving Geometric Selections

The pair pool $\mathcal{P}$ contains $250{,}000$ unique unordered pairs sampled without replacement from the $\binom{10{,}000}{2}$ distinct-example pairs. Sampling uses NumPy's default random generator with seed 42. Aligned example identities and the fixed sampling procedure give the same pair pool across checkpoints, conditions, training seeds, and layers.

At each checkpoint, control distances define thresholds at

$$q \in \{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}\%.$$

Thresholds use NumPy's default linear percentile interpolation, and all pairs at or below the threshold are included, including ties. These selections are cumulative subsets of the sampled pool, rather than disjoint distance shells or a fixed number of neighbors for each example. Examples may contribute to different numbers of selected pairs.

The control-selected set is used for both arms at each checkpoint, but is recomputed as the control representation evolves. Thus, the comparison holds pair identities fixed between arms at a given time, not across time. Selected pairs need not also be nearby in the pulse representation. Persistence of an aggregate response consequently describes a retained geometry–support relation, without identifying persistent individual pair contributions. The all-pair reference is the mean over $\mathcal{P}$, not an enumeration of every pair in the dataset.

## B.2 Interpreting the Local Excess Response

The all-pair reference includes the locally selected pairs. To make the consequence explicit, fix a checkpoint and layer, and write $L = \mathcal{P}^{\mathrm{c}}_{\le q}(t)$ and $\alpha = |L|/|\mathcal{P}|$. Let $\Delta_L$ be the mean pairwise pulse–control dissimilarity difference over $L$, and $\Delta_O$ its mean over $\mathcal{P} \setminus L$. When both sets are nonempty, the all-pair response and local excess satisfy

$$\begin{aligned} \Delta_G &= \alpha\Delta_L + (1-\alpha)\Delta_O, \\ S_{\le q}(t) &= \Delta_L - \Delta_G = (1-\alpha)(\Delta_L - \Delta_O). \end{aligned} \tag{7}$$

Here $\Delta_G = \Delta\widetilde{F}_{\le 100\%}(t)$. The selected fraction $\alpha$ need not equal $q/100$ exactly because of finite sampling and ties.

Local excess therefore compares the local response with the complementary response, multiplied by the fraction of pairs outside the selection. At full selection the complement is empty, and the original definition gives $S_{\le 100\%}(t) = 0$ directly. Since pairwise dissimilarity differences lie in $[-1, 1]$, the identity also gives $|S_{\le q}(t)| \le 2(1-\alpha)$. Attenuation near the all-pair endpoint is consequently partly built into the observable; the endpoint is not evidence of recovery. The identity does not require a monotone profile, since the conditional means change as the selection grows.

The sign of excess is distinct from the sign of the raw local response. For example, a local displacement of $+0.02$ and an all-pair displacement of $+0.03$ give an excess of $-0.01$: local dissimilarity has increased, but less than the all-pair mean. These values are illustrative. Neither response is adjusted by subtracting a pre-pulse value, and the all-pair subtraction does not by itself control for changes in support occupancy.

## B.3 Support, Geometry, and Numerical Conventions

Support dissimilarity counts differing active coordinates without weighting their magnitudes. For nonnegative hidden states $h$ and $g$, let $D$ be the symmetric difference of their supports. On each coordinate in $D$, exactly one activation is positive, and

$$\sum_{a \in D} (h_a - g_a)^2 \le \|h - g\|_2^2.$$

Thus, for nearby states, the differing active coordinates can be small even when many support bits differ. Their effect on predictions also depends on downstream weights and computation. The binary measurement alone does not determine their functional importance.

In exact arithmetic, consistently permuting unit coordinates or uniformly rescaling all hidden vectors in a layer by a positive constant leaves the support dissimilarities and percentile selections unchanged. Positive rescaling of individual units preserves support but can change Euclidean pair rankings. In the control-referenced comparison, such rescaling of the pulse states alone leaves the measurement unchanged because both its supports and the control-selected pairs remain fixed; rescaling the control coordinates can change the selection. These are properties of the measurement, not empirical robustness tests.

Geometry uses float32 post-ReLU activations and Euclidean distance, without unit-norm or width normalization. Support masks are formed by strict positivity before conversion to float32 and retained separately. There is no positive activation tolerance or magnitude weighting.

For two empty supports, the mathematical definition assigns zero dissimilarity. The implementation instead excludes a pair from both arms if either arm's support union is empty, using a shared validity mask. No pairs were excluded at any measured checkpoint, so the implementation and definition agree for these comparisons.

## C Reading the Response Profiles

**Temporal resolution.** Structural measurements use checkpoints $0, 10, \ldots, 200$. Behavioral differences are evaluated at matching recorded epochs; test-accuracy differences are reported in percentage points.

Smooth fields display the measured checkpoint–percentile grid. The reference surfaces interpolate linearly in epoch and in $\log q$; the objective-comparison fields use filled contours of the measured grid. Interpolated transitions do not locate events more precisely than the checkpoint spacing. Displayed interior zero fronts exclude the forced $q = 100\%$ boundary and wholly zero profiles.

**Seed aggregation.** The reference figures show the seed-42 CE comparison. In the objective comparison, behavioral–structural trajectories retain individual seeds, while fields and profiles average the two seed-specific excess responses at corresponding coordinates. The batch, momentum, timing, and depth comparisons likewise average the available seed-specific responses. Activations and distance distributions are not pooled across seeds: each comparison uses its own control geometry before aggregation.

Where shown, envelopes span the pointwise minimum and maximum of the two seed-specific responses. They describe the observed runs and are not confidence intervals. Sampling many pairs does not supply additional independent training replicates, and pairs can share examples.

**Retained profiles and pulse-relative time.** Pulse-relative time is $\tau = t - s$, where $s$ is the pulse start epoch. Retained profiles are evaluated at epoch 200, rather than averaged over a late-training window. At this endpoint, the early, intermediate, and standard pulses have been back at baseline for 140, 105, and 70 epochs, respectively. These profiles match training age but not elapsed time since restoration, so the timing comparison does not isolate training stage from recovery duration.

## D Matching and Pre-Pulse Differences

Pulse and control runs share their initial parameters and training examples, as checked through saved initial-parameter hashes and training-index lists. They are trained separately, rather than branched from a common pre-pulse checkpoint. Matching initialization and data therefore does not establish identical subsequent trajectories. Timing comparisons reuse the corresponding seed's reference control, and the two hidden layers of a depth-two network are measurements from the same run, not independent replications.

Some matched trajectories have nonzero measured differences before the scheduled pulse. The source of these pre-pulse differences remains unresolved. Table 2 reports the descriptive maximum

$$M_S = \max_{\substack{t<s,\ q \\ \text{measured grid}}} |S_{\le q}(t)|, \tag{8}$$

Table 2: Maximum absolute pre-pulse excess response, reported as $10^3 M_S$. Values are rounded to three decimal places; a displayed zero need not imply exact equality. A dash denotes a condition not evaluated with that seed.

| Condition / layer | Seed 42 | Seed 43 |
|---|---|---|
| Reference CE / L1 | 0.000 | 1.593 |
| MSE / L1 | 0.000 | 0.000 |
| Batch 1024 / L1 | 0.200 | 0.197 |
| Full batch / L1 | 0.033 | 0.013 |
| Momentum 0 / L1 | 1.640 | 0.460 |
| Momentum 0.5 / L1 | 0.000 | 0.000 |
| Depth 2 / L1 | 3.407 | 2.423 |
| Depth 2 / L2 | 2.340 | 1.496 |
| Early pulse / L1 | 1.382 | 1.032 |
| Intermediate pulse / L1 | — | 1.416 |

computed over the measured pre-pulse checkpoints and percentile grid for each condition, seed, and layer.

These maxima describe pre-pulse excess across the measured grid. They are not uncertainty estimates for a particular post-pulse value. No pre-pulse adjustment is applied, and small retained differences should be interpreted in light of the discrepancies in Table 2.

## E Motivating Observations with a Cycle-Based Measure

An initial exploration of representational frustration asked how local support compatibility changes when the distinctions a network is asked to fit are altered. We progressively corrupted Fashion-MNIST and CIFAR-10 training labels, replacing each selected label uniformly with one of the other classes, while leaving test labels unchanged. The one-hidden-layer ReLU MLPs considered here have widths $512$ and $8192$ and are evaluated after $200$ training epochs.

This exploration used a cycle-based summary of support compatibility. On at most $1024$ training examples, we constructed a mutual $10$-nearest-neighbor graph using Euclidean distances between hidden states. Each edge was assigned the Jaccard similarity of its activation supports, and each length-four cycle was assigned the geometric mean of its four edge similarities. Writing $\mathcal{C}_4$ for these cycles and $\mathcal{V}_4$ for those whose edges all have nonempty support unions, the summary is

$$F_{\mathrm{cyc}} = 1 - \frac{1}{|\mathcal{C}_4|} \sum_{C \in \mathcal{V}_4} \left( \prod_{(i,j) \in C} \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \right)^{1/4} . \tag{9}$$

All cycles were admissible in the measurements shown. The implementation stabilizes edge similarities at $10^{-12}$ when taking logarithms. This construction summarizes pairwise support overlap along geometric cycles; it does not measure a separate higher-order compatibility constraint.

Figure 5 shows an overall increase in cycle-based frustration with label corruption in all four settings, alongside contrasting fitting behavior. On Fashion-MNIST, the wider network fits the altered labels substantially more closely: with all training labels replaced, mean training accuracy is approximately $92.0\%$ at width $8192$, compared with $52.1\%$ at width $512$. Both CIFAR-10 networks fit the altered labels closely, reaching approximately $100.0\%$ and $97.1\%$, respectively, at full corruption. Clean test accuracy declines across all four settings. The structural response therefore accompanies substantially different degrees of success in fitting the altered supervision.

These contrasts motivated the possibility that support differentiation participates in accommodating distinctions among examples. They also leave a more specific question open: which aspects of the response reflect shared changes in support organization, and which depend on the learned solution? The cycle aggregate neither identifies the support differences that are functionally used nor resolves how incompatibility is distributed across geometric scales. This unresolved organization motivates

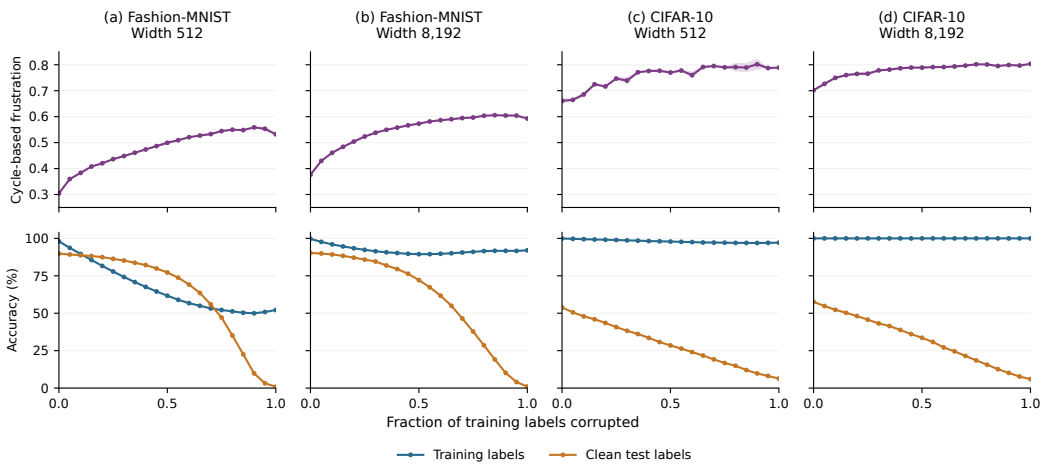


Figure 5: **Label corruption increases cycle-based frustration across contrasting fitting regimes.** Fashion-MNIST and CIFAR-10 ReLU MLPs with one hidden layer, measured at epoch 200. Top: cycle-based frustration. Bottom: accuracy on assigned training labels and clean test labels. Points show means over the same two training seeds in every setting; shading spans their values. Axes are shared within each row.

the distance-conditioned pairwise profile studied in the main text, which makes geometric scale an explicit part of the inquiry.