ON GATING DYNAMICS: SPECTRAL CONDENSATION AND DISTINCT PATHS OF RETURN

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the possibility that the Edge of Stability (EoS) accompanies a reorganization of neuron participation, we examine how ReLU gating evolves across training regimes. Binary neuron–sample supports let us distinguish how much activity is present, how participation is distributed, and how gates change over time. Across ReLU MLPs on FashionMNIST and CIFAR-10, support density initially rises while raw-support effective rank falls—a spectral condensation. Subsequent rank recovery and density decrease follow distinct paths of return. In shallow FashionMNIST, the higher-rate run recovers rank and renews gate switching near EoS, whereas the lower-rate run remains nearly frozen with little rank recovery. Renewed switching becomes largely balanced exchange, with the net on–off current exhibiting approximate period-two recurrence. Yet deep-layer density–rank return precedes the late increase in gate switching, and returns occur both with a larger low-participation population and, in CIFAR-10, with little such population and weaker class-associated sharing. The SGD examples also exhibit condensation and return, but with gate switching present from the outset. Together, these contrasts motivate a hypothesis that participation structure shapes the available paths of return from spectral condensation, while gate switching influences how those paths are traversed.

1 INTRODUCTION

At the Edge of Stability (EoS), gradient descent can keep learning while the largest Hessian eigenvalue hovers near or slightly above $2/\eta$ and the loss oscillates (Cohen et al., 2021). Work on self-stabilization and threshold-neuron learning connects this regime to both sustained progress and what a network learns (Damian et al., 2022; Ahn et al., 2023). These findings motivate a closer look at participation: does approaching the stability boundary accompany a reorganization of which neurons participate on which examples? ReLU gates make this participation explicit: each neuron–example pair is active when its preactivation is positive.

Learning rate also connects this question to sparsity: large learning rates have been associated with selective learning and sparse representations (Li et al., 2019; Andriushchenko et al., 2023; Barsbey et al., 2025). Average sparsity, however, leaves both arrangement and persistence unresolved: the same density can accompany different supports or conceal continual gate dynamics. We therefore follow the binary neuron–sample support through its density and singular spectrum, gate transitions, and individual participation histories. This builds on studies of activation patterns in expressivity and linear regions (Montúfar et al., 2014; Raghu et al., 2017; Serra et al., 2018; Hanin & Rolnick, 2019), input-specific subnetworks (Lakshminarayanan & Vikram Singh, 2020), optimization over activation patterns (Matena & Raffel, 2022), and training dynamics (Hartmann et al., 2021; Wang & Ma, 2023).

Across the studied ReLU MLPs on FashionMNIST and CIFAR-10, support density initially rises while raw-support effective rank falls: more neuron–example pairs participate as the spectrum becomes more concentrated. We call this spectral condensation and use return to describe subsequent rank recovery and density decrease. The returns vary in extent and in the relative timing of rank recovery and density decrease, with shallow FashionMNIST showing substantial recovery or nearly stalled rank depending on learning rate. These contrasts raise a common question: what allows participation to reorganize, and why does its course differ across regimes?

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.

In the shallow comparison, the higher learning rate produces renewed gate switching near measured EoS contact and substantial rank recovery, whereas the lower rate leaves both small. Renewed gate switching consists largely of balanced exchange, with the net on–off current developing approximate period-two recurrence. In deep FashionMNIST, however, density and rank turn well before the later rise in turnover near the stability boundary. Return therefore precedes the late rise in gate switching in this case, leaving open the role of less frequent switching earlier in training.

Participation profiles suggest why gate switching may play different roles: the shallow runs have similar profiles despite different recovery, deep FashionMNIST has more low-participation neurons and observed re-entry, and CIFAR-10 returns with few such neurons and weaker class-associated sharing. These contrasts motivate an abductive hypothesis: participation structure shapes the available paths of return from spectral condensation, while gate switching influences how those paths are traversed. We use structural slack to describe the room a participation pattern may leave for reorganization, through recruitment of rarely active neurons or rearrangement among neurons already active.

Selected SGD runs also show condensation and return, with gate switching present from the outset, extending the phenomenon beyond the temporal sequence observed under full-batch GD. Together, these findings reveal participation dynamics hidden by average sparsity and motivate an account of spectral return in which participation structure and gate switching act together.

2 GATING: STATE, FLOW, AND STRUCTURE

We use *gating* to describe whether a unit participates on a particular example. ReLU networks make this notion explicit: each preactivation induces a binary neuron–sample gate. For a fixed evaluation set $\mathcal{X} = \{x_s\}_{s=1}^{N}$ and a layer of width d_ℓ, define

$$A_{is}^{(\ell)}(t) = \mathbf{1}\Big\{ z_i^{(\ell)}(x_s;\theta_t) > 0 \Big\}, \qquad A^{(\ell)}(t) \in \{0,1\}^{d_\ell \times N}. \tag{1}$$

Rows describe neuron participation across examples and columns the participating neurons for each example. We use *participation* operationally to mean positive preactivation, not measured contribution to the final prediction. Below we suppress ℓ, write $d = d_\ell$, and set $M = dN$.

2.1 STATE

The support density

$$\rho_t = \frac{1}{M} \sum_{i,s} A_{is}(t) \tag{2}$$

measures how much support is active, but not how it is arranged (Fig. 1). We summarize the singular spectrum $A = U\Sigma V^\top$ by an effective rank $r_{\text{eff}}(A)$, which reflects both participation rates and their arrangement across examples. A decrease in effective rank indicates greater concentration of the normalized singular values, which we call *spectral condensation*. Condensation need not coincide with decreasing density.

For $A \neq 0$, we use the exponential of the entropy of its normalized singular values (Roy & Vetterli, 2007), computed on the raw, uncentered binary support:

$$r_{\text{eff}}(A) = \exp\left(- \sum_{j:p_j>0} p_j \log p_j \right), \qquad p_j = \frac{\sigma_j(A)}{\sum_k \sigma_k(A)}. \tag{3}$$

We set $r_{\text{eff}}(0) = 0$ by convention. This uses singular values, not their squares.

2.2 FLOW

Let $n_{ab}(t)$ count gates whose states on the same fixed examples are a at observation t and b at observation $t+1$, and define $r_{ab}(t) = n_{ab}(t)/M$. Total gate *turnover* and directional *current* are:

$$T_t = r_{01}(t) + r_{10}(t), \qquad J_t = r_{10}(t) - r_{01}(t). \tag{4}$$

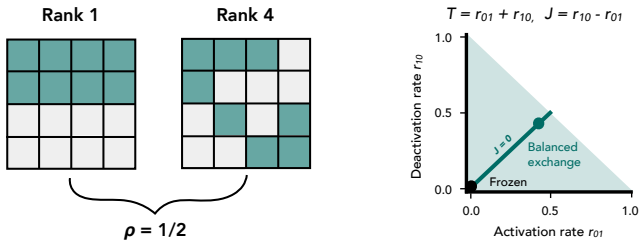

Figure 1: **Density does not determine support organization or gate dynamics.** Left: equal-density supports can have different rank. Right: gate flow decomposes into turnover $T = r_{01} + r_{10}$ and current $J = r_{10} - r_{01}$. The origin is frozen; along $J = 0$, substantial balanced exchange can occur without changing density.

Here T_t is the fraction of gates whose states differ between the two observations, while J_t is the excess of deactivation over activation, normalized by M. These endpoint comparisons do not count switches that reverse between observations. Counting active entries on the same evaluation set gives

$$\rho_{t+1} - \rho_t = -J_t. \tag{5}$$

Thus large turnover can coexist with little density change when the two directions nearly balance. We also report the directional bias $B_{\text{net}} = J/T$ for $T > 0$, read alongside turnover because a large bias can arise from very few switches; it is undefined when $T = 0$. The equations index a transition by its source state, whereas flow plots label it by its destination epoch.

2.3 Structure

The same matrix defines a bipartite neuron–sample graph. A neuron's degree $k_i = \sum_s A_{is}$ counts the examples in $\mathcal{X}$ on which it is active; we call it *isolated* on this set when $k_i = 0$, and denote the number of such neurons by I_t. Removing all zero-degree neuron and sample vertices defines the *active core*, represented by the remaining support submatrix A_+.

To distinguish neuron isolation from connectivity among the remaining active vertices, we analyze the normalized Laplacian of this active core,

$$\mathcal{L}_+ = I - D_+^{-1/2} \begin{bmatrix} 0 & A_+ \\ A_+^\top & 0 \end{bmatrix} D_+^{-1/2}. \tag{6}$$

Here D_+ is the diagonal degree matrix of the active-core bipartite graph. For a nonempty active core, the second-smallest eigenvalue $\lambda_{2,+}$ is positive exactly when the core is connected (Von Luxburg, 2007). We therefore report neuron isolates separately from the active-core Laplacian spectrum.

These observables describe the state, flow, and structure of gating, not a complete optimizer state: parameters and preactivation margins can evolve even when the binary support does not.

3 Experimental Setup

We train ReLU MLPs on fixed, balanced 10,000-example subsets of FashionMNIST and CIFAR-10. The primary architecture has two hidden layers of width 512 (2×512); depth and width comparisons use 4×512 and 2×1024. Full-batch GD minimizes $L = (2N)^{-1} \sum_s \|f_\theta(x_s) - e_{y_s}\|_2^2$, with $N = 10{,}000$, one-hot targets, constant learning rate, and no momentum or weight decay. SGD uses the same objective, with learning rates and batch sizes given in Table 1.

Supports on the full 10,000-example training subset are stored every 100 epochs, including initialization. Gate turnover and current are computed between consecutive epoch observations: shallow GD and SGD use supports recorded after each epoch on a fixed 2048-example probe; deep GD records all four layers on the full training subset before each update. An epoch contains one update under full-batch GD and multiple minibatch updates under SGD, so SGD turnover measures endpoint differences rather than all within-epoch switches. Flow axes use destination epochs. The

Table 1: **Main experimental comparisons. GD unless marked SGD.**

Setting	Learning rates	Purpose
FashionMNIST, 2×512	$.02,\ .04$	Learning rate and initialization
FashionMNIST, 4×512	$.01,\ .03$	Depth and gate switching
FashionMNIST, 2×1024	$.04$	Width and spectral return
CIFAR-10, 2×512	$.01,\ .03$	Cross-dataset comparison
Fashion/CIFAR, SGD 2×512	$.01$, $B = 256$	Stochastic comparison

density–current identity applies to the same evaluation set used for the corresponding flow measurements.

We report the normalized full-objective sharpness

$$\chi_t = \eta \lambda_{\max}(t)/2, \tag{7}$$

where $\lambda_{\max}(t)$ is the largest eigenvalue of the Hessian of L at θ_t, estimated numerically, and $\chi = 1$ is the local quadratic GD stability reference. Full-objective sharpness is evaluated every 25 epochs. For SGD, we estimate Batch Sharpness by averaging $g_B^\top H_B g_B / \|g_B\|_2^2$ over sampled batches, where $g_B = \nabla_\theta L_B$ and $H_B = \nabla_\theta^2 L_B$ are evaluated at the same parameters for the same batch loss L_B (Andreyev & Beneventano, 2024). Our estimator excludes batches with $g_B = 0$. This directional-curvature statistic is distinct from the largest full-objective Hessian eigenvalue; $2/\eta$ is shown as a reference scale, not as an inferred SGD transition.

4 Results

4.1 Condensation precedes distinct return trajectories

Training initially lowers raw-support effective rank, often while density increases (Fig. 2). More neuron–example pairs can therefore participate even as the normalized singular spectrum becomes more concentrated. We use *return* to describe subsequent rank recovery together with density decrease; the two changes need not begin at the same time.

In shallow FashionMNIST, learning rate separates two paths out of condensation. At $\eta = .04$, rank recovers after reaching its minimum around epoch 400, following observed EoS contact; at $\eta = .02$, density decreases while rank remains nearly stalled. By epoch 3500, rank has recovered by about 22 at the higher rate and less than 1 at the lower rate, a contrast that persists in other initializations. Deep L4 instead follows a pinched return without an observed EoS crossing, while CIFAR-10 recovers beyond its initial rank.

The paths also differ in the order of recovery. Density turns before rank in shallow FashionMNIST, the two turn together at the resolution of the deep L4 trajectory, and rank turns first in CIFAR-10. Density alone would hide both the extent of rank recovery and its timing within the return trajectory.

To quantify the rank component of return, let t_r mark the observed rank minimum and define

$$R = \frac{r_{\text{eff}}(t_{\text{end}}) - r_{\text{eff}}(t_r)}{r_{\text{eff}}(0) - r_{\text{eff}}(t_r)}, \tag{8}$$

when the denominator is positive. At the chosen endpoint, $R = 0$ denotes no net rank recovery, $R = 1$ recovery to the initial rank, and $R > 1$ rank overshoot. Density decrease and the timing of recovery remain separate features of the joint trajectory.

The joint trajectory (ρ, r_{eff}) distinguishes support amount from its spectral concentration. At the control snapshots, centered effective rank normalized by a within-row permutation baseline does not show the same rebound (Appendix E). Return in the *raw* support spectrum reflects participation rates as well as their arrangement; it need not restore the initial support organization.

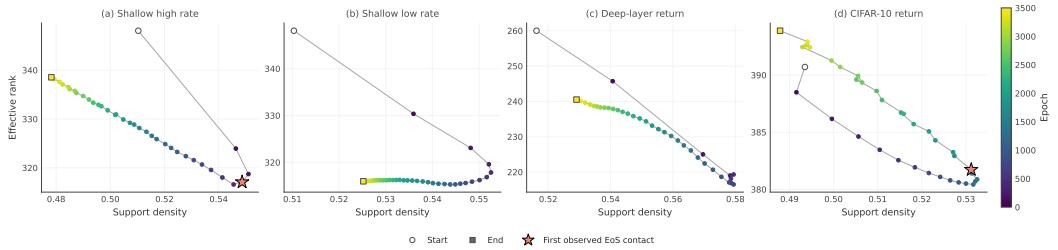

Figure 2: **Condensation gives way to distinct returns.** Density–raw-rank trajectories, seed 42, epochs 0–3500: Fashion 2×512, $.04/.02$, L2 (a/b); Fashion 4×512, $.03$, L4 (c); CIFAR 2×512, $.03$, L2 (d). Color: epoch; circles/squares: start/end; stars: first stored state at or after measured EoS contact. Shallow runs match Fig. 3.

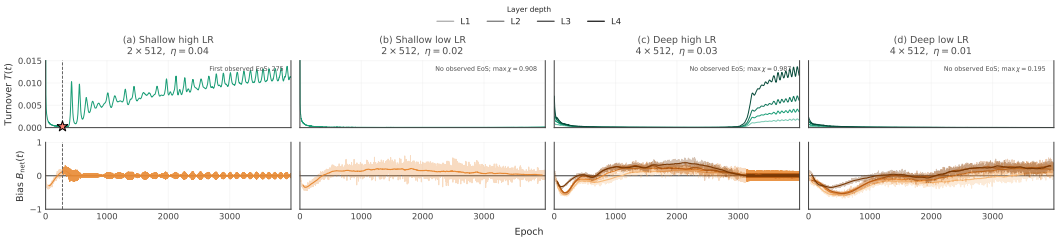

Figure 3: **Gate switching resumes as balanced exchange.** Fashion GD, seed 42, through destination epoch 3999. Shallow $.04/.02$: L2 on a 2048-example probe; deep $.03/.01$: L1–L4 on 10k training examples. Top: turnover. Bottom: raw directional bias (faint) and centered 101-epoch mean (dark); shade denotes depth. In the deep high-rate network, switching increases near the stability boundary.

4.2 Gate switching resumes as balanced exchange

A density–rank trajectory does not reveal how many gates switch beneath it. Across the Fashion-MNIST GD controls, turnover initially falls toward a nearly frozen support, yet the directional bias B_{net} makes broad excursions while very few gates switch (Fig. 3).

At the higher shallow rate, switching resumes near observed EoS contact as the smoothed directional bias approaches zero. Gates continue to switch even as the sustained preference for activation or deactivation fades. The lower-rate run remains nearly frozen. In the deep high-rate network, switching likewise increases across all four layers, most strongly in later layers, while the low-rate control shows little switching.

The deep case reveals that this late increase in switching is distinct from the density–rank return. In L4, return begins around epoch 700, when $\chi \approx .49$, more than 2000 epochs before switching rises sharply near the stability boundary. No crossing of $\chi = 1$ is observed (Appendix F). Return thus begins while switching is still sparse; the late increase cannot explain its onset.

Small density drift can therefore accompany two different kinds of support dynamics. During near-freezing, few gates change state. During balanced exchange, many gates switch but activation and deactivation nearly cancel. Directional bias must be read alongside turnover: a strong bias among a handful of switches need not produce a large current. Together, T and J distinguish little switching, switching dominated by one direction, and largely balanced exchange.

A small mean current can still conceal systematic variation from epoch to epoch.

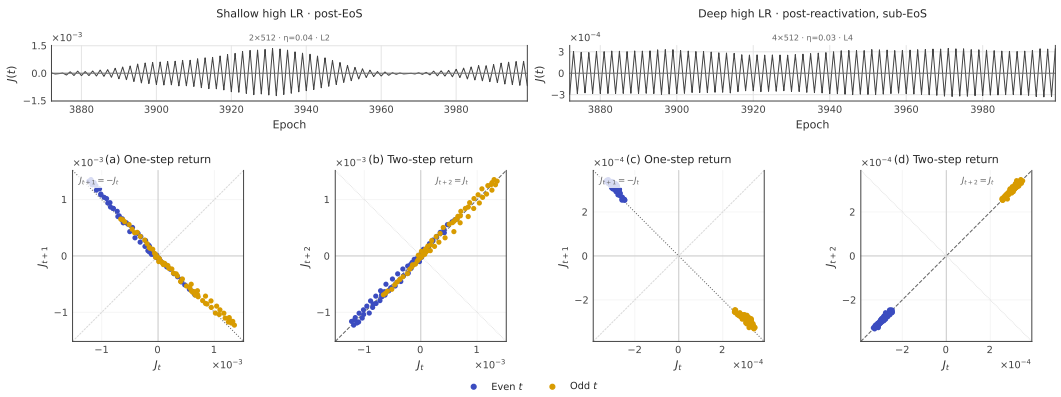

Figure 4: **Balanced exchange contains a period-two current.** Unsmoothed traces and return maps, destination epochs 3872–3999: shallow Fashion .04, L2 (left), and deep Fashion .03, L4 (right), seed 42. Colors distinguish epoch parity. The deep "sub-EoS" label denotes no sampled crossing; late measured $\chi = .912$–$.987$. Evaluation sets match Fig. 3.

4.3 Balanced exchange contains a period-two current

The small mean current conceals an alternating direction of switching. In the late high-rate shallow and deep examples, the current approximately obeys

$$J_{t+1} \simeq -J_t, \qquad J_{t+2} \simeq J_t. \tag{9}$$

Net deactivation in one epoch gives way to net activation in the next. In the return maps, consecutive currents lie near the negative diagonal, while currents two epochs apart lie near the identity (Fig. 4). Since $\rho_{t+1} - \rho_t = -J_t$, adjacent density changes nearly cancel. The sustained switching seen above thus accompanies little two-step density drift.

This alternation appears in both shallow initializations and in deep L4; lower-rate shallow training has a much weaker current and less regular alternation (Table 4). Early deep training provides a different contrast: its current is similar two epochs apart but does not alternate in sign. The period-two pattern is distinguished by the combination of one-step reversal and two-step recurrence.

In shallow training, this pattern follows observed EoS contact; in the deep example, it appears near the stability boundary without an observed crossing. The recurrence concerns the net balance of activation and deactivation: different gates can contribute two epochs apart, so the support itself need not repeat.

4.4 Participation profiles distinguish paths of return

The different return trajectories are accompanied by distinct participation profiles. Figure 5 relates these paths to class-associated support sharing and the prevalence of rarely active neurons.

Sharedness is the mean within-class minus mean between-class cosine similarity of sample support vectors, after centering each neuron's support by its mean participation. The marginal fraction counts neurons active on fewer than 1% of examples, including isolates. The two shallow learning rates produce similar sharing profiles, with only a few percent of neurons marginally active, despite markedly different rank recovery and gate switching. These two summaries alone therefore do not explain the shallow contrast.

To follow participation beyond a single threshold, write $q_i(t) = k_i(t)/N$ and define

$$F_t(q) = \frac{1}{d}\sum_{i=1}^{d} \mathbf{1}\{q_i(t) \le q\}, \qquad \Delta F_t(q) = F_t(q) - F_{\mathrm{ref}}(q). \tag{10}$$

The heatmaps show changes from a pre-turn reference selected by a common rule (Appendix E). A negative value means fewer neurons lie at or below that threshold, not that every neuron increases

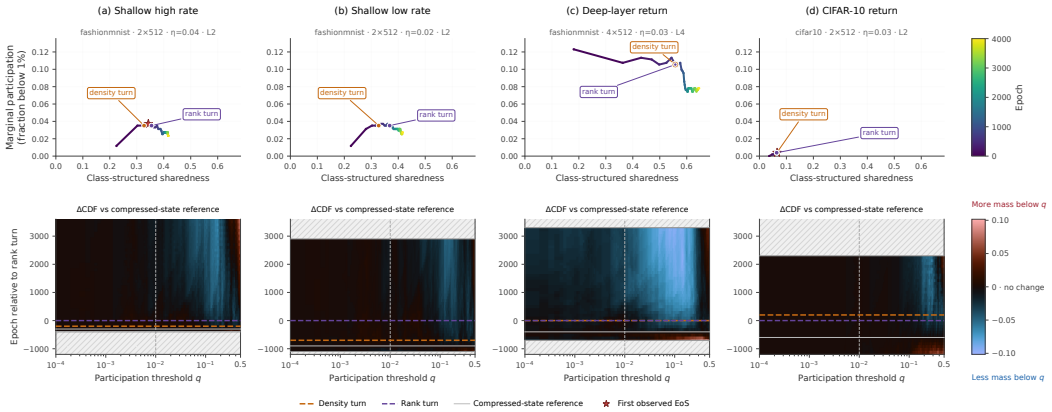

Figure 5: **Participation profiles distinguish return trajectories.** The experimental conditions and layers match Fig. 2. Top: class-associated sharedness versus the fraction of neurons active on fewer than 1% of examples. Bottom: change in the inclusive participation CDF relative to a pre-turn reference; red and blue indicate more and fewer neurons at or below threshold q, respectively.

its participation. The displayed thresholds cover the low-to-moderate participation range; fractions retain all neurons as their denominator. Unlike the marginal fraction, the CDF includes equality at the threshold.

In deep FashionMNIST L4, roughly a tenth of neurons are marginally active. Its heatmap shows fewer neurons at or below a range of low-to-moderate participation thresholds during return, before the late increase in gate switching. This redistribution can accompany decreasing density: changes in the lower-participation population need not follow the layer average. CIFAR-10, by contrast, returns with less than half a percent of neurons marginally active and a smaller class-sharing contrast. Return thus occurs with very different amounts of low participation and class-associated sharing; low class association need not mean little shared structure overall.

Individual participation and active-core connectivity. A changing participation distribution does not reveal which neurons leave or enter support. In deep L4, isolation is not always persistent: neurons with no active gates later participate again, at both learning rates. More of the neurons observed in isolation subsequently participate at the higher rate (Appendix D).The active core at each evaluated epoch show a single connected component despite these changes in individual participation. Whether this renewed participation contributes useful capacity is an open question.

4.5 SGD PRESERVES RETURN THROUGH A DIFFERENT TEMPORAL ROUTE

Spectral condensation followed by return also appears in the SGD examples (Fig. 6). In FashionMNIST, density and rank turn together at the resolution of the trajectory; in CIFAR-10, rank turns first and later exceeds its initial value. Their participation contrast also persists: FashionMNIST briefly develops an isolated neuron and a few-percent marginal population, while CIFAR-10 returns without observed isolation.

Gate switching follows a different temporal course. Epoch-level turnover is largest early and gradually declines, without the near-freezing and late increase seen under high-rate GD. The current fluctuates around zero with a shrinking RMS envelope, while Batch Sharpness remains high and noisy. Contact of full-objective sharpness with $2/\eta$ does not by itself identify an SGD stability transition.

These epoch-end comparisons span multiple minibatch updates. They record differences in gate states across an epoch, leaving switches that reverse within it uncounted; update-level alternation can therefore remain hidden. The SGD examples nevertheless show condensation and return alongside declining epoch-level turnover. The density–rank return thus extends beyond the particular timing of gate switching observed under GD.

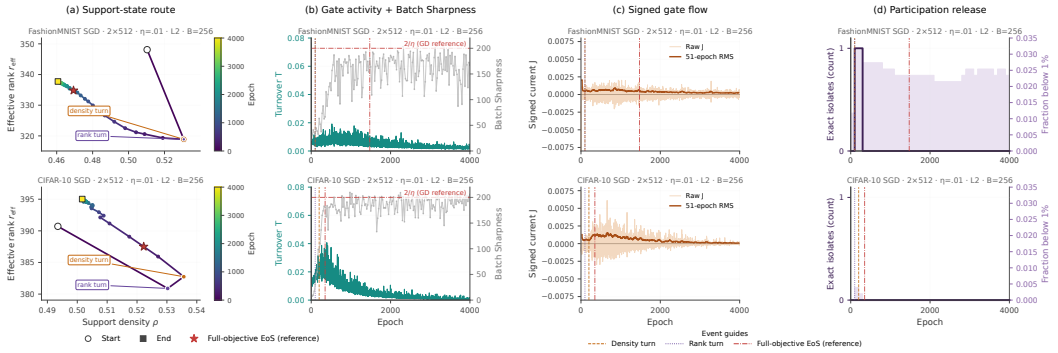

Figure 6: **SGD preserves return through a different temporal route.** Fashion (top), CIFAR (bottom): 2×512, .01, batch 256, seed 42, L2, epochs 0–4000. Left to right: density–rank path, turnover and Batch Sharpness, current with centered 51-epoch RMS, exact/marginal participation. Supports use 10k training examples; epoch-end flow uses a 2048-example probe. Star: full-objective GD contact; $2/\eta$: reference curvature scale, not an inferred SGD transition.

5 Discussion

Gate switching beneath stable density. Support density can look stationary while many neuron–example gates exchange states. Resolving that exchange reveals temporal organization: a small mean current can conceal approximate period-two alternation. Work on self-stabilization and trajectory alignment relates near-boundary optimization to oscillatory parameter dynamics (Damian et al., 2022; Song & Yun, 2023). Here, a related pattern appears in the net balance of activation and deactivation. Establishing whether the two share a mechanism requires relating parameter oscillations to gate transitions; recurring current alone does not imply recurring support.

The timing of return also matters. In shallow FashionMNIST, renewed switching near EoS accompanies substantial rank recovery, while deep density–rank return begins before the late increase in switching. SGD preserves condensation and return along a different epoch-level switching trajectory. These contrasts separate the path of return from one particular sequence of optimization events. They also complement work on stochastic stability, where batch directional curvature plays a different role from full-objective sharpness (Andreyev & Beneventano, 2024); the SGD observations here leave update-level switching unresolved.

Different returns, different opportunities for reorganization. Work linking large step sizes to sparse features makes participation a natural object of study (Andriushchenko et al., 2023). Our trajectories show why its average is only part of the account: density can rise as the support spectrum concentrates, and similar participation summaries can accompany different returns. A specific comparison comes from Hartmann et al. (2021), who report transient decreases in activation-pattern entropy in some architectures. Spectral condensation concerns a different object: the singular spectrum of neuron–sample support, rather than the frequency of distinct activation patterns. The centered controls further show that raw-rank recovery need not restore the earlier organization. Changing participation rates remain part of the phenomenon to be explained.

Our working hypothesis is that participation structure shapes the available paths of return from spectral condensation, while gate switching influences how those paths are traversed. Structural slack names the room a participation pattern may leave for reorganization. One possibility is recruitment of rarely active neurons; another is rearrangement among neurons already active. Deep FashionMNIST motivates the first possibility, while CIFAR-10 shows that a large marginal population is not required for return. Its weaker class-associated sharing also raises the possibility of fewer class-wide constraints on rearrangement, although this statistic does not measure those constraints directly. The shallow contrast supplies the complementary observation: similar profiles accompany very different recovery and switching. On this account, the same amount of gate switching need not have the same consequence for return: its effect depends on which participation relationships can change.

An alternative is that participation profiles mainly reflect optimization dynamics without themselves shaping the paths of return. Restricting recruitment of rarely active neurons, compared with comparable perturbations among already active neurons, would test one proposed route. A selective loss of recovery would support a role for recruitment; unchanged recovery would weaken that account. Changes in loss and curvature would also need to inform the interpretation.

From collective redistribution to individual participation. A participation distribution does not retain neuron identity: similar distributions can contain different neurons, and a shifted distribution does not identify which neurons supplied the change. The deep histories resolve one part of this ambiguity by showing that neurons observed in isolation can later participate. The active core is connected at each evaluated epoch, so low participation and a shared connected component can coexist. Connectivity describes how support is linked; it does not establish uniform participation or useful capacity.

Neural path features separate input-dependent gating from path weights (Lakshminarayanan & Vikram Singh, 2020), motivating a further question about the functional consequences of redistribution. In deep reinforcement learning, recycling neurons with low normalized activation can improve learning (Sokar et al., 2023). Our observations concern activation frequency: neurons inactive across the evaluation set can later become active through training alone. Whether this renewed participation contributes to predictions would help clarify its role in learning.

Together, the spectrum, switching, and neuron histories reveal how participation can reorganize beneath its average density. The common thread is spectral condensation followed by distinct paths of return, whose relationship to gate switching depends on the training regime. This raises the question of how participation structure and optimization dynamics interact to shape those paths.

6 Conclusion

Following binary support reveals spectral condensation and distinct paths of return hidden by average activity. Rank recovery and density decrease differ in their timing and relationship to gate switching: deep return begins before the late increase in switching, while SGD follows a different temporal route. In the high-rate GD examples, balanced exchange contains approximate period-two current. Participation profiles and neuron histories further distinguish collective redistribution from changes in individual participation.

The findings concern ReLU MLPs trained with MSE on FashionMNIST and CIFAR-10. Their temporal resolution is set by the observation intervals, especially for SGD, where each epoch spans multiple updates. The observables describe participation and its redistribution; their relationship to structural slack and predictive contribution is an open question.

Together, the findings motivate the hypothesis that participation structure shapes the available paths of return, while gate switching influences how those paths are traversed. The central question is whether participation structure helps shape these paths or mainly reflects the optimization dynamics that produce them. The contribution is to make this relationship visible: understanding learning involves explaining how participation is organized and reorganized alongside changes in activation values.

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 critically explore interpretations, connections 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 support, spectral, and gate-switching observables, and Section 3 describes the experimental setup. Appendices A and B detail data preparation, training conditions, measurement procedures, and the experiments shown in each figure. Appendix C presents initialization comparisons and defines the current-recurrence measures. Appendices D and E specify the participation, connectivity, and permutation analyses, including turn times and reference states. Appendix F describes the sharpness calculations and the criterion for renewed gate switching.

References

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.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd. *arXiv preprint arXiv:2412.20553*, 2024.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pp. 903–925. PMLR, 2023.

Melih Barsbey, Lucas Prieto, Stefanos Zafeiriou, and Tolga Birdal. Large learning rates simultaneously achieve robustness to spurious correlations and compressibility. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2055–2066. IEEE, 2025.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. *Advances in neural information processing systems*, 32, 2019.

David Hartmann, Daniel Franzen, and Sebastian Brodehl. Studying the evolution of neural activation patterns during training of feed-forward relu networks. *Frontiers in artificial intelligence*, 4: 642374, 2021.

Chandrashekar Lakshminarayanan and Amit Vikram Singh. Neural path features and neural path kernel: Understanding the role of gates in deep learning. *Advances in Neural Information Processing Systems*, 33:5227–5237, 2020.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in neural information processing systems*, 32, 2019.

Michael S Matena and Colin Raffel. A combinatorial perspective on the optimization of shallow relu networks. *Advances in Neural Information Processing Systems*, 35:22187–22198, 2022.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. *Advances in neural information processing systems*, 27, 2014.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. In *International conference on machine learning*, pp. 4558–4566. PMLR, 2018.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 32145–32168. PMLR, 2023.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. *arXiv preprint arXiv:2307.04204*, 2023.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of relu networks. *Advances in Neural Information Processing Systems*, 36:35654–35747, 2023.

A EXPERIMENTAL AND MEASUREMENT DETAILS

Data and training. We use balanced subsets of 10,000 training examples from FashionMNIST and CIFAR-10, with 1,000 examples per class and data-selection seed 42. The subset is fixed across initialization seeds. Images are normalized and flattened before entering the MLP. FashionMNIST uses mean .2860 and standard deviation .3530; CIFAR-10 uses channel means $(.4914, .4822, .4465)$ and standard deviations $(.2470, .2435, .2616)$.

Hidden layers use ReLU activations and biases, followed by a linear ten-output classifier. Training minimizes

$$L(\theta) = \frac{1}{2N} \sum_{s=1}^{N} \|f_\theta(x_s) - e_{y_s}\|_2^2, \qquad N = 10{,}000,$$

with one-hot targets. Learning rates are constant; there is no momentum, weight decay, or label noise. Full-batch GD takes one update using the gradient over all training examples. SGD retains the final incomplete batch, giving 40 updates per epoch for batch size 256 and 10 for batch size 1024. Table 2 specifies the conditions used in the main comparisons.

Support and effective rank. For each layer, $A_{is} = 1$ when neuron i has positive preactivation on example s, and zero otherwise. Full-subset supports are evaluated at initialization and every 100 epochs. Effective rank uses the singular values of the raw binary matrix:

$$r_{\text{eff}}(A) = \exp\left(- \sum_{j:p_j>0} p_j \log p_j \right), \qquad p_j = \frac{\sigma_j}{\sum_k \sigma_k}.$$

We set $r_{\text{eff}}(0) = 0$. Singular values are obtained from the smaller of $AA^\top$ and $A^\top A$. Before taking square roots, eigenvalues below $\epsilon \max(d, N) \max(1, \lambda_{\max})$ are set to zero, where ϵ is machine precision and $\lambda_{\max}$ is the largest eigenvalue of that Gram matrix. The entropy uses singular values, not their squares. Centered measurements use $A - q\mathbf{1}^\top$, where $q_i = N^{-1} \sum_s A_{is}$.

Gate switching and observation times. Shallow GD and SGD compare supports on a fixed 2048-example probe at consecutive epoch ends. Deep GD compares all four layers on the full training subset before consecutive updates. Turnover and current are normalized by the number of evaluated neuron–example pairs:

$$T_t = r_{01}(t) + r_{10}(t), \qquad J_t = r_{10}(t) - r_{01}(t).$$

On the same evaluation set, $\rho_{t+1} - \rho_t = -J_t$. Plot axes label each transition by its destination epoch. Epoch-end observations already use this index; a deep pre-update observation associated with training epoch e describes the state after update $e - 1$. SGD comparisons span an entire epoch and omit switches that reverse within it.

Directional bias is $B_{\text{net}} = J/T$ for $T > 0$. The main GD flow figure uses a centered 101-epoch mean of the per-epoch bias, using complete finite windows. The initialization-comparison figure uses a trailing 101-epoch mean of turnover. SGD current is shown together with a centered 51-epoch RMS envelope. The period-two analysis uses unsmoothed current.

Sharpness. Full-objective sharpness is the largest-algebraic Hessian eigenvalue, estimated by Lanczos with 50 iterations and sampled every 25 epochs in the main experiments. We use $\chi = \eta\lambda_{\max}/2$. For SGD, Batch Sharpness averages

$$\frac{g_B^\top H_B g_B}{\|g_B\|_2^2}, \qquad g_B = \nabla L_B(\theta), \quad H_B = \nabla^2 L_B(\theta),$$

over independently sampled batches at fixed parameters, omitting zero-gradient batches (Andreyev & Beneventano, 2024). The gradient and Hessian in each quotient use the same batch. This directional curvature differs from full-objective sharpness; $2/\eta$ supplies a reference scale in the SGD figure.

B EXPERIMENTAL CONDITIONS

Table 2 gives the training conditions, and Table 3 identifies the layers and comparisons used in each main figure.

Table 2: Conditions for the main comparisons. Architectures give hidden-layer count and width. Endpoints refer to support trajectories; when two seeds are listed, endpoints follow the same order. All use the fixed training subset described above.

Dataset	Hidden layers	η	Seed	Optimizer	Endpoint
FashionMNIST	2×512	.04	42,43	GD	4000,4000
FashionMNIST	2×512	.02	42,43	GD	4000,3900
FashionMNIST	4×512	.03	42	GD	4000
FashionMNIST	4×512	.01	42	GD	4000
CIFAR-10	2×512	.03	42,43	GD	3500,3400
FashionMNIST	2×512	.01	42	SGD, $B = 256$	4000
CIFAR-10	2×512	.01	42	SGD, $B = 256$	4000

Table 3: Layers and comparisons shown in the main figures. Unless stated otherwise, initialization seed is 42.

Figure	Experimental conditions
2	FashionMNIST: shallow GD at .04 and .02 (L2), deep GD at .03 (L4); CIFAR-10: shallow GD at .03 (L2). Trajectories shown through epoch 3500.
3	FashionMNIST: shallow GD at .04 and .02 (L2), deep GD at .03 and .01 (L1–L4). Destination epochs through 3999.
4	FashionMNIST: shallow GD at .04 (L2) and deep GD at .03 (L4), destination epochs 3872–3999.
5	The conditions and layers of Fig. 2, through their endpoints in Table 2.
6	FashionMNIST and CIFAR-10: shallow SGD at .01, batch size 256 (L2).

C INITIALIZATION DEPENDENCE AND CURRENT RECURRENCE

The shallow FashionMNIST contrast between substantial and nearly stalled rank recovery persists across the two initializations. Both CIFAR-10 trajectories recover beyond their initial rank.

One-step reversal and two-step recurrence. We compare early (1–128) and late (3871–3998) destination-epoch windows using unsmoothed current. For lag k, let P_k contain the times t for which both t and $t + k$ lie in the window. Define

$$E_1 = \sqrt{\frac{\sum_{t \in P_1} (J_{t+1} + J_t)^2}{\sum_{t \in P_1} J_t^2}}, \qquad E_2 = \sqrt{\frac{\sum_{t \in P_2} (J_{t+2} - J_t)^2}{\sum_{t \in P_2} J_t^2}}.$$

Small E_1 indicates sign reversal with similar magnitude; small E_2 indicates two-step recurrence. Ratios are omitted when RMS current is at most $1/(dN)$. Early deep current illustrates why both are needed: it changes little over two steps without alternating in sign. More generally, differencing independent, equal-variance density fluctuations produces lag-one current correlation $-1/2$, so negative correlation alone does not identify period two.

D INDIVIDUAL PARTICIPATION AND ACTIVE-CORE CONNECTIVITY

We follow each neuron's identity across support observations within a training trajectory. Isolation means zero participation on the fixed evaluation set; later activity means positive participation at a subsequent observation.

For connectivity, remove all zero-degree neuron and sample vertices. Two remaining neurons are adjacent when they share an example, as specified by the nonzero pattern of $A_+ A_+^\top$. The connected

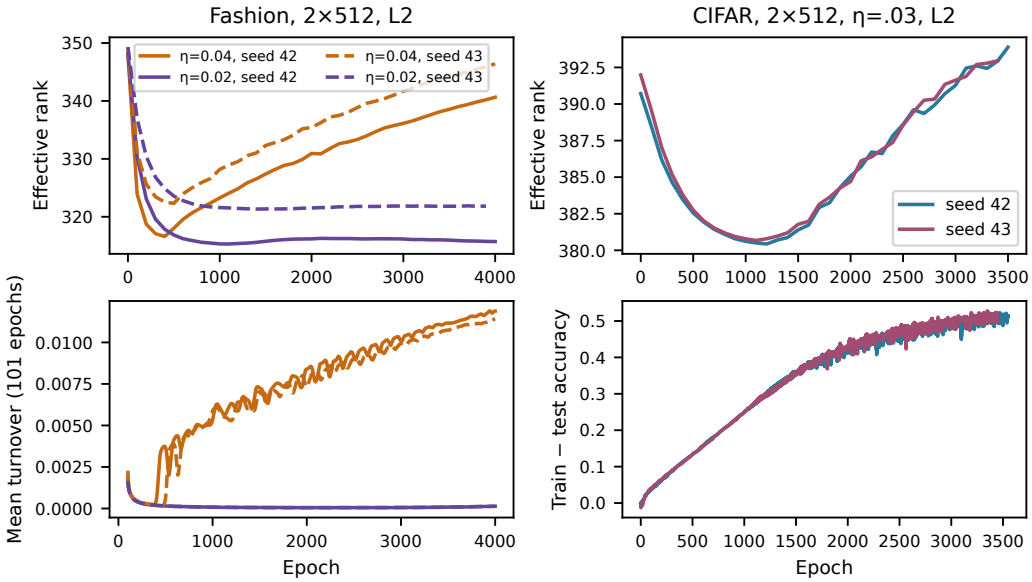

Figure 7: Rank recovery and gate switching across initialization seeds. Top: raw-support effective rank in shallow FashionMNIST and CIFAR-10. Bottom left: FashionMNIST probe turnover, averaged over trailing 101-epoch windows. Bottom right: CIFAR-10 training accuracy minus test accuracy. Curves show individual initializations on the same training subset.

Table 4: Current recurrence in FashionMNIST. Shallow measurements use L2 and deep measurements L4. RMS current and mean turnover specify the amplitude alongside the normalized recurrence errors.

Depth	η	Seed	Window	RMS J	E_1	E_2	Mean T
Shallow	.04	42	Early	.000687	1.49	.799	.00182
Shallow	.04	42	Late	.000652	.101	.0942	.0117
Shallow	.04	43	Early	.000643	1.44	.765	.00179
Shallow	.04	43	Late	.000437	.108	.115	.0114
Shallow	.02	42	Early	.000379	1.77	.567	.00145
Shallow	.02	42	Late	$1.21\,10^{-5}$	.838	.915	.000132
Shallow	.02	43	Early	.000357	1.71	.518	.00141
Shallow	.02	43	Late	$1.29\,10^{-5}$	.975	.791	.000126
Deep	.03	42	Early	.000449	1.99	.109	.00213
Deep	.03	42	Late	.000298	.0767	.0246	.0125
Deep	.01	42	Early	.000134	1.99	.167	.00111
Deep	.01	42	Late	$8.49\,10^{-6}$	1.96	.337	$2.85\,10^{-5}$

Table 5: Isolation and subsequent participation in deep FashionMNIST. "Ever isolated" counts distinct neurons; "Later active" counts those subsequently observed with positive participation, including neurons isolated at initialization. The minimum active-core gap is taken over initialization, the final-layer rank minimum, and the endpoint.

η	Layer	Isolate range	Ever isolated	Later active	Min. $\lambda_{2,+}$
.03	L1	0–0	0	0	.622
.03	L2	0–1	1	0	.638
.03	L3	7–16	21	11	.618
.03	L4	12–28	41	29	.597
.01	L1	0–0	0	0	.622
.01	L2	0–1	1	0	.659
.01	L3	7–17	21	8	.649
.01	L4	12–29	41	20	.628

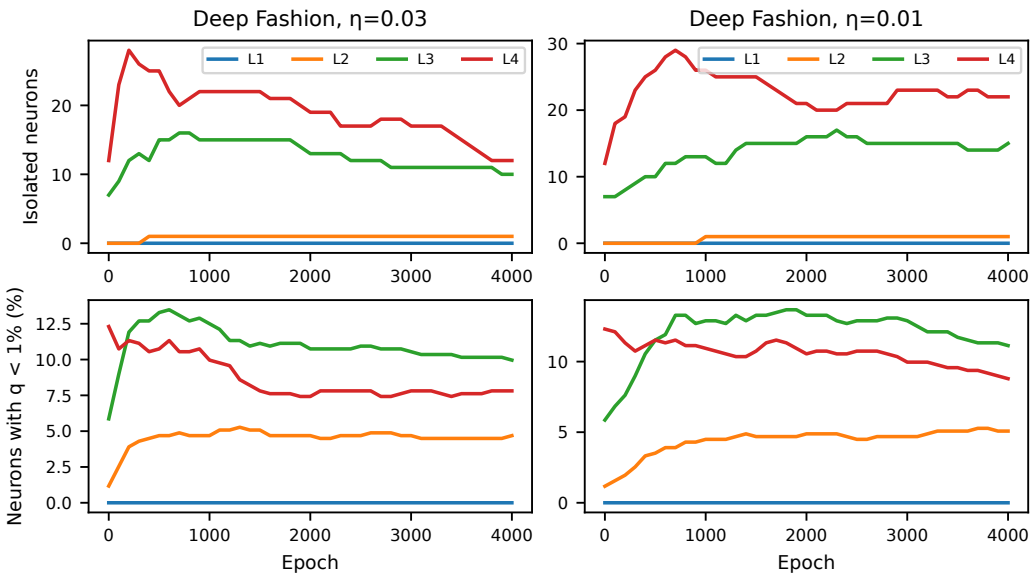

Figure 8: Layerwise participation in deep FashionMNIST, seed 42. Isolated neurons have $q_i = 0$; marginal neurons have $q_i < .01$, including isolates. Fractions use all 512 neurons in the layer. Points are evaluated every 100 epochs.

components of this neuron graph correspond to those of the active bipartite graph because every retained sample has a neuron neighbor. Writing

$$W = D_r^{-1/2} A_+ D_c^{-1/2}, \qquad \lambda_{2,+} = 1 - \sigma_2(W),$$

D_r and D_c are the neuron and sample degree matrices. The deep active cores are connected at the three stages in Table 5, even as individual neurons enter or leave support.

E Participation measures and additional comparisons

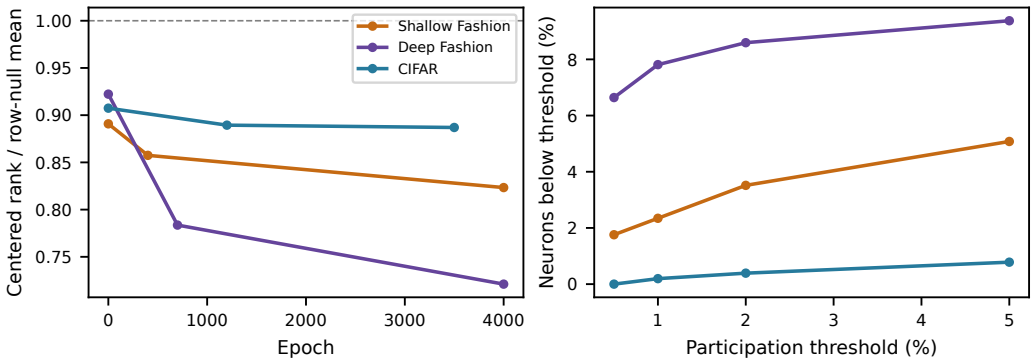

Figure 9: Centered-rank and participation-threshold comparisons. Left: centered effective rank divided by its mean under three independent within-row permutations, at initialization, the raw-rank minimum, and the endpoint. Each permutation preserves neuron participation rates while redistributing examples independently within rows. Right: fractions with participation strictly below each threshold, at epoch 4000 for FashionMNIST and 3500 for CIFAR-10.

Centered support and permutation baseline. The centered-rank ratio compares $r_{\text{eff}}(A - q\mathbf{1}^\top)$ with the mean centered effective rank of three independently row-permuted supports. This preserves

row degrees but not sample degrees. In shallow high-rate FashionMNIST, centered rank rises from about 400 to 403 after the raw-rank minimum, while its ratio to this baseline falls from .857 to .823. Thus raw-rank recovery need not be accompanied by recovery in the centered, baseline-normalized statistic.

Class-associated sharedness. For each sample, define $v_s = A_{:s} - q$. Sharedness is the mean cosine similarity for same-class pairs minus that for different-class pairs:

$$S = \frac{1}{|P_{\text{same}}|} \sum_{(s,u)\in P_{\text{same}}} \frac{v_s^\top v_u}{\|v_s\|\|v_u\|} - \frac{1}{|P_{\text{diff}}|} \sum_{(s,u)\in P_{\text{diff}}} \frac{v_s^\top v_u}{\|v_s\|\|v_u\|}.$$

The pair sets contain distinct samples with nonzero centered vectors. Pairs are weighted equally within each set. A class-association baseline is obtained by permuting class labels 20 times while holding support fixed. Sharedness measures association with these labels, not all possible shared structure.

Turn times and participation reference. Density turns mark a change from increasing to decreasing density sustained for two adjacent support intervals on each side. The SGD comparison permits an early turn with one preceding interval and two following intervals. Rank turns mark an interior minimum of raw effective rank. These times have the 100-epoch resolution of the support observations.

The pre-turn reference for the participation heatmaps is the support observation nearest

$$\frac{t_0 + \min(t_\rho, t_r)}{2},$$

where t_0 is initialization and t_ρ, t_r are the density and rank turn times; ties select the earlier observation. In panel order, the references are epochs 100, 200, 300, and 600. The heatmaps show $\Delta F_t(q) = F_t(q) - F_{\text{ref}}(q)$ with

$$F_t(q) = \frac{1}{d} \sum_i \mathbf{1}\{q_i(t) \le q\}.$$

This CDF includes equality, whereas the marginal fraction uses $q_i < .01$. Both include isolates. Fractions retain all neurons as the denominator even when only low-to-moderate thresholds are displayed. A negative difference means fewer neurons at or below that threshold relative to the reference.

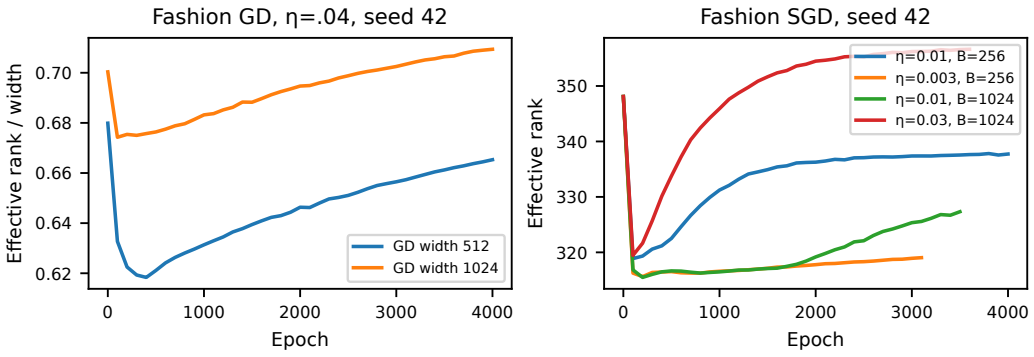

Figure 10: Width and SGD settings in FashionMNIST, initialization seed 42. Left: effective rank divided by layer width for 2×512 and 2×1024 GD networks at $\eta = .04$. Right: final-hidden-layer effective rank for 2×512 SGD networks with $(\eta, B) = (.01, 256), (.003, 256), (.01, 1024), (.03, 1024)$.

The width comparison holds learning rate fixed, rather than matching sharpness across widths. The SGD comparisons vary learning rate at fixed batch size and batch size at fixed learning rate. The four SGD trajectories extend to epochs 4000, 3100, 3500, and 4000, respectively.

F SHARPNESS AND THE TIMING OF DEEP-LAYER RETURN

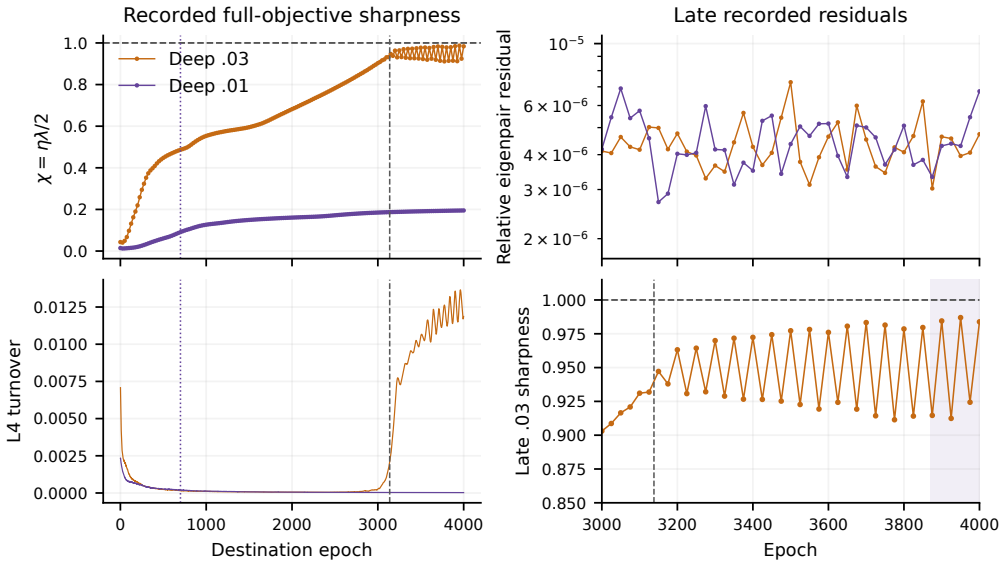

Figure 11: Deep FashionMNIST at learning rates .03 and .01, initialization seed 42. Left: full-objective sharpness and L4 turnover. The vertical guides mark density–rank return at epoch 700 and the later increase in mean-layer turnover at epoch 3138. Right: eigenpair residuals and late high-rate sharpness. Shading marks the current-recurrence figure's window, epochs 3872–3999. Sharpness is sampled every 25 epochs.

Deep density–rank return precedes the late increase in gate switching. At the high-rate return near epoch 700, $\chi \approx .49$. The later switching increase occurs near $\chi \approx .94$, with sampled sharpness remaining below one and reaching about .987. The low-rate comparison reaches about .195.

Switching onset. Let $\overline{T}$ be turnover averaged across the four layers. We identify a low-switching interval of at least 100 consecutive epochs with $\overline{T} \leq 10^{-4}$, followed by at least 50 consecutive epochs with $\overline{T} \geq 10^{-3}$. The first epoch of the latter interval is the switching onset, here 3138. During the late current-recurrence window, sampled χ ranges from .912 to .987.

Numerical accuracy. Sharpness uses all 10,000 training examples. The 50-iteration Lanczos estimates have relative eigenpair residuals below 10^{-5}. At epoch 4000, repeating the calculation in double precision with 100 iterations and two random starts gives $\lambda_{\max} \approx 65.5892$ and $\chi \approx .98384$, agreeing with the original estimate to within 5×10^{-6} in χ. The double-precision residuals are below 3×10^{-14}. These calculations support near-boundary dynamics at the sampled states; the sampling interval leaves intermediate crossings unresolved.