OBSERVATION DOES NOT IMPLY CONTROL: THE DYNAMICS OF DIRECTIONAL CURVATURE INTERVENTIONS AT THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Curvature estimation has made sharp directions increasingly accessible in neural-network optimization, but a measured direction need not respond predictably to intervention. We study directional gradient edits at the Edge of Stability (EoS). In a full-batch MLP, fixed-divisor edits provide an interventional test of known preconditioned-stability algebra: conditional on trajectories that exhibit the corresponding sign-flip orbit, late raw sharpness tracks its predicted location $2d/\eta$, while orbit survival is separate and is not explained by the reduced model. In stochastic ResNet-18/CIFAR-100 training, an engaged rank-1 attenuation arm and wider raw-divisor top-eigenspace rescalings through $k{=}20$ show no statistically resolved reduction in the tested sharpness diagnostics at the canonical operating point; the wider rescaling rule is not unconditional pure damping because below-target directions receive bounded sign-reversed amplification. We use SAM as a sharpness-reduction control: the tested $\pm H\mathbf{g}$ first-order corrections do not reproduce its ResNet response, and the sampled linSAM-to-SAM interpolation associates lower sharpness with increasing weight on the finite re-query residual bundle. This pathwise decomposition does not isolate a unique nonlinear mechanism and does not transfer in the same form to the LayerNorm ViT control. These results distinguish observing curvature from predicting the response to a specific curvature intervention, within the tested systems, measurements, and intervention rules.

1 INTRODUCTION

Modern curvature estimators make the sharp directions responsible for instability in neural-network optimization increasingly accessible. Measuring a sharp direction, however, does not by itself guarantee effective control under the intervention families tested here. Gradient descent at the Edge of Stability (EoS) provides a natural testbed for this question: the Hessian top eigenvalue $\lambda_{\max}$ climbs toward the linear-stability threshold $2/\eta$ and oscillates there while loss decreases non-monotonically (Cohen et al., 2021; Damian et al., 2023). This accessibility motivates a natural intervention: modify the gradient along sharp Hessian directions in an attempt to control sharpness; whether this intuition holds for neural-network optimization is the subject of this paper.

We organize the experiments around a known piece of threshold algebra, stated here for reference (Cohen et al., 2022; Islamov et al., 2026): a divisor-d edit along a sharp direction rescales the effective local step to $\eta_{\text{eff}} = \eta/d$, so the linear-stability location of a sign-flip period-2 orbit moves from $2/\eta$ to

$$\lambda_{\text{crit}} = 2d/\eta = 2/\eta_{\text{eff}}, \tag{1}$$

Lemma 1 establishes this orbit location in the 2-D self-stabilization reduction; our contribution is a feedback-free interventional test of the reduced-map orbit-location prediction on trajectories that exhibit the corresponding sign-flip behavior (§4), and of what replaces this picture in the stochastic regime (§5).

Across our experiments, two failure modes emerge within the tested intervention families:

- *Geometry relocation.* The edit acts but moves its own target: damping a sharp direction by $1/d$ relocates the sign-flip stability boundary from $2/\eta$ to $2d/\eta$, and surviving trajectories track the

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.

moved boundary upward instead of descending: the intervention moves the very stability constraint it was designed to control.

- *Stochastic mismatch between the measured curvature direction and the effective control response.* At the reference stochastic operating point, the edited eigenspace carries little of the current minibatch gradient and the intervention produces no resolved sharpness reduction; the small gradient share diagnoses weak immediate actuation but does not by itself explain the long-horizon null (§5.2).

The two modes are therefore limiting cases connected by an exploratory regime sweep, rather than disjoint categories assigned by regime label. Their boundaries are system- and protocol-dependent: a ViT control reproduces the unresolved damping response and even shows the relocation direction at its high-curvature point, yet the ResNet linSAM–SAM separation does not transfer in the same form. The organization summarizes the tested outcomes; it is not a predictive classifier for unseen systems or protocols, nor an assertion that sharpness is uncontrollable: full-batch rank-1 amplification and stochastic SAM both show that sharpness can be reduced by intervention in the tested settings.

The central message of this work is that curvature is not an external quantity available for direct manipulation: the optimizer dynamics that create a sharp direction are themselves altered when that direction is modified. In stochastic ResNet-18/CIFAR-100 training, the corrected rank-1 attenuation arm and the tested wider raw-divisor rescalings through $k{=}20$ show no statistically resolved sharpness reduction, and neither $\pm H\mathbf{g}$ first-order correction reproduces SAM's; along a direct linearized-to-full-SAM interpolation the four sampled arm means decrease (with no large change in the directly instrumented pre-update norm), providing a pathwise residual-bundle decomposition that a LayerNorm ViT control shows to be system-dependent rather than universal.

Contributions.

- **Intervention map across regimes.** We organize the tested directional interventions by two observed response patterns: relocation of the sign-flip orbit location in the full-batch reduction, and an unresolved sharpness response under the tested stochastic eigenspace rescalings: a descriptive organization, not a predictive regime classifier (the exploratory batch-size sweep shows that weak instantaneous actuation alone does not determine the long-horizon response). At the canonical ResNet-18/CIFAR-100 operating point, neither the engaged rank-1 attenuation arm nor the wider $k{\leq}20$ raw-divisor signed eigenspace rescalings (which are not unconditional pure damping) show a statistically resolved sharpness reduction; the edited subspace's small gradient share is an actuation diagnostic, not an explanation of the long-horizon null.
- **Conditional interventional test of orbit relocation.** The shift $2/\eta \to 2d/\eta$ is known stability algebra rather than a new threshold law. We test its network-level orbit-location consequence without adaptive feedback: among fixed-divisor trajectories that exhibit the corresponding stationary sign-flip behavior, late raw sharpness tracks the predicted location, while orbit survival is reported separately and is not explained by the reduced model.
- **Pathwise SAM decomposition.** Neither tested $\pm H\mathbf{g}$ first-order correction reproduces SAM's ResNet sharpness reduction. Across the sampled linSAM-to-SAM interpolation, increasing the finite re-query residual bundle is accompanied by decreasing sharpness. This identifies what differs along this particular path, not an isolated universal SAM mechanism; the ViT control gives a different decomposition.

2 Related work and positioning

EoS core and its stochastic counterpart. Prior EoS analyses characterize how sharpness emerges and stabilizes under a given optimizer: the Hessian spectrum governs the stability limit ($\lambda_{\max} \to 2/\eta$) (Lewkowycz et al., 2020; Cohen et al., 2021); progressive sharpening, self-stabilization, and saturation theorems (Li et al., 2022; Damian et al., 2023; Agarwala et al., 2023); simplified-model and bifurcation-theoretic accounts (Chen & Bruna, 2023; Ahn et al., 2023; Song & Yun, 2023); sign-flip/route-to-chaos descriptions (Kalra & Barkeshli, 2023; Kalra et al., 2025); implicit sharpness-minimization (Arora et al., 2022) and time-averaged central flows (Cohen et al., 2025): largely deterministic/full-batch analyses, with stochastic counterparts in batch sharpness (Andreyev & Beneventano, 2024; Andreyev et al., 2026) and preconditioned top eigenvalues (Cohen et al., 2022).

We study an interventional rather than algebraic extension of this literature: after modifying the optimizer along a measured sharp direction, does the original sharpness mechanism remain intact? Static preconditioning already implies that rescaling an update direction changes the relevant spectral stability quantity (Cohen et al., 2022), and recent work develops preconditioned GD as a generalized EoS setting (Islamov et al., 2026); we do not claim the $2d/\eta$ threshold shift itself as novel. Our full-batch contribution is the feedback-free fixed-divisor test of whether network trajectories exhibiting the relevant sign-flip behavior realize that location, together with an explicit separation between orbit location and orbit survival. Existing stochastic-EoS analyses characterize how minibatch noise and momentum modify the stability statistics of a fixed optimizer (Andreyev & Beneventano, 2024; Andreyev et al., 2026). Our batch-size map does not derive a competing stochastic equilibrium law; it is an exploratory intervention study over selected operating points and should not be read as a stochastic extension of the 2-D full-batch theorem.

Directional curvature edits. CDAT (Roulet et al., 2024) uses curvature along the update as a step-size denominator and embraces the $2/\eta$ threshold; critical sharpness (Kalra et al., 2026) defines λ_c via line search and states the Rayleigh identity we repurpose into a single-HVP monitor. Song et al. (2025) remove dominant Hessian-eigenvector components from the SGD update and observe constrained optimization with unconstrained sharpness: the $d \to \infty$ limit of our damping family. Our subspace-energy diagnostic differs operationally: it measures the current minibatch gradient's energy on the instantaneous top-k eigenspace of the current Hessian estimate, whereas Song et al. (2025) target a persistent dominant subspace aggregated over training; a small instantaneous share is not inconsistent with their bulk-subspace observations. Our finite raw-divisor rescalings are not equivalent to removing a persistent dominant subspace, and the stochastic wider-rank rule can change sign below target. We therefore use Song et al. (2025) primarily to position the broader question of Hessian-subspace interventions rather than to claim a direct algorithmic continuation between the two procedures. We vary the strength, sign, rank, sensing source, and stochasticity of the edit.

Measurement, normalization, and SAM. Warm-started power iteration tracks accurate estimates of $\lambda_{\max}$ at a few HVPs per step at 3B scale (Ramasinghe et al., 2026). Large batches find sharper minima (Keskar et al., 2017) with known caveats (Dinh et al., 2017); normalization reshapes sharpness (Wei et al., 2019; Lyu et al., 2022). SAM (Foret et al., 2021) is our escape benchmark; Long & Bartlett (2024) derive a SAM-specific EoS cutoff that SAM empirically tracks. We test vanilla SAM; scale-invariant and output-level variants (Kwon et al., 2021; Du et al., 2022) share a related amplifying structure, so our negative attenuation result does not transfer to them.

Closest to our stochastic analysis, Agarwala & Dauphin (2023) derive the sharpness level at which SAM stabilizes idealized dynamics as a function of η and ρ, Wen et al. (2023) isolate which parameter blocks SAM must perturb to retain sharpness reduction in deterministic settings, and Bartlett et al. (2023) analyze SAM's bounce-and-drift dynamics in minimal non-convex problems. Our ladder differs in aim: rather than characterizing SAM itself, we interpolate between a linearized and the full SAM update to test which update difference carries the endpoint separation along a controlled interpolation.

3 SETUP

Testbeds. Three testbeds carry the paper: (i) a tanh MLP $[20-256-256-1]$ teacher–student regression under full-batch GD; (ii) a CIFAR-adapted ResNet-18 (3×3 stride-1 stem, no max-pooling; 11.2M parameters) on CIFAR-100, trained under the deterministic 5000-step replayed mini-batch protocol (hyperparameters: Appendix F), a short constant-lr mechanistic protocol chosen to expose stationary sharpness dynamics and studied throughout §5; the 200-epoch external check (§5.4) tests the SGD/damping/SAM ordering under a standard recipe; and (iii) `timm` ViT-tiny (LayerNorm) on CIFAR-100 as a cross-system control (§5.4). Additional architectures used only for instrumentation and ablations (PreAct-ResNet, ConvNeXt, matched-step X-SAM projector ablation) are described in the appendices.

Sharpness quantities. Three sharpness quantities appear, and they are different objects: $\lambda_{\text{raw}} = \lambda_{\max}(H)$ is the raw Hessian top eigenvalue governed by the full-batch $2/\eta$ law; $\lambda_{\text{ref}} = \lambda_{\max}(H_R)$, the ARPACK top eigenvalue on a fixed 512-example evaluation batch (model

Symbol	Meaning	Used for
λ_{raw}	full Hessian top eigenvalue	full-batch GD boundary
λ_{ref}	fixed-batch ARPACK $\lambda_{\max}$	cross-method diagnostic
λ_{step}	batch gradient curvature	EoSS quantity (SGD–momentum); diagnostic for all arms

Table 1: The three sharpness quantities.

in eval mode), is the cross-method diagnostic used in the stochastic experiments, reported against the heavy-ball reference line $2(1+\mu)/\eta$, used as a fixed reference, *not* a theoretical constraint on H_R; and $\lambda_{\text{step}} = \mathbb{E}_B[\mathbf{g}_B^\top H_B \mathbf{g}_B / \|\mathbf{g}_B\|^2]$ is the batch sharpness, constrained by EoSS theory for SGD(-momentum) dynamics, which stabilizes at $2/\eta$ momentum-free and plateaus between $2(1-\mu)/\eta$ and $2(1+\mu)/\eta$ with momentum (Andreyev & Beneventano, 2024; Andreyev et al., 2026); we evaluate it as a diagnostic along all optimizer trajectories, including SAM's, without implying that SAM's update obeys the same $2/\eta$ boundary. Because H_R is evaluated through the model's stored BatchNorm state, λ_{ref} is a state-dependent protocol diagnostic rather than a parameter-only invariant of the network function. Appendix D.6 shows that changing only the BatchNorm-statistics convention can materially change its numerical level; cross-method comparisons are therefore interpreted only under the stated matched protocol. A three-fixed-batch calibration in Appendix D.1 shows comparable but non-identical scalar levels across batches, so λ_{ref} is used as a reproducible protocol-specific cross-method diagnostic, not as a population Hessian eigenvalue. One further distinction matters: the fixed-batch gradient used by the curvature monitor and the current stochastic training gradient used to measure intervention energy are different objects: the former aligns with the top eigenvector in the EoS regime, the latter nearly never does under bs-128 noise (§5.2). Table 1 summarizes the three.

Operational meaning of control. Throughout the paper, "control" means an intervention intended to reduce the reported raw or reference sharpness quantity; it does not denote control-theoretic controllability or closed-loop stability, nor do we identify raw Hessian sharpness with a parameterization-invariant generalization measure. In particular, a directional rescaling can regulate an effective or preconditioned stability quantity while raw $\lambda_{\max}$ increases. We therefore describe such an outcome as a failure only with respect to the stated raw-sharpness reduction objective.

Curvature and statistics. $\lambda_{\max}$ via power iteration + Pearlmutter HVP on the MLP path (cross-validated against ARPACK to $\leq 3 \times 10^{-7}$ relative); deterministic ARPACK at ResNet-18 scale. Headline reference-sharpness comparisons use $n{=}5$ seeds (mean$\pm$95% CI); instrumented batch-sharpness reruns use $n{=}3$ unless flagged in the run manifest. For ordinary paired arm comparisons, we first form one per-seed difference using matched seed IDs and the experiment's prespecified summary window, then report the mean paired difference with a two-sided 95% t confidence interval; we call a paired difference statistically resolved only when this 95% CI excludes zero. Practical-equivalence claims are evaluated separately with TOST at $\alpha{=}.05$ on the same paired per-seed differences. For TOST, the corresponding 90% CI is compared with the prespecified design margin, set before the rank-k equivalence experiments to one quarter of SAM's late-window reduction in the protocol at hand (numeric margins, their relation to the SGD reference levels, and the sensitivity analysis at stricter and looser margins, under which no tested arm satisfies equivalence at the stricter one-eighth margin, are in Appendix D.10). Thus 95% CIs support ordinary difference claims, whereas 90% CIs are used only for equivalence testing. The $m{=}4$ batch-sharpness estimate is intentionally compute-limited and consequently noisy (Monte-Carlo variance and a subsampling calibration: Appendix D.2). Training-time accuracy curves use a fixed 2,000-example test subset only for post-hoc visualization; all headline accuracy results use the full test set. A 95% CI containing zero is reported only as "not statistically resolved"; particularly for $n{=}3$ supporting arms, this is not evidence that the effect is absent or negligibly small. The appendix contains many secondary ablation comparisons. Unless a comparison is explicitly identified as prespecified, its confidence interval should be read as per-comparison descriptive evidence; we do not claim family-wise error control across the full ablation suite.

Protocol semantics. In the stochastic experiments, the corrected rank-1 arm is an engaged attenuation test, whereas the $k{>}1$ experiment uses the no-floor raw-divisor signed rescaling defined in §5.2. Current-minibatch-sourced eigenvectors are refreshed every $N{=}20$ steps; legacy sensing, clamp variants, and complete measurement semantics are detailed in Appendices C and D.

4 FULL-BATCH GD: FIXED-DIVISOR EDITS TEST A REDUCED-MAP ORBIT-LOCATION PREDICTION

4.1 REDUCED-MODEL PREDICTION

Lemma 1 (Reduced-map orbit location under directional rescaling; conditional on the 2-D self--stabilization recurrence of Damian et al. (2023))**.** *Assuming an isolated positive top eigenpair, window-fixed edit strength, accurate top-eigenvector sensing, and the reduced recurrence specified in Appendix A, every sign-flip period-2 orbit of that recurrence satisfies $\lambda^* = 2d/\eta$ with $d = 1/(1-\gamma) > 1$ for damping ($\gamma \in (0,1)$), and $\lambda^* = 2/((1+\alpha)\eta) < 2/\eta$ for the sign-reversed amplifying edit ($\gamma = -\alpha < 0$).*

Lemma 1 is an orbit-location statement, not a trajectory-level theorem: the reduced orbit is not proven attractive, and the network evidence in §4.2 is interventional.

As organized in §1, damping rescales the effective step along the sharp direction to η/d, moving the sign-flip condition to $2d/\eta$; the edit's contribution to the slow y-drift has zero cycle mean at odd order and only renormalizes the amplitude coefficient at even order (Appendix A), so the sign-flip orbit location remains $\lambda^* = 2d/\eta$.

The reduced orbit is indeed not guaranteed to attract: the corresponding orbit is repelling under the full two-dimensional Jacobian, so attraction in the network must involve higher-order effects omitted by the reduction (gap G4), and the lemma makes no claim that the orbit attracts nearby trajectories, that a network trajectory reaches it, or that the reduced recurrence explains the observed basin of survival. Section 4.2 tests only the orbit-location prediction, while separately reporting which fixed-divisor trajectories survive.

4.2 FIXED-DIVISOR INTERVENTIONAL TEST

The adaptive rule $d = \hat{\lambda}/(2/\eta)$ enforces $\hat{\lambda}/d = 2/\eta$ whenever the gate is active, so its trajectory cannot by itself test the predicted boundary shift. We therefore remove this feedback and hold $d = d_{\text{fix}}$ constant while keeping the edit and sensing procedure unchanged. On-orbit cells track the predicted location $\lambda^* = 2d_{\text{fix}}/\eta$ (Fig. 1a), and a fixed divisor matched to the adaptive arm's realized mean reaches the same sharpness scale, whereas intermediate divisors diverge on all seeds even though the adaptive controller remains stable. Thus orbit location and orbit survival are distinct effects: the fixed divisor tests the former, while the adaptive feedback affects the latter.

This separation persists as a conditional orbit-location observation across the $\eta \times d$ grid: seven cells exhibiting the corresponding stationary sign-flip behavior lie within 0.95–0.99 of the predicted location; two additional surviving cells are off the relation, and the remaining cells diverge (Fig. 1). Full per-cell survival counts and device-sensitive boundary cases are in Appendix C.5. We therefore do not interpret Fig. 1 as showing that fixed-divisor dynamics generally converge to $2d/\eta$: it shows that, conditional on the observed trajectory entering the corresponding sign-flip regime, its late location agrees closely with the reduced-map prediction. Orbit entry, attraction, and survival remain outside Lemma 1; the "on-orbit" and trajectory labels in Fig. 1 are descriptive summaries rather than outcomes of an independently preregistered orbit-classification test, so the fixed-divisor experiment is used as conditional evidence about orbit location only.

The survival window is itself structured (Fig. 1b): divergence concentrates at large ηd, stability is non-monotone in d, and marginal cells are device-numerics sensitive, so the higher-order dynamics bounding the orbit's basin (Remark 1, Appendix A) act well beyond the lemma's orbit-location content. A delayed-onset control addresses the possibility that the fixed-d failures are caused simply by intervening too early: switching the adaptive arm to fixed $d \in \{2.5, 3\}$ at its measured natural gate-opening step still yields no survivors at either d, while the adaptive rule remains stable from the same state: stabilization is attributable to the $\hat{\lambda}$-dependent feedback itself, not to delayed intervention or warm-up (Appendix C.5).

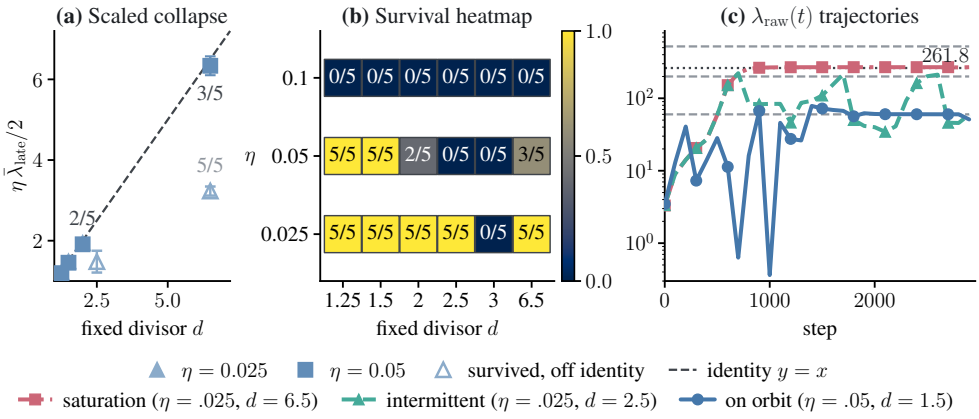

Figure 1: Unified fixed-divisor grid (MLP full-batch; one CPU round, float32; n=5 seeds per cell, 3000 steps; $\eta \in \{0.025, 0.05, 0.1\}$, $d_{\text{fix}} \in \{1.25, 1.5, 2, 2.5, 3, 6.5\}$). **(a)** Dimensionless late-window level $\eta\lambda_{\text{raw}}/2$ vs. d_{fix}: cells that settle onto a stationary orbit track the identity (filled; ratios 0.95–0.99); two cells at η=0.025 survive off it (open; ratios 0.59 and 0.50, the latter at $\lambda_{\text{raw}} \approx 260$). **(b)** Survival counts $n_{\text{surv}}/5$: divergence concentrates at large ηd (per-cell counts in §4.2). **(c)** Representative $\lambda_{\text{raw}}(t)$ trajectories (seed 0, late means match the grid; predicted levels $2d_{\text{fix}}/\eta$ dashed, rank-1 escape ceiling dotted; delayed-onset control: Appendix C.5). The three trajectory types in panel (c) are encoded redundantly by hue and line style/marker.

4.3 Sign reversal: accurate rank-1 attenuation raises the observed late-window sharpness in the tested full-batch system; amplification descends

Reversing the sign gives the complementary reduced-model prediction: amplification of the top-gradient component places the sign-flip orbit below $2/\eta$. Experimentally, rank-1 amplification lowers the late-window mean sharpness monotonically across $\alpha = 0.25/0.5/1$ (n=5; Fig. 7, Table 6; Appendix C). The observed values do not quantitatively follow the naive fixed-α orbit formula, a discrepancy that grows with the edit strength, so we interpret this experiment as consistent with the predicted sign and ordering over the tested strengths, reading the reduced recurrence as a local qualitative model rather than a globally accurate predictor of the edited equilibrium magnitude.

Accurate rank-1 *damping* (full batch) instead lets sharpness grow *past* $2/\eta$ (261.8 ± 5.4) while reaching a *lower* final loss than GD (0.090 vs. 0.111): in this testbed, lower loss does not require moving below the EoS threshold.

4.4 Robustness of the full-batch account

Several controls probe alternative explanations for the full-batch response: proxy-driven norm caps do not move λ_{raw}; inaccurate update-direction sensing leaves the original EoS cap intact, whereas accurate top-direction sensing produces the upward response; orthogonal damping and short sensing lags preserve the qualitative distinction; and a constructed separable multi-sharp control shows that spectral migration is not required to obtain the observed response in that controlled example; it is not a general multi-eigenvalue extension of the reduced theory (Appendix C). Section 5 shows that the same sign-based response does not extend to the canonical stochastic regime.

5 Stochastic deep networks: limits of the sign-based account

5.1 No resolved sharpness reduction from the tested low-rank rescalings

We evaluate the canonical operating point using two complementary diagnostics (trajectories: Fig. 13; per-arm seed counts are given in the figure captions). Our primary reported batch-curvature diagnostic is the *batch sharpness* λ_{step} (defined in §3), evaluated as our compute-limited Monte-Carlo estimate

(m=4 minibatches) along all arms, including SAM. Its estimator freezes BatchNorm running statistics in eval mode for reproducible HVP measurement (measurement protocol: Appendix D.2): we use the EoSS theory, which constrains this quantity for SGD(-momentum) dynamics, to motivate the measurement, not as an assertion that the measured scalar is exactly the train-mode stability variable of every optimizer arm; for SGD it is centered near the momentum-free boundary (late-window distribution: Appendix D). The stochastic rescaling controller retains the target $\lambda_{\mathrm{targ}}=2/\eta$ because it defines the intervention inherited from the full-batch construction; under $\mu=0.9$ this target is an intervention convention, not a claimed unique or exact stochastic stability boundary: the momentum plateaus $2(1\mp\mu)/\eta$ make the reference non-universal under momentum (Appendix D). We reserve *attenuation* or *damping* for an edit whose realized directionwise scale is in $(0, 1)$, use *raw-divisor signed rescaling* when the no-floor rule may reverse sign below target, and use *amplification* for the fixed-sign $\mathbf{g} + \alpha(\mathbf{g}^\top v)v$ controls. Across the central arms the ordering is (Fig. 2; instrumented reruns on n=3 shared seeds): SAM resolves below SGD in λ_{step} (paired $\Delta_{\mathrm{SAM-SGD}}$ -4.30, 95% CI $[-6.70, -1.89]$); rank-1 damping with fixed-batch ARPACK sensing shows no statistically resolved difference from SGD; every gradient-side amplifying action along positive-curvature directions has an above-SGD point estimate; the same qualitative ordering holds for λ_{ref} (full estimates in Appendix D). The damping null is not an engagement failure: ARPACK-based sensing verifies persistent engagement of the edited directions on every substantive step (an early power-iteration sensor defect and its diagnosis are recorded in Appendix D). A rank-1 attenuation control with periodically refreshed current-minibatch-sourced sensing (the most recently sensed direction is reused between N=20-step ARPACK refreshes on training minibatches) likewise leaves the outcome in the SGD band, so the null is not specific to a single fixed reference eigenspace, though this refresh cadence does not test per-step same-minibatch oracle sensing (engagement diagnostics: Appendix D).

5.2 Top-eigenspace edits show weak per-step alignment at the canonical operating point

How strongly do the edits actuate the update? Under bs-128 noise the instantaneous training gradient aligns with *any* top-eigenvector estimate at $|\cos| \approx 0.008$ (median, late window), so the edited subspace's share of the per-step gradient energy is minute: even the top-20 eigenspace carries $< 0.3\%$ under the live in-run convention (per-k shares: Table 9). A separate eval-mode checkpoint convention against a fixed eigenspace gives substantially larger shares (top-100 $\approx$24%); the two quantities condition on different batches, model modes, and eigenspaces and must not be compared as estimates of the same random variable, and the effect of k>20 edits remains untested (Appendix D.4).

The energy share diagnoses the immediate update perturbation, not the long-horizon null itself: across the batch-size sweep the share stays near 10^{-4} over $B \in \{128, \ldots, 2048\}$ (Table 8), yet the same edit's paired outcome shifts from unresolved at B=128 to a statistically resolved $+44.0$ at the quasi-deterministic end: weak per-step actuation can coexist with both null and non-null outcomes (full sweep: Appendix D).

A complementary coordinate is the realized gradient-side edit magnitude $\|\mathbf{g}' - \mathbf{g}\|/\|\mathbf{g}\|$ (computed exactly from the plant logs): the widest rescaling tested, rank 20, actuates about 6% per fired step, substantially smaller than the measured linSAM and SAM corrections, yet magnitude alone does not predict the outcome, since SAM produces a statistically resolved downward shift in λ_{step} whereas linSAM has an upward point estimate; per-step actuation therefore remains a diagnostic rather than a sufficient explanation of the response (every k, the angles, and the momentum-buffer caveat: Appendix D, Table 10).

The wider-rank experiment extends the number of rescaled Hessian directions under the same raw-divisor rule; it is not a pure-rank ablation of unconditional attenuation, and its outcome reflects the full rank-dependent rescaling rule (engagement changes with k; per-direction firing fractions: Table 9) rather than an isolated causal effect of rank. For direction i the implementation uses $d_i = \hat{\lambda}_i/\lambda_{\mathrm{targ}}$ without a $\max(d_i, 1)$ floor and $f_i = \mathrm{clip}(1 - 1/d_i, -1, 1)$, so directions with $d_i > 1$ are attenuated whereas directions with $d_i < 1$ receive bounded sign-reversed amplification; we therefore refer to the k>1 arms as raw-divisor signed eigenspace rescalings rather than pure rank-k damping. The corrected rank-1 arm is different in practice: its sensed divisor remains above one over the substantive engaged window (Appendix D.14), so that arm is an attenuation test. Across $k \leq 20$, none of the raw-divisor rescaling arms has a statistically resolved downward λ_{ref} contrast; among them, only k=10 satisfies the prespecified equivalence criterion at this sample size and margin (TOST, α=.05; Fig. 19, Table 9).

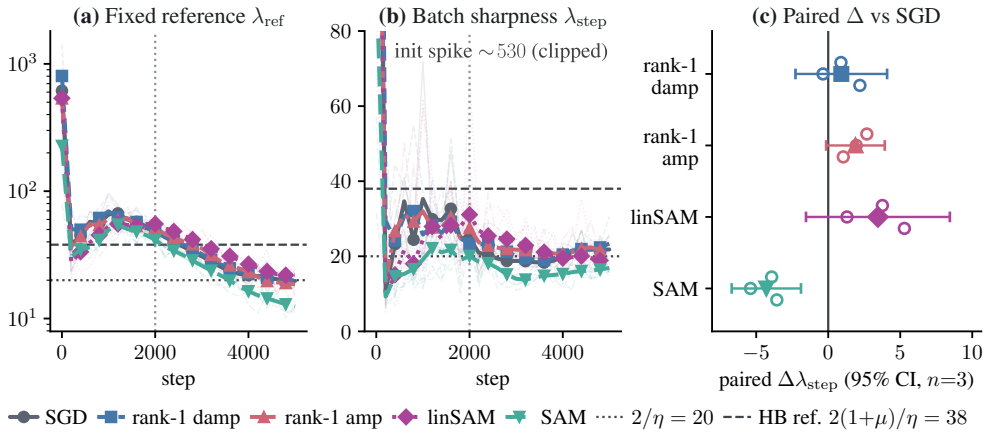

Figure 2: Batch sharpness across optimizer arms (deterministically replayed mini-batch protocol, 5000 steps; per-seed trajectories thin, seed mean bold; late window ≥ 2000; n per arm: SGD/rank-1 damping/amplification n=3, SAM/linSAM n=5). **(a)** Fixed-reference λ_{ref}; **(b)** the EoSS batch sharpness $\lambda_{\text{step}} = \mathbb{E}_B[\mathbf{g}_B^\top H_B \mathbf{g}_B / \|\mathbf{g}_B\|^2]$; **(c)** late-window paired differences $\Delta = \text{arm} - \text{SGD}$ on shared seeds (open markers: per seed; marker+bar: mean and 95% CI). SAM sits below the SGD level (paired $\Delta_{\text{SAM}-\text{SGD}}$ -4.30, 95% CI $[-6.70, -1.89]$), rank-1 damping shows no statistically resolved difference from SGD, and every gradient-side amplifying action along positive-curvature directions has a positive point estimate. The $2/\eta$ and $2(1+\mu)/\eta$ lines are reference lines only; optimizer identity is encoded redundantly by hue, line style, and marker.

Rank-k *amplification* never reproduces SAM's shift toward lower sharpness: the trend is toward higher sharpness, with the k=20 increase statistically resolved under the 95% paired-CI criterion. Directly measured outcome spectra show little systematic spectral migration around the edit (Appendix D), and the regime sweep of Table 8 is treated only as an exploratory operating-point map, not a controlled causal trajectory from stochastic to deterministic dynamics: the alignment measurements are diagnostic rather than causal, and weak alignment is not claimed as the mechanism of the long-horizon null.

5.3 Interpolating between linSAM and SAM: the finite re-query residual bundle carries the ResNet transition

SAM serves here only as a contrasting sharpness-reduction control: it reduces sharpness where the attenuating edits above do not, and its update expands to first order as $\mathbf{g} + (\rho/\|\mathbf{g}\|)H\mathbf{g}$, which amplifies positive-curvature components (and attenuates negative-curvature ones). Because $H\mathbf{g}$ reweights the full spectrum with sign determined jointly by Hessian eigenvalues and gradient components (including negative-curvature directions), this correction is not the high-rank limit of the attenuation experiments: it shows only that this specific $+H\mathbf{g}$ edit is insufficient to reproduce SAM's reduction under this protocol, not that high-rank attenuation as such fails. The linearized arm (ρ=0.05 matched, one HVP per step, no perturbed-point re-query) lands above SGD by a statistically resolved paired contrast and $\sim$12 above SAM (Fig. 16). The SAM–linSAM ordering persists across the tested radii $\rho \in [0.005, 0.1]$, and the re-query's nonlinear Taylor residual remains comparable in magnitude to the first-order correction throughout (Fig. 17); linSAM is therefore an operational first-order comparator, not evidence that SAM is well approximated by its first-order Taylor expansion over the tested radii. BN-protocol controls with bit-identical buffer trajectories show that matching the BN-state evolution alone is insufficient to reproduce SAM's level (Appendix D); these experiments form a sequence of increasingly different intervention families, not a magnitude-matched nested sufficiency test, and the first two do not reproduce SAM's shift at this operating point.

The opposite-sign correction $\mathbf{g} - (\rho/\|\mathbf{g}\|)H_B\mathbf{g}$, with the same nominal $\rho/\|\mathbf{g}\|$ scaling, likewise shows no resolved reduction relative to SGD and leaves batch sharpness at the SGD level (realized-magnitude caveat and full statistics: Appendix D.11); thus none of the tested low-rank rescalings or full-rank $\pm H\mathbf{g}$ corrections reproduces SAM's downward shift in this protocol.

We next interpolate directly between the BN-matched linSAM and SAM, $g_\tau = g_{\text{lin}} + \tau\,(g_{\text{SAM}} - g_{\text{lin}})$, while holding the BatchNorm protocol and RNG stream fixed. Across the four sampled τ values, both reported sharpness means decrease; this associates the endpoint separation with increasing weight on the finite re-query residual bundle along this path, without isolating its nonlinear constituents (per-τ values: Appendix Fig. 18), a pathwise decomposition rather than a universal mechanistic account of SAM. The directly logged pre-update norm varies by only 2.7% over $\tau \in \{0, 0.25, 0.5\}$ (cross-arm-validated at $\tau{=}1$; Appendix D.11), so the norm diagnostic is not used to causally isolate the residual mechanism.

5.4 External controls

A LayerNorm control. The ViT experiment is a cross-system comparison rather than an architecture-isolated ablation: it tests only whether the ResNet decomposition is universal, not attribution of the difference to any one factor. On ViT-tiny at a matched high-curvature operating point ($n{=}5$), SAM separates below SGD, while the linearized control retains only a small resolved difference from SAM yet satisfies the prespecified equivalence test at this sample size and margin: unlike in the BatchNorm ResNet, the finite re-query is not required to match SAM within that margin (full statistics and SAM's $\sim$2–3 accuracy-point cost: Appendix D). Under the 200-epoch cosine recipe, SAM remains resolved below SGD, rank-1 attenuation remains unresolved from SGD, and final accuracy differences are unresolved (Appendix D.8).

6 Discussion and limitations

Observing curvature does not imply controlling the dynamics that generate it: an intervention can change the stability geometry being observed, or elicit no resolved long-horizon response in the stochastic regime. We study sharpness as a dynamical and stability variable rather than a parameterization-invariant generalization metric (§3), and every conclusion concerns the stated parameterization and measurement protocol. In the stochastic protocol, the tested eigenspace rescalings and $\pm H\mathbf{g}$ first-order corrections do not reproduce SAM's resolved sharpness reduction; along the tested interpolation, the endpoint difference is carried by the finite re-query residual (Table 2). No single sign rule holds across regimes, and we make no optimization-performance claim from the observed sharpness separation (§5.4).

Limitations. Lemma 1 concerns only the location of a specified orbit in the reduced map; attraction and survival are not explained, and the full-batch evidence comes from one synthetic testbed. In the stochastic experiments, λ_{ref} is fixed-batch/eval-mode and λ_{step} uses frozen BatchNorm statistics; $2/\eta$ is a controller convention rather than a claimed momentum boundary. The $k{>}1$ no-floor rule is a signed raw-divisor rescaling rather than unconditional damping, and the current-minibatch-sourced sensor is refreshed every $N{=}20$ steps rather than per step. Several supporting cells use $n{=}3$, so unresolved differences are not equivalence. The batch-size sweep is exploratory, and the linSAM–SAM interpolation identifies only a residual bundle along one tested path. Extended scope and measurement limitations are in Appendix G.

7 Conclusion

Directional curvature interventions can change the dynamics that generate the curvature being measured, but the resulting response is intervention- and regime-specific. In the full-batch MLP, trajectories that exhibit the corresponding sign-flip regime track the reduced-map orbit-location relation $2/\eta \to 2d/\eta$, while attraction and survival remain outside the theory. In stochastic ResNet training, the engaged rank-1 attenuation arm and the wider raw-divisor rescalings through $k{=}20$ show no statistically resolved reduction in the reported sharpness diagnostics, and the wider rule is not unconditional pure damping. SAM provides a contrasting response, but the linSAM–SAM experiment identifies only a pathwise residual bundle: increasing its weight accompanies lower ResNet sharpness, whereas the ViT control does not require the same decomposition within the prespecified equivalence margin. Measuring a sharp direction is therefore not sufficient to predict how a particular directional intervention will change raw sharpness. Such predictions must account for the optimizer, measurement convention, sensing protocol, and the dynamics that generate the curvature signal.

AI USE STATEMENT

Generative AI tools were used in three capacities, under author direction and verification. (i) *Writing*: review of the manuscript's language and expression, with suggestions for revision; all modifications were made by the authors, who take full responsibility for the submission. (ii) *Literature search*: assistance in locating and summarizing related work; all cited references were consulted directly by the authors in their original publicly available form. (iii) *Research execution*: AI coding assistants were used for code assistance during implementation and in summarizing experimental outputs; AI tools provided suggestions during discussions of experimental design; all training runs, experimental decisions, statistical tests, and analyses were performed and verified by the authors, and all reported numbers derive from the authors' run manifests. Generative AI was not used to formulate mathematical claims, provide proof ingredients, write proofs, generate datasets, or interpret experimental results; no other required-disclosure tasks apply.

REPRODUCIBILITY STATEMENT

All headline comparisons use deterministic paired seeds. A pointer map to the supplement: Lemma assumptions and proofs → Appendix A; fixed-divisor configurations and seeds → Appendix C and the run manifests; stochastic estimators and ARPACK tolerances → Appendix D; complete hyperparameters and per-run compute → Appendix F; code entry points → supplementary README.

REFERENCES

Atish Agarwala and Yann N. Dauphin. SAM operates far from home: eigenvalue regularization as a dynamical phenomenon. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 152–168. PMLR, 2023.

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 169–195. PMLR, 2023.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability". In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for SGD, 2024. URL https://arxiv.org/abs/2412.20553. arXiv:2412.20553v5.

Arseniy Andreyev, Advikar Ananthkumar, Marc Walden, Tomaso Poggio, and Pierfrancesco Beneventano. Momentum further constrains sharpness at the edge of stochastic stability, 2026. URL https://arxiv.org/abs/2604.14108.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 639–668. PMLR, 2022.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024. PMLR, 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391. PMLR, 2023.

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. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability, 2022. URL https://arxiv.org/abs/2207.14484. arXiv preprint arXiv:2207.14484.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=sIE2rI3ZPs.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=gcOZC-dA9x.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Y. F. Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2205.14083.

Hongru Duan, Yongle Chen, and Lei Guan. X-SAM: Boosting sharpness-aware minimization with dominant-eigenvector gradient correction, 2026. URL https://arxiv.org/abs/2601.10251.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert Gower. Non-euclidean gradient descent operates at the edge of stability. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.05002.

Dayal Singh Kalra and Maissam Barkeshli. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/a23598416361c7a9860164155e6ddd0b-Abstract-Conference.html.

Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=VZN0irKnl0.

Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H. Miller, and Michael Shvartsman. A scalable measure of loss landscape curvature for analyzing the training dynamics of LLMs, 2026. URL https://arxiv.org/abs/2601.16979.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.04836.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 2021. URL http://proceedings.mlr.press/v139/kwon21b.html.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism, 2020. URL https://arxiv.org/abs/2003.02218. arXiv preprint arXiv:2003.02218.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/40bb79c081828bebdc39d65a82367246-Abstract-Conference.html.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024. URL https://jmlr.org/papers/v25/23-1285.html.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dffd1c523512e557f4e75e8309049213-Abstract-Conference.html.

Sameera Ramasinghe, Ajanthan Thalaiyasingam, Hadi Mohaghegh Dolatabadi, Chamin Hewa Koneputugodage, Gil Avraham, Violetta Shevchenko, Yan Zuo, Karol Pajak, and Alexander Long. Taming curvature: Architecture warm-up for stable transformer training. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=DuNf2vPTTK.

Vincent Roulet, Atish Agarwala, Jean-Bastien Grill, Grzegorz Swirszcz, Mathieu Blondel, and Fabian Pedregosa. Stepping on the edge: Curvature aware learning rate tuners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=SEflLHIhhJ.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e2a9256bd816ab9e082dfaa22f1f62a2-Abstract-Conference.html.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=v6iLQBoIJw.

Mingwei Wei, James Stokes, and David J. Schwab. Mean-field analysis of batch normalization, 2019. URL https://arxiv.org/abs/1903.02606. arXiv preprint arXiv:1903.02606.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? In *International Conference on Learning Representations (ICLR)*, 2023.

Note to Table 2: Legacy labels in Table 2 follow the original run naming: for $k>1$, "damping" denotes the signed raw-divisor rescaling of §5.2; "required" denotes only the pathwise residual-bundle statement, not an isolated mechanism.

Table 2: Claims, formal status, and tested scope.

Claim	Formal status	Empirical status / tested scope
Boundary shift $2/\eta \to 2d/\eta$	Orbit location in the 2-D reduced map (Lemma 1); not an attraction theorem	On-orbit fixed-divisor cells track $2d/\eta$; two surviving cells remain off the predicted orbit (§4.2, Fig. 1); full-batch testbed
Low-rank top-eigenspace rescaling (damping, $k \leq 20$)	—	No resolved λ_{ref} reduction through $k \leq 20$; no resolved λ_{step} reduction for fixed-batch-sensing rank-1 damping; $s_{20} < 0.3\%$ energy share; bs-128 ResNet-18/CIFAR-100
Full-rank first-order curvature corrections ($\pm H\mathbf{g}$)	—	Neither correction reproduces SAM's reduction ($+H\mathbf{g}$ raises λ_{ref} above SGD, $+5.27$, 95% CI $[+1.94, +8.61]$; $-H\mathbf{g}$ shows no statistically resolved reduction, $+2.07$, 95% CI $[-1.55, +5.70]$; §5.3)
Re-query residual required for the transition	—	Along the tested interpolation, the transition toward SAM's lower-sharpness regime cannot be reproduced without the SAM–linSAM residual; BN ResNet-18; not universal in the tested ViT

A Proof and extensions of Lemma 1: the 2-D self-stabilization reduction

This appendix develops the reduced-model account behind the full-batch experiments. Within the two-dimensional self-stabilization reduction, damping the top-gradient component does not create a sub-threshold sign-flip orbit; reversing the sign moves the corresponding orbit below $2/\eta$. Thus the relevant distinction in this reduction is attenuation versus amplification rather than local versus global action. The formal result concerns orbit location, while the accompanying experiments test several alternative explanations for the observed network trajectories, including spectral migration and sensing failure. Section A.1 states the account and its direct tests, Section A.2 derives the reduced recurrence, Section A.3 states the formal orbit-location lemma and its stability limitations, and the remaining controls (§A.4) delimit which parts of the account extend beyond the reduction.

A.1 The self-stabilization cap, and why rank-1 damping has no sub-threshold sign-flip orbit

Claim 1 (reduced-model interpretation; informal)**.** *The EoS threshold $2/\eta$ is a self-stabilization* cap*: the period-2 oscillation of the iterate along the top eigenvector u_1 generates, via the cubic identity $\nabla S = \nabla^3 L(u_1, u_1)$ of Damian et al. (2023), a restoring force that balances progressive sharpening and pins $S = \lambda_{\max}$ at $2/\eta$. In this reduction, a rank-1 gradient edit that* damps *the gradient component along an estimate $\hat{v}$ of u_1 creates no sub-threshold sign-flip orbit; empirically it lands in one of two regimes, neither of which descends below the threshold:*

- Imperfect damping *($\hat{v}$ lags u_1, e.g. the free $\hat{v} = \Delta\theta_{t-1}/\|\cdot\|$): the parallel-gradient cancellation is incomplete, the period-2 oscillation persists, and self-stabilization caps $\lambda_{\max}$ at $2/\eta$.*
- Accurate damping *($\hat{v} \approx u_1$ via power iteration): the directional rescaling is applied to the correct sharp direction. At the original boundary $\lambda = 2/\eta$, the effective top-direction step lies inside its previous sign-flip boundary, weakening the oscillatory balance that supported the original cap. For fixed d, progressive sharpening can then continue until the sign-flip condition is re-established at $\lambda = 2d/\eta$; under adaptive self-gating, the reduced model does not determine a finite raw-sharpness endpoint. Thus accurate damping releases the original $2/\eta$ cap rather than eliminating sign-flip dynamics altogether.*

Within the reduced sign-flip account, a sub-threshold orbit requires the amplifying sign. SAM provides a separate global benchmark whose first-order correction has this sign along positive-curvature directions; the "spectral migration" alternative, that damping the current top direction causes another eigendirection to become dominant, is not *what the experiments show.*

This matches *all* empirical controls, including the following controls, several designed specifically to probe migration: (a) the free-$\hat{v}$ optimizer locks at $\lambda_{\max}^{\text{final}} = 40.12 \pm 0.12 = 2/\eta$ (regime (i): the lagged $\hat{v}$ preserves the oscillation, so the cap holds); (b) the "re-acquire the correct v_{top}" variant does *not* test accurate v_{top} every step: the refreshed vector is overwritten by the next step's free $\Delta\theta$ at the end of each step, so it reduces to variant (a) save one step. Its 40.019 ± 0.002 therefore supports regime (i). The accurate-v_{top} test is the rank-k sweep below; (c) the free-$\hat{v}$ diagnostic $\hat{\lambda} = \hat{v}^\top H \hat{v}$ *does* collapse to $\sim 10^{-4}$ while true $\lambda_{\max}$ stays at 40, but this is a *$\hat{v}$-tracking failure*, not a migration signature: the trajectory's only sharp axis is u_1, and once the lagged $\hat{v}$ rotates out of it (as the active window changes the update direction) $\hat{\lambda}$ measures a subdominant axis: a measurement artifact, not sharpness flow; (d) SAM escapes ($21.35 \pm 0.04 < 2/\eta$) because its perturbation $\varepsilon = \rho \mathbf{g}/\|\mathbf{g}\|$ acts in *every* direction simultaneously, so there is no eigenvector whose cap is selectively removed; the matched-step X-SAM projector ablation (Appendix E, n=10) shows the rank-1 half of an eigenvector-aware SAM fires every step yet has no detectable effect on $\lambda_{\max}$ at this power (paired $\Delta = +0.62$, 95% CI $[-0.61, +1.85]$; reductions beyond $\sim$0.61 reference-sharpness units are excluded).

The rank-k sweep is inconsistent with migration on the 1-D-sharp landscape. Replacing the free $\hat{v}$ by *accurate* warm-started power iteration of the top-k eigenvectors (block power iteration + modified Gram–Schmidt) and damping all k components, the optimizer enters regime (ii): with $k=1$ the edit accurately rescales the only sharp direction, releasing the original $2/\eta$ cap; raw sharpness then escapes upward to 261.8 ± 5.4 (lower loss 0.090 than GD's 0.111, but *sharper*, opposite to the control objective). Crucially $k = 1, 2, 5, 10$ are bit-identical: the spectrum is 1-D sharp ($\lambda_1 = 261.8 \pm 5.4$, $\lambda_2 \approx 0.6$; over two orders of magnitude), so there is *nowhere for sharpness to migrate to*; damping more directions changes nothing. The k-invariance is predicted by the cap-removal mechanism (one accurately damped direction suffices to release the cap on the only oscillating direction) and is inconsistent with migration (which predicts more damping $\Rightarrow$ lower $\lambda_{\max}$).

ResNet-18 rank-k is an outcome-level null, not an engagement failure. One might worry the MLP k-invariance above merely confirms the 1-D-sharp geometry rather than telling against migration. The ResNet-18 rank-k sweep (heavy-tailed spectrum, not 1-D sharp) addresses this worry only partially: as established in §5.1, with deterministic ARPACK-based sensing the edits fire as designed (verified engagement at every tested $k \leq 20$), yet no damping arm produces a statistically resolved reduction of the reference $\lambda_{\max}$ relative to SGD at this power: the null is at the outcome level, not the engagement level. (An earlier power-iteration false-convergence had initially left the clamped optimizer effectively dormant; after the sensing correction, the absence of a statistically resolved reduction persists at this operating point, so the earlier dormancy does not explain the observed null.) The sweep therefore contributes the scale reproduction of the *ordering* (no descent below the SGD band at any attenuating k; only full SAM escapes), not an independent test of migration; the evidence against migration rests on the MLP multi-sharp test below.

The multi-sharp-landscape test is inconsistent with migration where it could plausibly act. As established in §4.4: on a constructed multi-sharp landscape (two directions pinned at $2/\eta$), damping only u_1 lets u_1 escape upward while λ_2 stays pinned (there is no flow from u_1 to u_2) and, unlike the 1-D case, k *does* matter, each direction whose cap is removed escaping independently and revealing the *per-direction* structure of the cap. The multi-sharp construction is therefore used as a stress test of the per-direction account; we do not infer the typical dimensionality of the sharp Hessian subspace from trajectory-PCA dimensionality. Across both landscapes the verdict is identical: no descent below the threshold, and no migration.

The non-separable multi-sharp experiment provides evidence consistent with per-direction capping in one constructed landscape; it does not establish general multi-eigenvalue decoupling: the regime Damian et al. (2023) flag as open (§8.2).

A.2 Supporting heuristic: the damped self-stabilization dynamics

The cap-removal account is the dynamics-level avatar of Damian et al. (2023). Let $S(\theta) = \lambda_{\max}(H)$ with simple top pair (λ_1, u_1), x_t the iterate's displacement along u_1, $y_t = S - 2/\eta$, $\alpha_{\text{ps}} = -\nabla L \cdot \nabla S > 0$ the progressive-sharpening coefficient (their Def. 2), and $\beta = \|\nabla S\|^2$. Their keystone

identity $\nabla S = \nabla^3 L(u_1, u_1)$ (their Lemma 2) yields the 2-D recurrence (their eq. 3)

$$x_{t+1} = -(1 + \eta y_t)\, x_t, \qquad y_{t+1} = y_t + \eta\Big(\alpha_{\mathrm{ps}} - \tfrac{\beta}{2} x_t^2\Big), \tag{2}$$

whose period-2 fixed point $|x_t| \to \delta = \sqrt{2\alpha_{\mathrm{ps}}/\beta}$ pins $S = \lambda_1 \approx 2/\eta$ by balancing progressive sharpening α_{ps} against the cubic restoring force $\frac{\beta}{2}x^2$ generated by the iterate moving along u_1. This is the self-stabilization *cap* of Claim 1: the oscillation x_t is what holds $\lambda_{\max}$ at $2/\eta$.

Effect of rank-1 damping (heuristic). The optimizer edits the gradient to $g' = g - \gamma(g^\top \hat{v})\hat{v}$ with $\gamma = 1 - 1/d \in (0, 1)$, $d = \max(\hat{\lambda}/(2/\eta), 1)$. Taking $\hat{v} \approx u_1$ and treating d as fixed over a local window, the displacement along u_1 evolves as

$$x_{t+1} = \big(1 - (1-\gamma)\eta\lambda_1\big)x_t = \big(-1 + 2\gamma - (1-\gamma)\eta y_t\big)x_t, \tag{3}$$

while the restoring-force recurrence in y retains the form $y_{t+1} = y_t + \eta(\alpha_{\mathrm{ps}} - \frac{\beta}{2}x_t^2)$: the edited optimizer's first-order sharpness increment is $-\eta\langle \nabla S, g'\rangle = \eta\alpha_{\mathrm{ps}} + \eta\gamma(g^\top u_1)(u_1^\top \nabla S)$, and expanding $g^\top u_1 = \lambda_1 x + \frac{1}{2}(u_1^\top \nabla S)x^2 + O(x^3)$ (the same along-axis truncation that yields (3)) decomposes the correction into an x-odd ripple of zero mean over the sign-flip cycle (the class of terms the *undamped* reduction of Damian et al. (2023) already discards in freezing α_{ps} to a constant) and an x-even remnant $\frac{\gamma}{2}(u_1^\top \nabla S)^2 x^2$ that only renormalizes the amplitude-setting coefficient β. Since the orbit *location* y^* follows from the x-recurrence alone, neither piece moves $\lambda^* = 2d/\eta$. Solving (3) for the period-2 fixed point ($x_{t+1} = -x_t$) gives an equilibrium excess

$$y^* = \tfrac{2\gamma}{(1-\gamma)\eta}, \qquad \tfrac{\partial y^*}{\partial \gamma} = \tfrac{2}{(1-\gamma)^2\eta} > 0 \ \text{ for all } \gamma < 1 : \tag{4}$$

the sign-flip orbit location is *strictly increasing in the sign of the edit*. Damping ($\gamma \in (0, 1)$) gives $y^* > 0$; the sign-flipped *amplification* edit ($\gamma = -\alpha < 0$, i.e. $g' = g + \alpha(g^\top \hat{v})\hat{v}$) gives $y^* = -\frac{2\alpha}{(1+\alpha)\eta} < 0$ with equilibrium sharpness $\lambda^* = 2/((1+\alpha)\eta) < 2/\eta$: *descent below the threshold by a rank-1 single-direction edit* (Corollary 3). The reduction thus does not support a local-vs-global dichotomy; it supports an *attenuation-vs-amplification* one, of which SAM's global perturbation is one first-order instance of the amplifying sign (Remark 6).

What the heuristic predicts, and what it does not. The content to keep is the *sign* of the prediction, not a magnitude: $y^* > 0$ says accurate damping moves the sign-flip orbit location *past* $2/\eta$ (upward, opposite to the control goal), and this is what the accurate-v_{top} experiment confirms (an upward escape on the 1-D landscape, ~ 65 on the multi-sharp one). Symmetrically, $y^* < 0$ at $\gamma < 0$ says amplification moves the sign-flip orbit location *below* $2/\eta$, and this is what the sign-flip control confirms (§4.3): late-window equilibria 30.1 ± 0.5 / 22.5 ± 1.3 / 14.3 ± 2.7 at $\alpha = 0.25$ / 0.5 / 1 (n=5), all strictly below $2/\eta = 40$ and monotone in α, though each sits *below* the naive fixed-point value $\lambda^* = 2/((1+\alpha)\eta)$ (32 / 26.7 / 20); the magnitude disclaimer below applies symmetrically to both signs. Neither sign predicts the escape magnitude: under the self-gating rule $d = \hat{\lambda}/(2/\eta)$, $\gamma = 1 - 1/d$, the fixed-point condition $y^* = 2\gamma/((1-\gamma)\eta)$ is satisfied for *every* $\lambda_1 > 2/\eta$ (substituting γ reduces it to the identity $y^* = y$); the magnitude is set by where unopposed progressive sharpening equilibrates, a quantity the 2-D reduction does not fix. The free-lagged-$\hat{v}$ variant, by contrast, falls *outside* this derivation: once the active window changes the update direction, $\hat{v} \not\approx u_1$, so $\hat{\lambda} = \hat{v}^\top H \hat{v}$ reads a subdominant axis and the self-gating d is set by the wrong curvature; the u_1-oscillation that (3) targets is essentially untouched, the cap holds, and $\lambda_{\max} \approx 2/\eta$. The two regimes of Claim 1 are thus two cases of *whether the derivation's $\hat{v} \approx u_1$ premise holds*: accurate $\hat{v}$ enters the $y^* > 0$ regime (cap removed, upward escape); lagged $\hat{v}$ never engages the u_1 axis and the cap is never at risk.

Scope of the reduced-model interpretation. Claim 1 is a reduced-model interpretation, not a trajectory theorem. Its multi-direction premise is empirically supported in the constructed multi-sharp test but remains theoretically open in general (**G1**). The reduced-dynamics heuristic (2)–(3) postulates per-direction decoupling, which is the multi-eigenvalue extension Damian et al. (2023) explicitly defer to future work (§8.2), and is formally singular at the rotation instant $\lambda_1 = \lambda_{j^*}$ (their Lemma 2 needs $\lambda_{\max}$ simple); there one works with the individual gradients $\nabla\lambda_i$ rather than ∇S. We tested the "spectral migration" alternative (sharpness flowing to an undamped direction) directly on a multi-sharp

landscape and it did not manifest (λ_2 stayed pinned while u_1 escaped); we therefore present migration only as a hypothesis that these tests do not support, and the cap-removal account as the one the data support. The core algebraic content of Claim 1 (that no period-2 fixed point of the damped recurrence has $y \leq 0$) is made rigorous, with a stability analysis, in Lemma 1 below; the gaps that remain before a full trajectory theorem are itemized there (**G1**–**G5**). The division of labor is deliberate and should be read as such: Lemma 1 is a *reduced-map statement* (it constrains sign-flip period-2 orbits of the 2-D recurrence under **A1**–**A3**, not trajectories of the network) and every mechanistic statement that reaches beyond that reduction (optimizers with sensed $\hat{v}$, non-sign-flip transients, the stochastic regime) is carried by the interventional experiments of §4–§5.1, not by the lemma.

A.3 Formal lemma: no flat sign-flip orbit under rank-1 damping (2-D reduction)

We upgrade the heuristic above to a formal statement on Damian et al.'s 2-D recurrence. The addition over (3)–(2) is a *conditional y-contraction* analysis: not only is the period-2 fixed point at $y^* > 0$, the two-step y-direction multiplier (conditional on x remaining on the period-2 orbit) is contractive under a checkable condition. We label this *conditional y-contraction*, not stability: the full 2-D map is normally repelling (Remark 4), and orbit stability would additionally require the higher-order $|x|$-bounding of gap **G4**.

Setup. Recall $S(\theta) = \lambda_{\max}(H)$, $y = S - 2/\eta$, the top pair (λ_1, u_1), x_t the iterate's displacement along u_1, $\alpha_{\mathrm{ps}} = -\nabla L \cdot \nabla S > 0$ and $\beta = \|\nabla S\|^2 > 0$. The y-recurrence is

$$y_{t+1} = y_t + \eta\Big(\alpha_{\mathrm{ps}} - \tfrac{\beta}{2} x_t^2\Big), \tag{5}$$

and the rank-1-damped x-recurrence (gradient edit $g \mapsto g - \gamma(g^\top \hat{v})\hat{v}$, $\gamma = 1 - 1/d \in (0,1)$, $\hat{v} = u_1$) is

$$x_{t+1} = \big(-1 + 2\gamma - (1-\gamma)\eta y_t\big) x_t. \tag{6}$$

Let F be the 2-D map defined by (6)–(5).

Assumption 1 (**A1**: positive isolated top pair)**.** Along the analysis window, $\lambda_1(\theta) \geq m > 0$ and $\lambda_1(\theta) - \lambda_2(\theta) \geq \Delta > 0$, so the top eigendirection and ∇S are well defined (Damian et al., 2023).

Assumption 2 (**A2**: window-fixed edit strength)**.** The edit strength is constant over the analysis window: damping $\gamma \in (0,1)$ (it evolves on $O(1/\eta)$ steps; x-dynamics is $O(1)$); the sign-flipped amplifying edit $\gamma = -\alpha < 0$ is treated under the same window assumption in Corollary 3.

Assumption 3 (**A3**: accurate sensing)**.** $\hat{v} = u_1$ (an idealization; within the reduced top-direction equation, an angular sensing error φ reduces the effective damping to $\gamma \cos^2 \varphi$, cf. gap **G5**; off-axis effects lie outside the reduction).

Lemma 1, restated in detailed form. Under **A1**–**A3**, for the damped map F with $\alpha_{\mathrm{ps}}, \beta, \eta > 0$, $\gamma \in (0,1)$:

- **(a)** *Algebraic.* Every *sign-flip* period-2 fixed point of F, one with $F(x_*, y_*) = (-x_*, y_*)$ and $|x_*| > 0$, satisfies $y^* = \frac{2\gamma}{(1-\gamma)\eta} > 0$ and $|x_*| = \sqrt{2\alpha_{\mathrm{ps}}/\beta}$. In particular, *no sign-flip period-2 orbit has $y^* \leq 0$.* (The map F^2 admits additional non-sign-flip period-2 orbits for $\kappa > 4$, with y alternating between two negative values; these require $y < -\frac{2}{\eta}$, i.e. $S = \lambda_{\max} < 0$, contradicting **A1** ($\lambda_1 \geq m > 0$). We exclude them on these *physicality* grounds, not on stability grounds: the sign-flip orbit the lemma analyzes is *itself* a full-map repeller (Remark 4), so instability cannot be the criterion that separates the two orbit families.)
- **(b)** *Conditional y-contraction (not stability).* With $\kappa := 2(1-\gamma)\eta^2\alpha_{\mathrm{ps}} > 0$, the two-step y-direction multiplier *conditional on x remaining on the period-2 orbit* is $\mu_y = 1 - \kappa$, satisfying $|\mu_y| < 1$ iff $(1-\gamma)\eta^2\alpha_{\mathrm{ps}} < 1$. This is a *conditional y-contraction*, a contraction of the slow y-direction on the invariant graph $x \in \{\pm x_*\}$, not an orbit-stability result: the full 2-D map is normally repelling (Remark 4), and orbit stability additionally requires the higher-order $|x|$-bounding of gap **G4**, which the 2-D reduction cannot supply.

Proof. *(a)* Impose the period-2 ansatz $x_{t+1} = -x_t$ in (6): $-x_* = (-1 + 2\gamma - (1-\gamma)\eta y^*)x_*$, so $y^* = \frac{2\gamma}{(1-\gamma)\eta} > 0$ (positive since $\gamma \in (0,1)$). Monotonicity $\partial y^*/\partial\gamma = 2/((1-\gamma)^2\eta) > 0$: stronger damping strictly raises the sign-flip orbit location. The y-recurrence (5) on the orbit (y invariant over two steps) forces the increment to vanish, giving $|x_*| = \sqrt{2\alpha_{\mathrm{ps}}/\beta}$, identical to the undamped amplitude $\delta = \sqrt{2\alpha_{\mathrm{ps}}/\beta}$ (Damian et al., 2023): damping relocates y^* but not the orbit amplitude.

(b) Singular-perturbation (slow–fast) analysis: y is slow (per-step drift $O(\eta)$), x is fast (per-step sign-flip). Perturb $y \to y^* + \varepsilon$ keeping x on the fast manifold. Writing $m(y) := -1 + 2\gamma - (1-\gamma)\eta y$ (so $m(y^*) = -1$): step 1 gives $x_1 = (-1 - (1-\gamma)\eta\varepsilon)x_*$ and $y_1 = y^* + \varepsilon$ (the y-increment vanishes by $|x_*| = \sqrt{2\alpha_{\mathrm{ps}}/\beta}$); step 2 then yields, using $x_1^2 = (1 + 2(1-\gamma)\eta\varepsilon)x_*^2 + O(\varepsilon^2)$ and $\beta x_*^2 = 2\alpha_{\mathrm{ps}}$,

$$y_2 = y^* + \varepsilon - \underbrace{2(1-\gamma)\eta^2\alpha_{\mathrm{ps}}}_{=:\,\kappa}\,\varepsilon + O(\varepsilon^2) = y^* + (1-\kappa)\,\varepsilon + O(\varepsilon^2). \tag{7}$$

Thus $\mu_y = 1 - \kappa$, conditionally y-contractive (on the invariant graph $x \in \{\pm x_*\}$) iff $|1-\kappa| < 1 \iff (1-\gamma)\eta^2\alpha_{\mathrm{ps}} < 1$. □

Remark 1 (scope of stability). The orbit above is repelling for the full 2-D Jacobian (Remark 4 below): a complete stability argument would require the higher-order $|x|$-bounding of gap **G4**, which the 2-D reduction cannot supply, and we leave this open. The result is used as an orbit-location prediction only.

Remark 2 (degenerate and unstable fixed points outside scope). F also has a degenerate period-1 fixed point at $y = -2/\eta$ (i.e. $S = 0$, $\lambda_1 = 0$), outside the EoS regime $\lambda_1 \approx 2/\eta$. When $(1-\gamma)\eta^2\alpha_{\mathrm{ps}} \geq 1$ (i.e. $\kappa \geq 2$) the conditional y-multiplier $|\mu_y| \geq 1$ and the equilibrium loses even conditional stability, plausibly the route to the most violent upward escapes under strong damping plus strong progressive sharpening; our experiments (η=0.05, moderate α_{ps}) sit comfortably at $\kappa < 1$.

Remark 3 (regime (i) lies outside the lemma's premise). A3 places the lemma in regime (ii) (accurate sensing). Under regime (i) (lagged or free $\hat{v}$) the sensing direction drifts off u_1, the effective damping becomes $\gamma_{\mathrm{eff}} = \gamma\cos^2\varphi$ (gap **G5**), and in the limit $\hat{v} \perp u_1$ the u_1-oscillation is untouched and the cap holds at $2/\eta$. Regime (i) is therefore not a counterexample to the lemma but the regime where its premise fails to engage: the iterate sits at the undamped equilibrium $y^* = 0$ by default.

Remark 4 (full 2-D Jacobian: the lemma claims only conditional y-contraction). DF^2 at (x_*, y_*) has eigenvalues $(1-\kappa) \pm 2i\sqrt{\kappa}$ of modulus $1 + \kappa > 1$: the full 2-D map is *normally repelling*. We stress that the conditional y-contraction part of the restatement below does *not* claim asymptotic stability of the orbit: it claims only the *conditional* y-direction multiplier $\mu_y = 1 - \kappa$ (contractive under $(1-\gamma)\eta^2\alpha_{\mathrm{ps}} < 1$), computed on the invariant graph $x \in \{\pm x_*\}$ with x slaved. The standard EoS intuition—that the fast x-amplification acts as a restorative "amplitude servo" checked by the y-dynamics (when y overshoots y^*, $|x|$ grows, the cubic restoring term $\beta x^2/2$ exceeds progressive sharpening α_{ps}, and y is driven back down)—is a *conjectured* closing of gap **G4**: bounding $|x_\infty|$ requires the quartic-and-higher landscape terms the 2-D reduction omits, and is exactly what a trajectory theorem would need to supply. We therefore claim only the conditional y-multiplier result (the conditional y-contraction part of the restatement below, the EoS-relevant notion), not full 2-D asymptotic stability. Crucially, our negative claim (no sign-flip period-2 orbit has $y^* \leq 0$, the algebraic part of the restatement below) does not depend on this stability story at all; it is pure algebra on the fixed-point ansatz.

Corollary 2 (upward escape; direction under adaptive γ, point prediction under fixed γ)**.** *At equilibrium $S_* = 2/\eta + y^* = 2d/\eta$ with $d = 1/(1-\gamma)$; $\gamma \uparrow 1$ drives $S_* \to \infty$. The strength of this statement depends on how γ is set. Under the* adaptive *self-gating rule (γ tracking $\hat{\lambda}$), the lemma pins the* direction *($y^* > 0$, upward) and the parametric dependence ($S_* = 2d/\eta$) but* not *the finite value S converges to: with γ tied to $\hat{\lambda}$ the fixed-point condition degenerates to the identity $y^* = y$, and the magnitude is set by where unopposed progressive sharpening equilibrates, beyond the 2-D reduction. Under a* fixed *divisor (γ constant, the fixed-divisor control), $S_* = 2d_{\mathrm{fix}}/\eta$ is the reduction's genuine point prediction: among the seven cells that settle onto the predicted orbit, Fig. 1 gives late-window realized-to-predicted ratios of* 0.95–0.99*, i.e. within* 5% *of $2d_{\mathrm{fix}}/\eta$. The fixed-d period-2 orbit remains only conditionally stable: $d \in \{2.5, 3\}$ diverges on* 5/5 *seeds at η=0.05, while d=6.5 survives* 3/5 *and the adaptive rule is* 5/5 *stable at the same operating point.*

Fig. 3 displays the control-law-side bookkeeping of this orbit location under the adaptive rule.

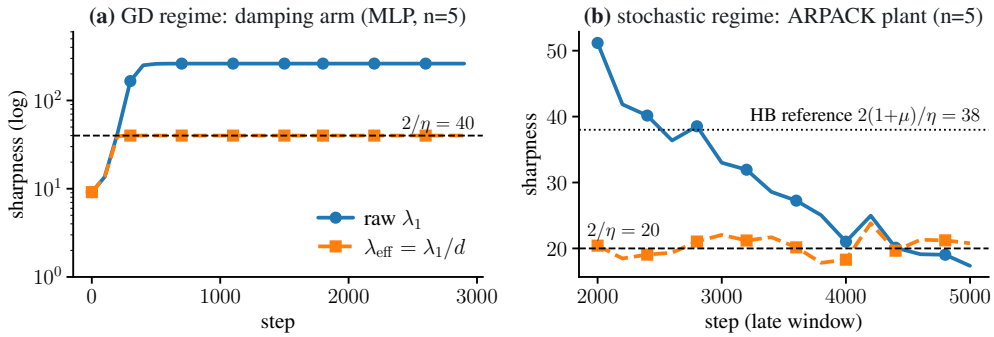

Figure 3: Control-law bookkeeping of the preconditioned sharpness λ_{eff} (Corollary 2) under the *adaptive* self-gating rule. **(a)** GD regime (MLP, accurate rank-1 damping, n=5, seed-mean): the raw sensed $\hat{\lambda}_1$ escapes past $2/\eta$=40 to ~262≈6.6 · $2/\eta$ while $\hat{\lambda}_1/d$ sits at $2/\eta$ throughout the climb (log scale). This panel is *definitional* whenever the clamp fires ($d = \hat{\lambda}_1/(2/\eta)$ by construction): it verifies the control law's internal consistency, and the interventional test of the λ_{eff} claim is the fixed-divisor control of Fig. 1. **(b)** Stochastic regime (ResNet-18/CIFAR-100, accurate ARPACK-sensing damping arm, n=5, late window): raw λ_1 (independent tol-10^{-6} outcome ARPACK; no statistically resolved difference from SGD) while λ_1/d sits at $20.40 \pm 0.65 = (1.02 \pm 0.03) \cdot 2/\eta$, here a measurement-consistency statement (independent outcome vs. stale sensing), not an identity; the heavy-ball reference line at 38 and the SGD raw level 29.2 are shown for contrast.

Corollary 3 (rank-1 amplification descends below the threshold (sign-flip control))**.** *Run the same fixed-point algebra of the algebraic part of the restatement below at the opposite sign of the edit:* $\gamma = -\alpha < 0$ *(i.e.* $g' = g + \alpha(g^\top u_1)u_1$, $\alpha > 0$*), with* ***A1–A3*** *otherwise unchanged. Every sign-flip period-2 fixed point then satisfies* $y^* = \frac{2\gamma}{(1-\gamma)\eta} = -\frac{2\alpha}{(1+\alpha)\eta} < 0$*, i.e.* $S_* = 2d/\eta$ *with* $d = \frac{1}{1+\alpha} \in (0,1)$*:* $\lambda^* = 2/((1+\alpha)\eta) < 2/\eta$, and $S_* \downarrow 0$ monotonically as $\alpha \uparrow \infty$ *(never below* 0*). The conditional* y*-contraction part of the restatement below becomes* $|\mu_y| < 1$ *iff* $(1+\alpha)\eta^2\alpha_{\text{ps}} < 1$, harder *to satisfy as* α *grows, unlike damping, where* $\gamma \uparrow 1$ *makes it easier.* Experimentally confirmed *(§4.3): rank-1 amplification with accurate warm-started sensing descends below the threshold at every tested* α*, with late-window equilibria monotone in* α*, while the same sensing with* $\gamma > 0$ *escapes past* $2/\eta$ *(Lemma 1); the realized equilibria sit below the naive* $\lambda^* = 2/((1+\alpha)\eta)$ *values, consistent with Corollary 2's direction-only disclaimer. (At ResNet-18 scale, rank-1 amplification shows no statistically resolved effect, whereas increasing the amplified rank to* k=20 *and applying the full-rank first-order correction both shift sharpness upward rather than reproducing the full-batch descent (§5.1), so the corollary's downward escape is specific to the GD-regime dynamics.) The lemma's negative statement is thereby* sharpened, not contradicted*: what cannot descend below the threshold is rank-1* damping *(*$\gamma \in (0,1)$*), not rank-1 actuation in general.*

Remark 5 (partial multi-eigenvalue extension (separable case)). If H is separable, $H = \sum_i \lambda_i u_i u_i^\top$ with no off-diagonal coupling, the Damian recurrence decouples into independent (x_i, y_i) pairs and Lemma 1 applies *per direction*: damping any subset of eigenvectors leaves each undamped direction's cap at $y_i^* = 0$ or moves each damped direction to $y_i^* > 0$. The multi-sharp test (§4) provides an empirical check consistent with per-direction independence on one non-separable landscape, the regime Damian et al. (2023) leave theoretically open. A general multi-eigenvalue extension remains open (gap **G1**).

Connection to projection-based precedents. Song et al. (2025) project the SGD update onto the spectral "bulk" (removing the top-k eigenvector components) and report that sharpness grows unconstrained while loss and accuracy are preserved. This is the $\gamma \to 1$ ($d \to \infty$) limit of our rank-1 damping: Corollary 2's equilibrium $S_* = 2d/\eta$ diverges as $d \to \infty$, so their observation is regime (ii) (cap removed) taken to its extreme: the unconstrained growth is the predicted divergence of the orbit location itself, not a separate phenomenon. Our partial-damping ($\gamma < 1$) experiments diagnose the cap at finite S_*; their full-projection ($\gamma = 1$) removes it entirely. Their "pseudo-alignment" reading and our "cap-removal" reading describe the same trajectory through different lenses; only the cap-removal lens predicts the γ-dependent divergence quantitatively.

Remark 6 (Connection to SAM). SAM's update $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t + \rho \nabla L/\|\nabla L\|)$ does *not* enter the rank-1 recurrence (6): expanding to $O(\rho)$ (Andriushchenko & Flammarion, 2022), its effective gradient is $\nabla L^{\mathrm{SAM}} \approx (I + \frac{\rho}{\|g\|} H)g$, rescaling the top direction by $1 + \rho\lambda_1/\|g\| > 1$ (curvature-amplifying), which lowers the stability boundary SAM tracks to $\lambda^*_{\mathrm{SAM}} = \frac{\|g\|}{2\rho}\big(\sqrt{1 + \frac{8\rho}{\eta\|g\|}} - 1\big) < 2/\eta$ (Long & Bartlett, 2024), i.e. $y^*_{\mathrm{SAM}} < 0$. Rank-1 damping rescales the top direction by $1/d < 1$ (curvature-attenuating), raising the boundary to $y^* > 0$. This is the formal content of the attenuation-vs-amplification dichotomy: the two edits move the EoS threshold in opposite directions along the *same* monotone $y^*(\gamma)$ of (4), and reaching the sub-threshold region $y < 0$ requires the amplifying sign. The amplifying sign does *not* have to be global: rank-1 amplification on the top direction alone reaches $y^* < 0$ (Corollary 3). At first order, SAM contains a curvature-weighted correction whose positive-curvature components have the amplifying sign; this motivates the sign comparison with Corollary 3, but full SAM is not identified with this linear amplification: §5.3 shows that the finite re-query residual bundle carries the observed endpoint difference along the tested ResNet interpolation.

The assumptions above define the regime in which the reduced recurrence is interpretable; we do not empirically certify each assumption pointwise along every network trajectory. In particular, spectral isolation, slow variation of the reduced coefficients, and piecewise-constant sensing are modeling conditions rather than measured invariants of the full network dynamics.

Gap inventory (2-D theorem → trajectory theorem).

- **G1** *Multi-eigenvalue decoupling.* The lemma projects dynamics onto (u_1, rest); the multi-eigenvalue extension Damian et al. (2023) defer (§8.2) is open.
- **G2** *Rotation singularity.* A1 fails at eigenvalue crossings $\lambda_1 \leftrightarrow \lambda_{j^*}$, where ∇S is singular; patching is open (empirically the trajectory crosses in $O(\eta^{-1})$ steps with the qualitative verdict intact).
- **G3** *Time-varying γ.* A2 freezes γ; the extension to piecewise-constant γ_t (refresh interval N) needs a perturbation analysis.
- **G4** *Higher-order x-bounding.* Remark 4's fast-direction amplification is bounded by quartic-and-higher landscape terms the 2-D reduction omits.
- **G5** *Sensing error.* A3 is ideal; with $\hat{v}$ at angle φ to u_1 the effective damping is $\gamma_{\mathrm{eff}} = \gamma \cos^2 \varphi$ and the lemma holds with $\gamma \to \gamma_{\mathrm{eff}}$.

A.4 Two further controls confirm cap robustness

The lemma concerns damping the *top* direction. Two further controls test the symmetric direction and the effect of sensing lag; both confirm that descent below the threshold did not occur for any single-direction *attenuating* or lagged edit we test (the amplifying sign is tested separately, Corollary 3). The empirical data for both live in §4 (Fig. 5); here we state the structural content only.

Corollary 4 (orthogonal damping leaves the cap intact)**.** *The C4 edit $\Delta\theta = -\eta(g_\parallel + g_\perp/d)$ with $g_\parallel = (g^\top u_1)u_1$ preserves the gradient component along u_1. Its x-recursion is therefore the undamped Damian map $x_{t+1} = -(1 + \eta y_t)x_t$, whose period-2 fixed point is $y^* = 0$. The orthogonal damping rescales only the complement and does not enter the top-direction dynamics. The cap holds for every d because the x-recursion is the undamped Damian map, whose sign-flip fixed point is $y^* = 0$* regardless *of how the y-drift coefficient is affected: the edit-induced correction to the drift is odd in x (zero mean per cycle) and its even remnant only renormalizes the amplitude-setting β, which does not enter the cap location. Experimentally confirmed in §4 (Fig. 5); the widening of per-seed variance at large d reported there is a measured instance of gap **G4**, where freezing the complement weakens the higher-order $|x|$-bounding the 2-D reduction omits.*

Together with Lemma 1 (top-direction damping $\Rightarrow y^* > 0$, upward escape) and Corollary 4 (orthogonal damping $\Rightarrow y^* = 0$, cap intact), the verdict is symmetric: damping either direction alone fails to achieve sub-threshold escape; descending below requires the amplifying sign: global (Remark 6) or rank-1 local (Corollary 3).

Observation 1 (short sensing lags preserve the response; the gradient-direction proxy pins). The delayed-update sensor $\hat{v}_k = \Delta\theta_{t-k}/\|\cdot\|$ conflates two distinct imperfections: *lag*, and using the update (gradient) direction rather than the eigenvector. Separating them shows that only the second pins (§4, Fig. 5): the lagged-but-accurate sensor $\hat{v}_k = \hat{u}(t{-}k)$ stays in regime (ii), because u_1 rotates on the slow $O(1/\eta)$ timescale (Damian et al., 2023) and $\hat{u}(t{-}k) \approx u_1(t)$ (we test only short lags $k \leq 3$). The lagged update-direction sensor $\hat{v}_k = \Delta\theta_{t-k} \propto -g$, by contrast, suffers a ***self-induced sensing failure***: the edit's own damping rotates the update direction, the lagged $\hat{v}$ drifts off u_1, its Rayleigh quotient collapses, the self-gating clamp sets $\gamma{=}0$, and the system reduces to undamped Damian GD with $y^* = 0$ (Remark 3): the operative cause is inaccuracy, not staleness.

The cap is thus robust to the tested short sensing *lags* ($k \leq 3$) but not to sensing *inaccuracy*: only accurate top-direction sensing achieves cap removal, and the failure of the delayed-update proxy is attributable to sensing-direction inaccuracy rather than to these tested lags. (Without the self-gating clamp, a stale Rayleigh quotient flips $\gamma < 0$ and anti-damps to divergence: an optimizer pathology, not cap removal; §4.)

B THE PROXY LADDER AND MONITOR COMPARISONS

This appendix is supporting instrumentation for the control study of §4 and §4.1: it characterizes the cheap sharpness monitors used as diagnostics throughout, and delimits, via a one-HVP warm-started comparator (§B.4), when the gradient Rayleigh quotient is the right monitor for the regime at hand.

B.1 ZERO-COMPUTE (TIER-0) PROXIES ARE NORMALIZATION-DEPENDENT AND SEED-UNSTABLE

On plain nets, every layer's weight norm tracks $\lambda_{\max}$ (mean Pearson 0.65–0.71 over 14 runs: the main teacher–student run plus 13 difficulty/lr/width sweeps), with the *last-hidden* layer best (best layer in $10/14$ runs), refining Li et al. (2022). Under BatchNorm the correlation *flips sign* (-0.8): BN renormalizes activations, decoupling $\|\mathbf{W}\|$ from curvature; the affine scale γ inherits the signal. However, multi-seed ($n = 5$) the BN-γ proxy is **unstable**: detrended $+0.247 \pm 0.340$ (per-seed from $+0.55$ to -0.05), CI spanning zero. On LayerNorm-based Transformers, zero-compute norms plateau: gradient norm $\|\mathbf{g}\|$ is marginally best ($+0.52$ on both ViT and ConvNeXt) but only ~ 0.08 above $\|\gamma\|/\|W_{\text{attn}}\|$ (~ 0.44), and splitting attention from MLP weights does not help. (A 16-point detrended correlation can swing ± 0.25 across reasonable windows because $\|\gamma\|$ moves $< 1\%$ while $\lambda_{\max}$ swings $20\times$; dense sampling gives $+0.44$.)

B.2 A SINGLE-HVP GRADIENT RAYLEIGH QUOTIENT: A HIGH-CORRELATION TRAJECTORY MONITOR IN THE EOS REGIME

The quantity $\hat{\lambda}_{\mathbf{g}} = \mathbf{g}^\top H\mathbf{g}/\mathbf{g}^\top\mathbf{g}$ (one Hessian-vector product) is algebraically the critical sharpness λ_c on a quadratic (Kalra et al., 2026) and a Rayleigh lower bound on $\lambda_{\max}$, tight iff $\mathbf{g}$ aligns with v_{top}. Empirically on the ResNet (Table 3, Fig. 4): detrended $r = +0.898 \pm 0.141$ over the whole trajectory and $+0.981 \pm 0.030$ in the EoS regime ($|\cos(\mathbf{g}, v_{\text{top}})| > 0.7$), where the $\sim 17\%$ relative gap is the expected Rayleigh lower-bound slack (the EoS-regime alignment is higher, $|\cos| \to 0.85$–0.9, giving $\cos^2 \approx 0.8$), not measurement noise. This dominates every Tier-0 proxy and the BN-γ baseline at $+0.247 \pm 0.34$. Cost: one HVP for the deployed monitor (vs ~ 5 HVPs for exact warm-started measurement Ramasinghe et al., 2026, vs 5–6 forward passes for the λ_c line search Kalra et al., 2026).

Table 3: Single-HVP $\hat{\lambda}_{\mathbf{g}}$ vs Tier-0 proxies on the ResNet ($n = 5$ seeds). Detrended Pearson r with $\lambda_{\max}$.

proxy	whole trajectory	EoS regime ($\lvert\cos(\mathbf{g}, v_{\text{top}})\rvert > 0.7$)
$\hat{\lambda}_{\mathbf{g}}$ (1 HVP, gradient)	$\mathbf{+0.898 \pm 0.141}$	$\mathbf{+0.981 \pm 0.030}$
$\hat{\lambda}_{\Delta\theta}$ (1 HVP, update)	$+0.896 \pm 0.144$	$+0.975 \pm 0.047$ (err 0.238 ± 0.026)
$\|\text{BN}\,\gamma\|$ (0 HVP)	$+0.247 \pm 0.340$ (unstable)	—

The Rayleigh bound $\hat{\lambda}_{\mathbf{g}} \leq \lambda_{\max}$ held at $248/248$ informative measurement points across the $n{=}5$ monitor runs (the five exceptions are the initialization step, where both quantities are $\sim 10^{-2}$ and the reference reads noise), and two eigenvector-free candidates for a deployment gate (thresholding $\hat{\lambda}_{\mathbf{g}}$ against $2/\eta$, and a self-referential saturation rule) failed to isolate the aligned regime with useful precision, so none is claimed.

B.3 The monitor probe is the gradient, not the update

For $\lambda_{\max}$ *monitoring*, the raw gradient beats the momentum-smoothed update: $|\cos(\mathbf{g}, v_{\text{top}})| = 0.563 \pm 0.017 > |\cos(\Delta\theta, v_{\text{top}})| = 0.511 \pm 0.024$, and the EoS relative error of $\hat{\lambda}_{\mathbf{g}}$ (0.169 ± 0.026) is below that of $\hat{\lambda}_{\Delta\theta}$ (0.238 ± 0.026). The whole-trajectory correlations of $\hat{\lambda}_{\mathbf{g}}$ and $\hat{\lambda}_{\Delta\theta}$ are tied (~ 0.90): the distinction shows up in alignment and EoS error, not in global correlation. Mechanistically, the momentum buffer is an EMA of past mini-batch gradients whose directional noise *dilutes* alignment with v_{top}. This bifurcates our use from CDAT/Kalra: step-size *tuning* wants the (preconditioned) update direction; $\lambda_{\max}$ *monitoring* wants the raw gradient. This distinction also limits the causal interpretation of the stochastic intervention-strength analysis: our controller edits and energy shares are defined on the instantaneous gradient, whereas the optimizer state that determines the parameter displacement contains accumulated momentum.

B.4 A one-HVP warm-start comparator: when the quotient is the right monitor

The quotient's one-HVP cost is shared by a natural alternative: a warm-started power-iteration tracker that carries its previous eigenvector estimate forward and refines it with a single HVP per measurement (Ramasinghe et al., 2026). We compared the two estimators at identical cost and measurement cadence (one measurement per 50 steps; ground truth = warm-started power iteration / ARPACK $\lambda_{\max}$; EoS gate $|\cos(\mathbf{g}, v_{\text{top}})| > 0.7$) on two families: (A) the MLP teacher–student testbed (full-batch GD, offline exact-HVP reconstruction, $n{=}5$ seeds, 59 points per seed) and (B) the CIFAR-10 PreAct-ResNet (0.175M parameters, BatchNorm + momentum, mini-batch; online tracker, $n{=}5$ seeds, 51 points per seed; Table 4).

The regime contrast is the point. In the full-batch slow-drift family (A) the tracker dominates at equal cost: it rides the warm-started eigenvector (mean $|\cos(v_{\text{ws}}, v_{\text{top}})| = 0.915$, vs. $|\cos(\mathbf{g}, v_{\text{top}})| = 0.427$ for the gradient) and its median level error is ~ 0.03 where the quotient's Rayleigh slack is ~ 1: outside the EoS window the gradient is nearly orthogonal to v_{top}, so the quotient reads the spectral bulk. Under mini-batch churn the roles reverse: between 50-step measurements the BatchNorm parameters and the iterate move enough that the carried vector goes stale (mean overlap 0.582), and the tracker decorrelates (0.118 ± 0.139 overall; -0.061 ± 0.272 in the EoS window) exactly where the quotient is strongest (0.987 ± 0.008). The quotient's value is therefore not universal cost superiority (at matched cost the tracker wins in slow-drift regimes) but its gradient-coupled semantics and EoS-regime coverage: it monitors the direction the optimizer actually uses, precisely in the regime this paper studies. Ground truth costs 3.9 HVPs per measurement in family A and 13.8 in family B.[1]

[1] Scope caveats: both families are toy-scale (an MLP teacher–student testbed and a 0.175M PreAct-ResNet); the cadence is one measurement per 50 steps, not per-step; the MLP comparator uses offline exact-HVP reconstruction, validated by a loss gate $\leq 2 \times 10^{-7}$ and agreement of the reconstructed $\hat{\lambda}_{\mathbf{g}}$ with its online value to $\leq 3.5 \times 10^{-5}$; and $4/295$ MLP tracker estimates are marginally negative (isolated steps $\sim$850–1050 on seeds 1–3, minimum -0.075), which we report as-is rather than clip.

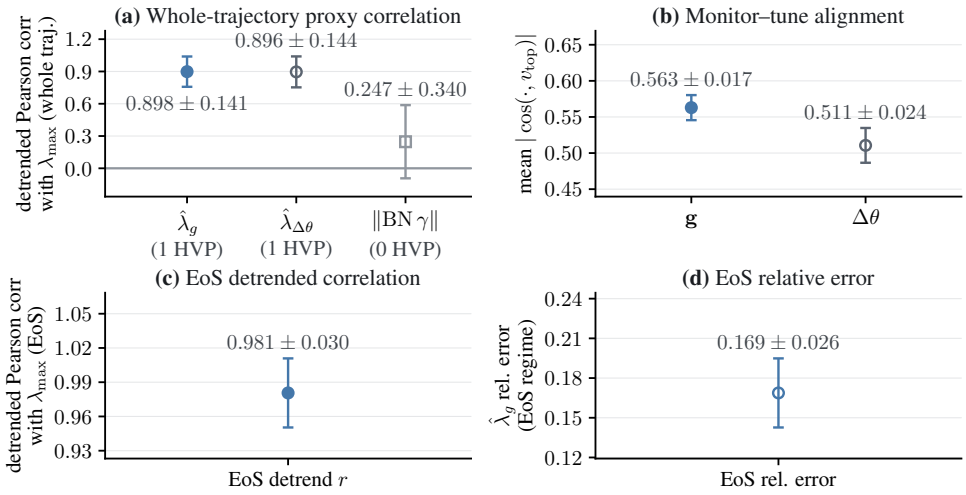

Figure 4: Gradient Rayleigh-quotient monitoring on the ResNet ($n = 5$ seeds). **(a)** Whole-trajectory detrended Pearson correlation with $\lambda_{\max}$; **(b)** mean absolute alignment with v_{top}; **(c)** EoS-window detrended correlation; **(d)** EoS relative level error. Bars show the across-seed mean±95% CI. The one-HVP gradient Rayleigh quotient $\hat{\lambda}_{\mathbf{g}}$ is robust in the EoS window, the Tier-0 $\|\mathrm{BN}\,\gamma\|$ proxy is seed-unstable, and $\mathbf{g}$ aligns with v_{top} better than $\Delta\theta$.

Table 4: One-HVP warm-started power-iteration tracker ($\hat{\lambda}_{\mathrm{ws}}$) vs. the gradient Rayleigh quotient $\hat{\lambda}_{\mathbf{g}}$ at identical cost and cadence (measurement every 50 steps; ground truth = warm-started power iteration / ARPACK $\lambda_{\max}$; EoS gate $|\cos(\mathbf{g}, v_{\mathrm{top}})| > 0.7$; mean±95% CI over n=5 seeds; rel. err. = median relative level error; ms = wall-clock per measurement; early = pre-EoS window, not defined for the online family B).

family	estimator	Pearson	early	EoS	rel. err. (med.)	ms
A: MLP full-batch	$\hat{\lambda}_{\mathrm{ws}}$	0.981±0.004	0.979±0.008	0.976±0.033	0.027±0.014	5.8
A: MLP full-batch	$\hat{\lambda}_{\mathbf{g}}$	0.537±0.070	0.464±0.237	0.913±0.121	0.982±0.001	6.0
B: PreAct-ResNet	$\hat{\lambda}_{\mathrm{ws}}$	0.118±0.139	—	−0.061±0.272	0.321±0.026	29.8
B: PreAct-ResNet	$\hat{\lambda}_{\mathbf{g}}$	0.919±0.067	—	0.987±0.008	0.344±0.033	33.0

C Robustness controls: full protocols and results

C.1 Protocol map and reading guide

The empirical logic is deliberately staged. Section 4 asks only whether a known orbit-location prediction survives a feedback-free full-batch intervention; Section 5 then drops the reduced-model interpretation and treats the stochastic experiments as protocol-specific intervention outcomes; Section 5.3 uses SAM only as a contrasting control family, not identified with the full-batch amplification mechanism.

Protocol map. The full-batch fixed-divisor experiments use unconditional fixed $d>1$ attenuation with accurate warm-started sensing. The canonical stochastic rank-1 result is reported for both the legacy gated sensor and corrected ARPACK sensing; mechanistic conclusions use the corrected result. The wider stochastic rank experiments use the no-floor raw-divisor signed rescaling of §5.2. The clamp-free control of §C.4 is a separate bridge experiment, not unconditional damping. SAM/linSAM experiments use current-minibatch first-order or re-query gradients, not the eigenspace-rescaling rule.

C.2 PROXY-DRIVEN NORM CAPS (C1, C2)

(C1) Control on a plain net: weight-norm cap. Project $\|\mathbf{W}_{\text{lh}}\|$ onto a tight ball when it exceeds budget $1.01{\cdot}\|\mathbf{W}_{\text{lh}}\|_{\text{init}}$. The projection fires on $\sim$750 of 3000 steps (per-seed $745/751/746/745/723$, $n{=}5$), yet $\lambda_{\text{raw}}^{\text{final}} = 40.26{\pm}0.43 = 2/\eta$ (no escape); the peak is not reduced within seed-noise.

(C2) Control on a BN net: γ-cap. Project all BN-γ onto $1.02\cdot\|\gamma_{\text{init}}\|$. $\lambda_{\text{raw}}^{\text{final}} = 81.6{\pm}19.2$ vs baseline $73.0{\pm}12.9$ (no escape; accuracy slightly worse, 0.754 vs 0.768).

C.3 CAP ROBUSTNESS: ORTHOGONAL DAMPING AND SENSING LAG

Damping the orthogonal complement also fails (C4: cap-robust). The rank-1 edit damps the *top* direction; the symmetric control damps *everything else*, $\Delta\theta = -\eta(g_{\|} + g_{\perp}/d)$, keeping the top-direction gradient intact (Corollary 4). Sweeping $d \in \{1, 1.5, 2, 4, 8\}$ ($n{=}5$, MLP full-batch), λ_{raw} pins at $2/\eta \approx 40$ for every d ($40.2, 40.3, 38.9, 43.4, 40.1$; $0/5$ seeds diverge). The orthogonal damping slows convergence (final loss $0.11 \to 0.59$ as $d \to 8$) and shrinks the along-u_1 oscillation ($|x_t|$ $0.18 \to 0.03$), but the top-direction self-stabilization is untouched, exactly as the 2-D reduction predicts ($g_{\|}$ preserved $\Rightarrow$ the x-recursion is the undamped Damian map $\Rightarrow$ $y^* = 0$). At $d \geq 4$ the per-seed spread widens sharply (CI ± 9 at $d{=}4$ vs ± 0.3 at $d \leq 1.5$, ± 1.7 at $d{=}2$): freezing the complement removes directions that help bound the u_1 oscillation, and the the constant-α 2-D reduction loses precision: a measured instance of gap **G4**, neither a descent below the threshold nor an upward escape ($0/5$ diverge at every d). With the rank-1 result (top damping $\Rightarrow$ escape past $2/\eta$), the two single-direction controls fail by *opposite* mechanisms: one removes the cap upward, the other leaves it in place. Neither descends below the threshold.

Short sensing lags ($k \leq 3$) preserve the response; the failure of the delayed-update proxy is attributable to sensing-direction inaccuracy rather than to these tested lags. The delayed-update sensor $\hat{v}_k = \Delta\theta_{t-k}/\|\cdot\|$ conflates lag with using the update (gradient) direction rather than the eigenvector. Separating them ($n{=}5$, MLP): the lagged-but-*accurate* sensor $\hat{v}_k = \hat{u}(t{-}k)$ escapes upward to the §4.3 escape value for $k{=}1, 2, 3$, numerically identical to $k{=}0$; u_1 rotates on the slow $O(1/\eta)$ timescale, so $\hat{u}(t{-}k) \approx u_1(t)$ and the cap is still removed (Observation 1). The lagged *update-direction* sensor $\hat{v}_k = \Delta\theta_{t-k} \propto -g$ pins at $2/\eta$ for $k{=}1, 2, 3$ ($40.10 \pm 0.13, 40.03 \pm 0.02, 40.03{\pm}0.03$): the operative cause is a *self-induced sensing failure*, not staleness: the optimizer's own damping rotates the update direction off u_1 and disengages the edit (Remark 3; reduced-model account in §4.1, Fig. 5). With *free lagged* $\hat{v} = \Delta\theta_{t-1}/\|\cdot\|$ ($|\cos(\Delta\theta, v_{\text{top}})| \to 0.998$, Damian et al., 2023), $\lambda_{\text{raw}}^{\text{final}} = 40.12 \pm 0.12 = 2/\eta$: the lagged $\hat{v}$ imperfectly cancels $\mathbf{g}_{\|}$, the period-2 oscillation persists, and the cap holds (Claim 1, regime i; Fig. 14). (Without the clamp, the stale Rayleigh quotient flips $\gamma < 0$ and *anti-damps*: a divergence we observe at $k{=}1$, step ~ 540, an optimizer pathology.) The cap is thus preserved under the tested short lags ($k \leq 3$); the delayed-update proxy's failure is attributable to sensing-direction inaccuracy rather than to these tested lags.

C.4 CLAMP-FREE BRIDGE: THE NULL IS NOT A CLAMP ARTIFACT

Removing the self-gating clamp changes nothing (clamp-free C3). The X-SAM projector ablation (§4.4) revealed a potential confound in the clamp design: the clamped X-SAM projector variant fired its rank-1 correction on 0% of steps where the matched-step (unclamped) one fires every step: the $\max(\cdot, 1)$ clamp can manufacture dormancy. The stochastic-regime rank-1 arm uses the *same* clamp and showed no statistically resolved difference from SGD partly through this viewpoint (it fired on $\leq 4\%$ of steps, $\approx$ SGD), so that null could be attributed to the clamp rather than the landscape. We bridge this with a *clamp-free* arm on the same CIFAR-100/ResNet-18 setup, deterministic CUDA, $n{=}5$ seeds (`--no-clamp`): no $\max(\cdot, 1)$ gate, every direction fires every step (fired fraction 100%), the raw divisor $d = \hat{\lambda}/\lambda_{\text{targ}}$ allowed below 1.

What "clamp-free" edits (Algorithm 1, clamp-free branch). The semantics must be stated precisely: this is *not* an "always damp" arm. With $\alpha = \langle \mathbf{g}, \hat{v} \rangle$ the $\hat{v}$-component of the gradient, the edit is $\mathbf{g} \mapsto \mathbf{g} - f\,\alpha\,\hat{v}$ with $f = 1 - 1/d$ *clipped to* $[-1, 1)$. When $\hat{\lambda} > \lambda_{\text{targ}}$ ($d > 1$), $f \in (0, 1)$ and the edit damps, leaving α/d. When $\hat{\lambda} < \lambda_{\text{targ}}$ ($d < 1$), the factor flips sign ($f < 0$): instead of opposing the $\hat{v}$-component, the edit reinforces it: bounded sign-flip anti-damping. The clip keeps the edit magnitude at most $|\alpha|$ and the post-edit $\hat{v}$-component at most $2|\alpha|$, so the anti-damping is

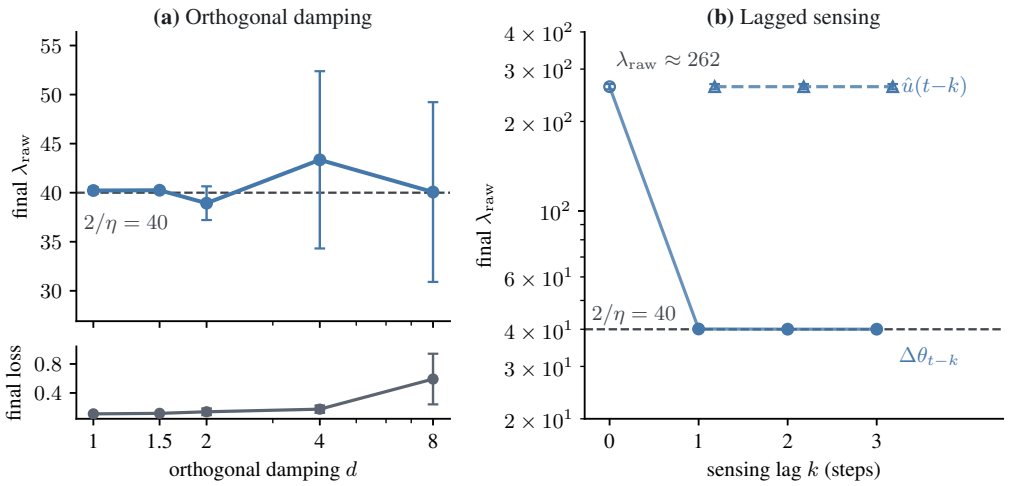

Figure 5: Two further controls on the MLP full-batch setup (n=5 seeds). **(a)** Orthogonal damping (C4): λ_{raw} pins at $2/\eta \approx 40$ for every tested $d \in \{1, 1.5, 2, 4, 8\}$ ($0/5$ diverge; loss rises $0.11 \to 0.59$, lower strip); damping the complement does not touch the top-direction cap (Corollary 4). In (a), markers show across-seed means and error bars 95% CIs of the final-step λ_{raw} (n=5). **(b)** Sensing lag vs. inaccuracy: the lagged-but-accurate sensor $\hat{u}(t-k)$ escapes upward to 261.8 ± 5.4 ($\triangle$), identical to k=0; the lagged update-direction sensor $\Delta\theta_{t-k}$ pins at $2/\eta$ ($\circ$). Short lags ($k \leq 3$) preserve the response; the gradient-direction proxy, not staleness, pins (Observation 1).

bounded and cannot explode. This bound is exactly what separates the bridge from the un-clipped lagged variant of §C.3 (Fig. 5), whose raw factor ($\gamma \to -\infty$ as the stale Rayleigh quotient collapses) anti-damped into divergence at k=1, step ~ 540: there the sign flip was unbounded and diverged, here it is clipped: the two clamp-free statements are consistent, and it is the bound, not the gate, that separates them.

No upward escape without the clamp. We compare four arms under one deterministic driver (Table 5), reporting the late-window (step $\geq$ 2000) single-HVP gradient Rayleigh quotient $\hat{\lambda}_{\mathbf{g}}$ on the deterministic 512-sample eval batch (n=5 seeds, mean$\pm$95% CI). At this stage the power-iteration $\lambda_{\max}$ estimate was not yet trustworthy (the false convergence later diagnosed in §5.1), so we report the deterministic per-model proxy $\hat{\lambda}_{\mathbf{g}}$, a lower bound on $\lambda_{\max}$ (§6), and keep all comparisons within-protocol.

Table 5: Clamp-free bridge, CIFAR-100/ResNet-18, deterministic CUDA (n=5 seeds): late-window (step $\geq$ 2000) $\hat{\lambda}_{\mathbf{g}}$ mean$\pm$95% CI, and the fired fraction of the rank-1 correction. **Within-protocol comparison only**: $\hat{\lambda}_{\mathbf{g}}$ is a single-HVP lower-bound proxy on a fixed eval batch and its absolute scale (13–18) is *not* comparable to the ARPACK λ_{ref} scale (22–29) of the stochastic-regime control in §5.1.

arm	late $\hat{\lambda}_{\mathbf{g}}$	fire%
SGD	18.73 ± 1.29	—
rank-1 (clamped C3)	18.13 ± 1.47	4
rank-1 (clamp-free C3)	$\mathbf{17.14 \pm 1.21}$	100
SAM	12.91 ± 0.59	—

The clamp-free arm does *not* escape upward (Fig. 6): its differences are $\Delta = -1.59$ vs. SGD and -0.99 vs. the clamped arm, both well inside the $\sim$1.5-unit CIs, nowhere near the upward escape of the accurate-v_{top} arm on the full-batch MLP (the §4.3 escape value; Fig. 15). The unresolved rank-1 outcome is therefore robust to the clamp at this tested operating point: with the gate removed and the correction firing every step, sharpness still shows no statistically resolved difference from SGD. (The clamp-free arm does sit $\sim$1.6 below SGD, the direction Duan et al. predict (Duan et al., 2026), but

well inside the CI and far smaller than SAM's -5.8; the no-resolved-contribution verdict stands.) The same driver re-validates the deterministic three-way: SAM (12.91 ± 0.59) lies below both rank-1 arms (paired Δ vs. SGD -5.8 above), so the stochastic-regime ordering of §5.1 survives the stricter determinism standard introduced by the X-SAM projector ablation.

Robustness of the null. The two residual suspicious readings of the stochastic-regime rank-1 result, clamp artifact and dormant sensor, are excluded independently: the null survives with the clamp removed *and* with the sensor fixed. The ARPACK-sensing arm of §5.1 fires at full force (no clamp floor, refreshed v_{top} every $N{=}20$ steps) yet still shows no statistically resolved sharpness reduction relative to SGD at this tested operating point.

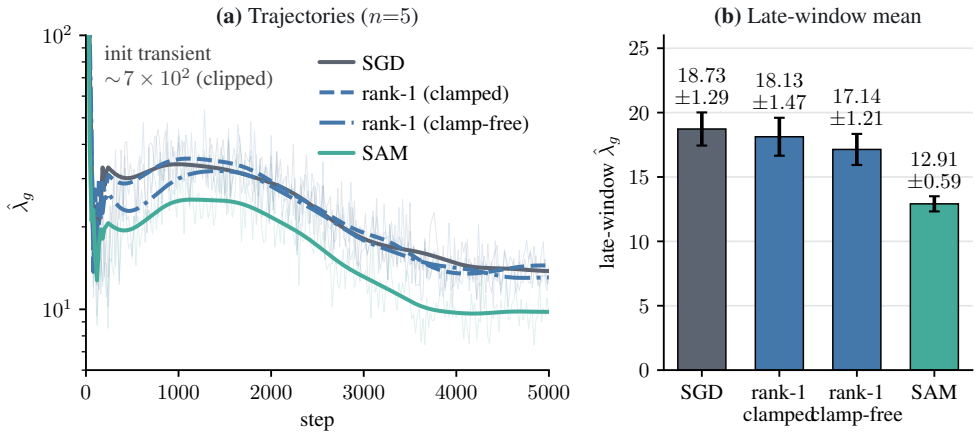

Figure 6: Clamp-free bridge (deterministic CUDA, $n{=}5$ seeds): late-window $\hat{\lambda}_{\mathbf{g}}$ of SGD, clamped rank-1, clamp-free rank-1, and SAM: (a) cross-seed mean trajectories Gaussian-smoothed in log space for display (faint thin lines: unsmoothed means; the two-sample init transient is left unsmoothed); (b) late-window mean$\pm$95% CI (arms distinguished by linestyle and hue, so the separation survives grayscale printing). The clamp-free arm fires every step yet stays within CI of SGD and of the clamped arm (no upward escape, no clamp artifact) while SAM sits below them.

C.5 Sign-flip control: table and algorithm

Fixed-divisor grid, round 13 (Fig. 1; §4.2). Protocol: the §3 tanh MLP testbed (teacher–student synthetic, $n{=}1024$, dim 20, student $[256, 256]$), full-batch GD, data seed 1, model seeds 0–4. All 18 cells of the unified round share one CPU backend (float32, torch 2.9.1, 8 threads; $n{=}5$ seeds, 3000 steps; late window step ≥ 1500). Seven cells settle onto the predicted orbit ($\eta{=}0.025$: ratios $0.973/0.990/0.971$ at $d{=}1.25/1.5/2$; $\eta{=}0.05$: $0.957/0.969/0.956/0.975$ at $d{=}1.25/1.5/2/6.5$). At $\eta{=}0.05$, $d{=}6.5$ survives $3/5$ at ratio 0.975, consistent with the earlier GPU round ($3/5$, 0.997; its single-η levels $47.9/59.2/74.7/259.3$ against $50/60/80/260$, ratios $0.96/0.99/0.93/1.00$, agree within a few percent). Two $\eta{=}0.025$ cells survive off the identity: $d{=}2.5$ is intermittent (ratio 0.592; per-seed late means 100–150), and $d{=}6.5$ sits at $\lambda_{\text{raw}} \approx 258$ (per-seed 244–269, sd 9), an η-independent saturation level coinciding with the $\eta{=}0.05$, $d{=}6.5$ survivors' absolute level (243–260) and with the rank-1 damping escape ceiling 261.8 ± 5.4 (§4.3); we read it as a ceiling on the damped escape set by the higher-order dynamics, not as a slow approach to $2d/\eta = 520$. All remaining cells diverge ($\eta{=}0.1$: $0/5$ throughout; $\eta{=}0.05$, $d \in \{2.5, 3\}$: $0/5$; $\eta{=}0.025$, $d{=}3$: $0/5$).

Round-13 reproducibility. Relative to the previous CPU round, interior cells reproduce to $\leq 1\%$ in the late-window mean, while boundary (knife-edge) cells swing in survivor count and are not bit-reproducible: CPU reduction order matters exactly at the divergence boundary ($\eta{=}0.05$, $d{=}2$: $2/5$ here vs. $3/5$ previously). The paper uses only this unified round. We therefore do not use survivor counts in these marginal cells as quantitative evidence for a universal basin boundary; their role is only to show that orbit survival is distinct from orbit location.

Delayed-onset control (§4.2). The fixed-d survival grid is shown in Fig. 1b. To isolate whether the failures are caused by intervening too early, per seed, the adaptive arm ran to its measured natural gate-opening step T_{gate} (seeds 0–4: $200/194/188/195/206$; mean 196.6), then switched to fixed

$d \in \{2.5, 3\}$ from exactly that step: $0/5$ survive at both d (divergence in every seed), while the adaptive reference is $5/5$ stable from the same state (late $\bar{\lambda} = 258.2$, $\eta\bar{\lambda}/2 = 6.46 \approx d$=6.5).

For the sign-flip amplification control of §4.3, two qualifications matter: (i) the amplified equilibrium is *intermittent*, not a clean pin: λ_{raw} swings between deep dips ($\sim$0–2) and transient spikes ($\sim$43–47) on a $\sim$500-step scale, with window means tight across seeds (per-seed 29.56–30.69 at α=0.25), consistent with the arm's harder conditional y-contraction as α grows; and (ii) the escape does not trade optimization for sharpness: minimum losses match GD's (0.092–0.095 vs. 0.092), though final losses run modestly higher under the intermittency (0.126 ± 0.009 / 0.133 ± 0.028 / 0.197 ± 0.115).

Table 6: Sign-flip control on the MLP full-batch setup (n=5 seeds): late-window (step ≥ 1500) λ_{raw}, mean$\pm$95% CI, and the fraction of measurement steps below $2/\eta$=40. The 2-D reduction (Corollary 3) predicts equilibria $\lambda^* = 2/((1+\alpha)\eta)$ for the amplifying arms: direction and monotonicity, not magnitude (Corollary 2).

arm	late λ_{raw}	predicted λ^*	frac$<2/\eta$
GD	37.49 ± 1.76 (final 39.92 ± 0.40)	40 (pin)	0.56 ± 0.32
rank-1 damping (accurate, full batch)	261.8 ± 5.4	— (upward escape)	0.00
SAM (full perturbed-point update)	21.52 ± 1.01	—	1.00
rank-1 amp, α=0.25	30.14 ± 0.50	32.0	0.88 ± 0.04
rank-1 amp, α=0.5	22.45 ± 1.27	26.7	0.97 ± 0.07
rank-1 amp, α=1	14.25 ± 2.74	20.0	0.97 ± 0.05

The magnitude gap grows monotonically with the edit strength ($-5.8\%/-15.9\%/-28.8\%$ at $\alpha = 0.25/0.5/1$), consistent with the weak-edit locality of the 2-D reduction rather than a sign error in the fixed-point algebra: the reduction freezes $(\alpha_{\text{ps}}, \beta, \lambda_1)$ at the undamped operating point, and the predicted orbit's displacement from that point grows with α. Two observations support this reading: at the weakest edit the gap is within the realization offset the *undamped* arm itself shows in the same window (GD 37.49 vs. the pin 40), and the documented intermittency (deep sub-threshold dips, spikes capped just above threshold, CI widening with α as the conditional y-contraction tightens) is the orbit-level signature of the x-odd correction terms the constant-coefficient reduction omits (size $\propto \alpha$; Appendix A), whose asymmetric time-average sits below the sign-flip value. Both vanish in the weak-edit limit $\alpha \to 0$, where the prediction reduces to the undamped pin. A fresh six-point grid run in a single session ($\alpha \in \{0.05, 0.1, 0.2, 0.25, 0.5, 1\}$; n=5 seeds; late-window means $37.5/33.5/31.7/28.6/23.9/15.2$ against predictions $38.1/36.4/33.3/32.0/26.7/20.0$) verifies the weak-edit limit directly: the realized-to-predicted gap at $\alpha = 0.05$ is -1.5%, within the in-session undamped arm's own realization offset (39.5 vs. the pin 40, -1.2%), and its magnitude grows with the edit strength ($-7.9/-5.1/-10.7/-10.2\%$ in the interior, -23.8% at $\alpha = 1$). This fresh grid is an independent validation run; Fig. 7 and Table 6 report the original sign-flip run.

C.6 Multi-sharp landscape test

Damping only u_1 lets u_1 escape upward while λ_2 stays pinned ($40.1 \to 65.2$, seed 0); here k matters: k=1 : 64.0 $<$ k=2 : 75.2 $\approx$ k=5 : 77.4, each removed cap escaping independently. Figure 8 summarizes the final levels and the two-eigenvalue trajectories for this multi-sharp control.

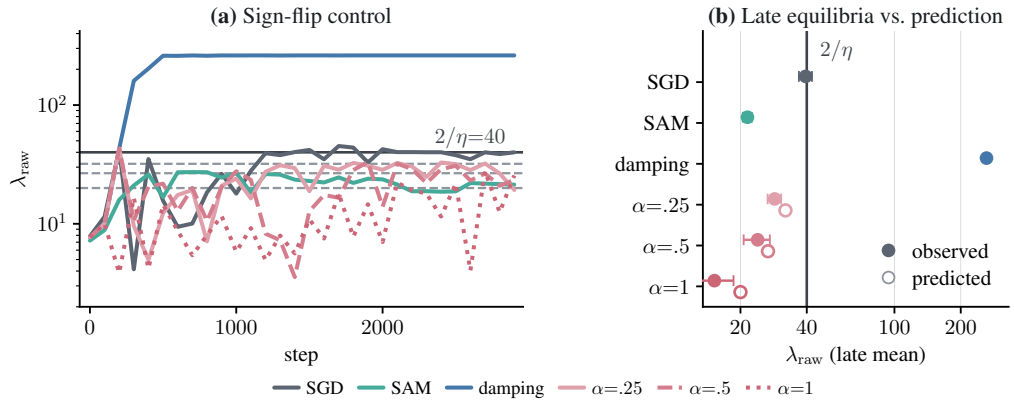

Figure 7: Sign-flip control on the MLP full-batch setup (n=5 seeds, using the accurate-sensing MLP protocol of §4.3 and Table 6). **(a)** λ_{raw} trajectories (bold seed-mean): the attenuating rank-1 damping edit (accurate sensing) escapes past $2/\eta$=40 to ∼262, while the sign-flipped amplifying arms at $\alpha \in \{0.25, 0.5, 1\}$ descend below $2/\eta$ with equilibria monotone in α; SGD pins at $2/\eta$, and SAM descends below it. **(b)** Late-window (step $\geq$ 1500) mean±95% CI against the 2-D prediction $\lambda^* = 2/((1+\alpha)\eta)$ (thin gray dashed lines): the sign and monotonicity match, while the realized equilibria sit below the naive fixed-point values (direction-only, Corollary 2).

Algorithm 1 Rank-1 gradient edit (C3), the clamp-free bridge, and the fixed-divisor interventional control

Require: step size η, gradient $\mathbf{g} = \nabla L(\theta)$, unit $\hat{v}$ (free-lagged, power-iterated, or ARPACK-sensed), target λ_{targ}=$2/\eta$, variant $\in$ {clamped, clamp-free, fixed-d}

$\alpha \leftarrow \langle \mathbf{g}, \hat{v} \rangle$; $\mathbf{g}_\perp \leftarrow \mathbf{g} - \alpha\,\hat{v}$ ▷ split $\mathbf{g}$ along $\hat{v}$
$\hat{\lambda} \leftarrow \hat{v}^\top H \hat{v}$ ▷ one Hessian-vector product
$d \leftarrow \hat{\lambda}/\lambda_{\text{targ}}$
if variant = clamped (default C3) **then**
  $d \leftarrow \max(d,\, 1)$ ▷ self-gating clamp: fire only above target
  $\tilde{\mathbf{g}} \leftarrow \mathbf{g}_\perp + \alpha\,\hat{v}/d$
else if variant = fixed-d (interventional control, §4.1) **then**
  $d \leftarrow d_{\text{fix}}$ (constant) ▷ no $\hat{\lambda}$ feedback: tests λ_{eq}=$2d/\eta$
  $\tilde{\mathbf{g}} \leftarrow \mathbf{g}_\perp + \alpha\,\hat{v}/d$
else if variant = clamp-free **then**
  $f \leftarrow \text{clip}(1 - 1/d,\, -1,\, 1)$ ▷ bounded sign-flip anti-damping below target
  $\tilde{\mathbf{g}} \leftarrow \mathbf{g}_\perp + (1-f)\,\alpha\,\hat{v}$
end if
$\theta \leftarrow \theta - \eta\,\tilde{\mathbf{g}}$
return θ

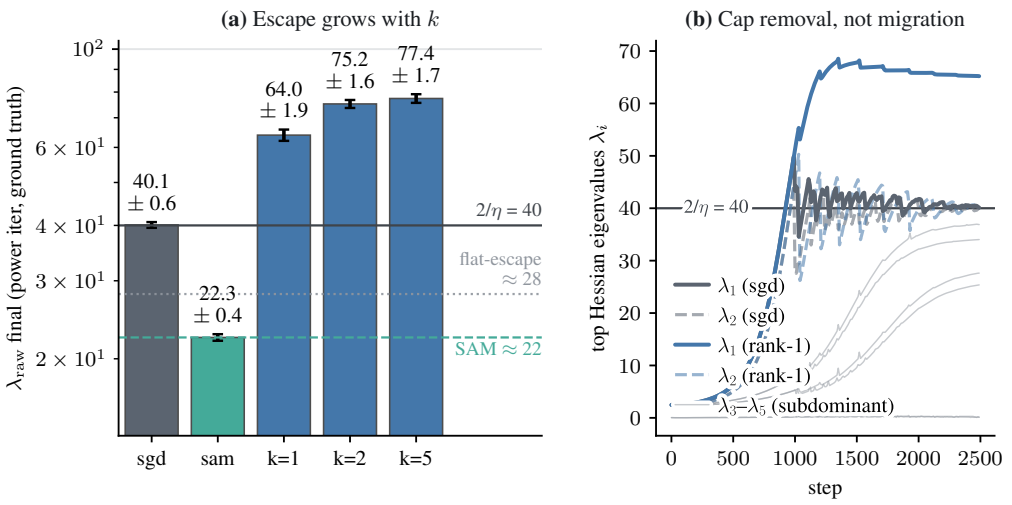

Figure 8: Multi-sharp landscape test ($n = 5$ seeds). A constructed multi-output isotropic teacher–student pins *two* directions at $2/\eta$ ($\lambda_1 = 40.1$, $\lambda_2 = 40.0$). Left: final λ_{raw} across seeds (mean$\pm$95% CI); damping only u_1 (rank-1) lets u_1 escape upward while λ_2 stays pinned: no flow in this constructed test, inconsistent with migration where it could plausibly act. Right: seed-0 top-eigenvalue trajectories; $k > 1$ now matters, revealing the per-direction structure of the self-stabilization cap. Full values and protocol are given in §C.6.

D STOCHASTIC-REGIME MEASUREMENT AND ROBUSTNESS DETAILS

D.1 CROSS-BATCH SENSING AND DESCRIPTIVE BOUNDARY ORGANIZATION

What the fixed-batch reference measures. A cross-batch sensing study. All reference measurements above (and the control law's sensing) use one fixed 512-example batch, so two questions arise: is that batch's λ_{ref} representative, and is its top eigenvector the sharp direction of the *current* minibatches the optimizer actually sees? We re-ran vanilla SGD with an instrumented protocol (n=3, same deterministic driver) that measures, at every outcome step, the ARPACK top-3 on two additional fixed 512-example batches (seeds 124/125) and on the *current training minibatch* (bs 128). Late-window results: (i) the fixed batches *agree at the eigenvalue level* (late λ_{ref} 28.61 ± 3.42 / 28.12 ± 0.83 / 32.20 ± 2.40 across the three batches), so the reference protocol is sound as a scalar measurement; (ii) the current-minibatch $\lambda_{\max}(H_B)$ sits markedly higher, 37.0 ± 9.7 (per-measurement median 37, IQR $[30, 48]$), straddling the heavy-ball reference line at 38 (Fig. 9); and (iii) the top eigenspace is substantially batch-specific late in training: the reference batch's v_1 overlaps the current minibatch's v_1 with $|\cos|$ median only 0.26 (IQR $[0.13, 0.51]$; vs. 0.47–0.53 against the other fixed batches; top-3 subspace min-cosines 0.10–0.31). Early and mid training the overlap is high ($|\cos| = 0.89$/0.90 median): the shared sharp direction fragmentates as the spectrum flattens. Consequence: a fixed-batch-sensed $\hat{v}_1$ is *not* the current minibatch's sharpest direction late in training, so the rank-1 edit acts on a direction with small gradient share per step, which is one candidate mechanism for the single-direction null below. The current-minibatch-sensing control tests it by intervention (below).

Thresholds: which boundary each diagnostic straddles. The two reference lines, the momentum-free $2/\eta$ and the heavy-ball reference line $2(1{+}\mu)/\eta$=38, sit at different relations to the three diagnostics: the batch sharpness (20.51) is centered near $2/\eta$, the fixed-batch λ_{ref} (28.6) sits below 38, and the current-minibatch $\lambda_{\max}(H_B)$ (37.0) straddles the heavy-ball reference line. One reading must be ruled out explicitly: taking the minibatch $\lambda_{\max}(H_B)$ "straddling 38" as an independent confirmation of stochastic EoS conflates $\lambda_{\max}(H_B)$ with the batch sharpness; the directional quantity the EoSS theory constrains sits near $2/\eta$ in this protocol, not at the heavy-ball reference line: an empirical point in the batch-size-dependent regime family of momentum EoSS (Andreyev et al., 2026), in which the batch sharpness plateaus at the noise-dominated small-batch value $2(1{-}\mu)/\eta$ ($= 2$ here) and rises to the large-batch heavy-ball plateau $2(1{+}\mu)/\eta$ ($= 38$) as batch size grows. We do not read $2/\eta$ as a universal momentum-EoSS threshold: at our particular operating point (bs 128, μ=0.9) the measured batch sharpness lies near the momentum-free boundary $2/\eta$, between the two plateaus; mapping that interpolation is outside our scope (§6). The ordering $\lambda_{\mathrm{step}} < \lambda_{\mathrm{ref}}(\mathrm{fixed}) < \lambda_{\max}(H_B)$ is also consistent with the weak-actuation diagnostic from §5.2: the stochastic gradient follows directions with bulk-scale curvature, while edits confined to the spectral edge intersect only a small fraction of the per-step gradient.

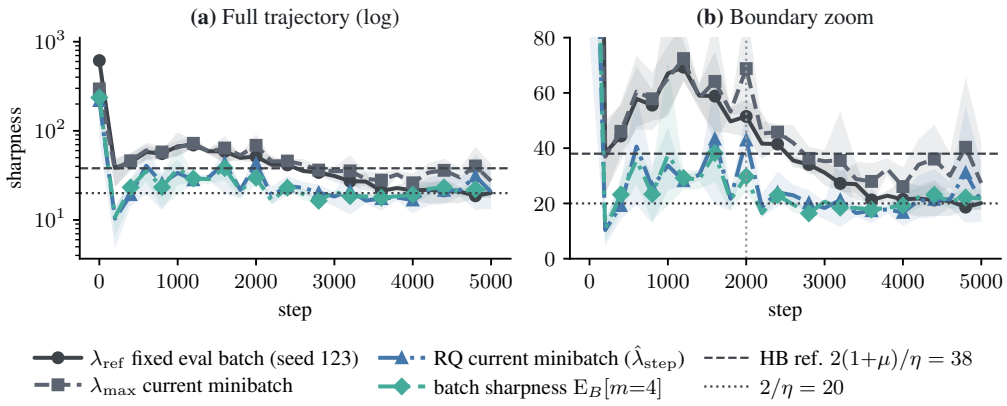

Figure 9: Batch sharpness vs. spectral sharpness for SGD on ResNet-18/CIFAR-100 ($n{=}3$ seeds, deterministic protocol; bold curves are seed means and bands are min–max across seeds). **(a)** full trajectories (log scale); **(b)** boundary-region zoom with the momentum-free reference $2/\eta$ (dotted) and the heavy-ball reference line $2(1{+}\mu)/\eta{=}38$ (dashed). The EoSS batch sharpness $\mathbb{E}_B[\mathbf{g}_B^\top H_B \mathbf{g}_B / \|\mathbf{g}_B\|^2]$ ($m{=}4$ minibatches per measurement) is centered near $2/\eta{=}20$ (20.51 ± 2.10 late), while the current-minibatch $\lambda_{\max}(H_B)$ (37.0 ± 9.7) and the fixed-batch reference λ_{ref} (28.6 ± 3.4) are distinct, larger quantities; late scalar summaries are mean$\pm$95% CI across the three seeds.

D.2 Batch-sharpness estimator: m-convergence

The $m{=}4$-minibatch average used for λ_{step} throughout trades Monte-Carlo variance for cost. At matched end-of-training checkpoints of an independent re-execution of the canonical SGD protocol (not a bit-replay: nondeterministic kernels; statistically the same family), the estimator reads 23.11 ± 1.93 at $m{=}4$ vs. 24.40 ± 0.43 at $m{=}16$ and 24.55 ± 0.92 at $m{=}64$ (mean$\pm$sd, $n{=}3$ seeds): $\approx$1.4–1.5 units below the $m{=}64$ reference and $\sim 4.5\times$ the seed spread of $m{=}16$. Every headline batch-sharpness contrast (e.g. SAM$-$SGD -4.30, 95% CI $[-6.70, -1.89]$) exceeds this difference; unresolved contrasts are reported as such.

Subsampling calibration of the $m{=}4$ estimator. We replicate the in-run estimator exactly (same minibatch sampling, eval-mode BN-frozen convention), draw an $m{=}128$ pool of per-batch values at each of the three end-of-training SGD checkpoints of the instrumented rerun, and bootstrap $B{=}2000$ subsamples of size $m{=}4$ with replacement, matching the i.i.d. sampling of the in-run estimator; a disjoint-partition companion (32 groups of 4, no reuse) gives the same conclusion. Table 7 reports the results: the subsample distributions are centered on the pool estimate (group-z mean 0.00, sd 0.93–0.94, 94–97% of $|z| < 2$), with dispersion matching the i.i.d. prediction $\hat{\sigma}/2$. The earlier $m{=}4$ vs. $m{=}64$ scan gap (23.11 vs. 24.55) is $0.37\times$ the standard error of an $m{=}4$ estimate: finite m therefore inflates variance without evidence of systematic downward bias.

Table 7: Subsampling calibration of the $m{=}4$ batch-sharpness estimator at the three end-of-training SGD checkpoints ($m{=}128$ pool per checkpoint; $B{=}2000$ bootstrap subsamples of size $m{=}4$, with replacement; theory sd is the i.i.d. prediction $\hat{\sigma}/\sqrt{m} = \hat{\sigma}/2$ with $\hat{\sigma}$ the pool sd).

checkpoint (seed)	$\hat{\lambda}_{128}$	boot mean$\pm$sd	theory sd
instrumented SGD rerun, seed 0	26.13	26.14 ± 4.37	4.36
instrumented SGD rerun, seed 1	24.90	24.99 ± 4.52	4.41
instrumented SGD rerun, seed 2	23.01	23.11 ± 2.83	2.90

D.3 Spectral redistribution

No persistent spectral redistribution detected at the sampled checkpoints. Within the measured top-10 spectrum and the available checkpoint cadence, we detect no persistent redistribution sufficient

to explain the rank-1 outcome (Fig. 10; n=3): the damped direction remains the outcome top eigenvector in 98% of late measurements (median $|\cos(\hat{v}_{\text{ctrl}}, v_1)|$=0.978; within the top-5: 100%), the spectrum shows no resolved change: late λ_2/λ_1 is 0.845 under rank-1 damping vs. 0.853 under SGD, with $\lambda_2 \approx 25.3$ vs. 24.4, and the top-10 subspace turnover between consecutive measurements is identical (median min-cosine 0.069 vs. 0.070). Nothing flows: damping $\hat{v}_1$ neither demotes the damped direction nor lifts λ_2 into a vacancy. The stochastic-regime null is therefore not sharpness "escaping around" the edit: at this operating point the edit has weak per-step actuation ($\sim 10^{-4}$ gradient share), but §5.2 and the regime sweep show that this diagnostic alone does not determine the long-horizon response. (This is the stochastic-regime, measured counterpart of the MLP multi-sharp migration refutation, §4.4.)

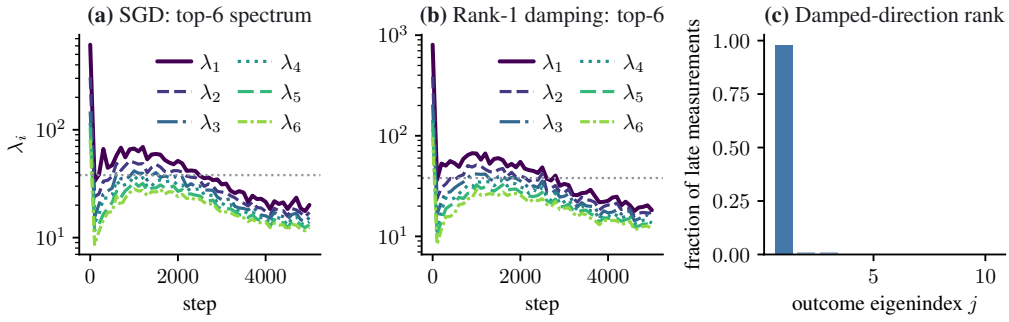

Figure 10: Spectral redistribution test (n=3 seeds, top-6 outcome spectrum, seed means). **(a)** SGD vs. **(b)** accurate rank-1 damping, both on log-scale y-axes: the spectra show no resolved difference (late λ_2/λ_1: 0.845 vs. 0.853). No sharpness flows into the complement. Eigenvalue index is encoded redundantly by color and line style. **(c)** The damped direction $\hat{v}_1$ remains the outcome top eigenvector in 98% of late measurements (median $|\cos|$=0.978): the edit neither demotes the damped direction nor lifts λ_2; the top of the spectrum is simply unperturbed, consistent with the $\sim 10^{-4}$ per-step gradient share of the edit.

D.4 Subspace share at k=50/100

Extending the measurement toward the class count evaluates, at matched canonical checkpoints (eval mode), the cumulative minibatch-gradient energy share $\mathbb{E}_B[\sum_{i\le k}(\mathbf{g}_B \cdot \hat{v}_i)^2/\|\mathbf{g}_B\|^2]$ over m=64 fresh minibatches against the checkpoint top-k eigenspace (k up to 100, single ARPACK solve per checkpoint, ∼9 min per checkpoint). This convention differs definitionally from the in-run measurement of §5.2: the checkpoint share is an expectation over the minibatch distribution against a fixed eigenspace of the checkpoint Hessian, whereas the in-run s_k is the current single training gradient's alignment with the live-sensed subspace at that step (train mode, $k \le 20$). Both are correct under their own conventions: the gap (s_{20}=0.133 vs. $\lesssim$0.3%) is definitional, not a discrepancy. The checkpoint-share measurement is also the operationally closer analogue of the persistent dominant subspace aggregated over training that Song et al. (2025) target (§2); a small in-run share is not inconsistent with either. Aggregate over n=3 seeds (mean±sd): $s_{20} = 0.133 \pm 0.011$, $s_{50} = 0.189 \pm 0.009$, $s_{100} = 0.239 \pm 0.010$: no majority of gradient energy resides in the top eigenspace even at the class count, leaving ≈76% outside the top-100 directions.

D.5 Batch-size × momentum regime sweep

Exploratory batch-size × momentum map: asymmetric re-emergence of the damping effect (Fig. 11). The deterministic and stochastic regimes so far sit at two distant operating points. An exploratory sweep ($B \in \{32, 128, 512, 2048\} \times \mu \in \{0, 0.9\}$, n=3, canonical 5000-step protocol, lr not rescaled with B; sensing runs at every cell, the edit arms run at the critical cells $\{128/\mu$=0, $512/\mu$=0.9, $2048\times\}$; the $(128, 0.9)$ anchor is the cell above) samples intermediate operating points between them with four measurements per cell: the gradient-energy share of the edited top-1 subspace, the paired outcome of each edit sign, the EoSS batch sharpness, and the reference λ_{ref} (pre-registered predictions and window rules in the supplement; because fixed steps confound epochs ($3.2 \to 205$ epochs across B) and momentum shifts convergence speed by $(1-\mu)^{-1}$, $B \le 128$

cells use the late window and $B \geq 512$ cells the training-phase mid window). We consequently treat this experiment as an *exploratory phase map*, not as an estimate of a universal regime boundary. The main observation is *asymmetric* (Fig. 11b): rank-1 *damping* shows a statistically resolved shift toward higher sharpness where the dynamics turn quasi-deterministic—paired $\Delta{=}+44.0$, 95% CI $[+39.6, +48.4]$ at $B{=}2048/\mu{=}0.9$, in the direction the 2-D reduction predicts for damping (the boundary moves *up*)—while at $B{=}128/\mu{=}0$ the point estimate is also positive ($+49.7$, 95% CI $[-4.2, +103.5]$) but is not statistically resolved at the 95% level, and rank-1 *amplification* shows no statistically resolved effect toward lower sharpness at *any* tested (B, μ). Strikingly, the gradient-energy share barely moves across the sweep ($2.1{\times}10^{-4}$ at $B{=}128$ vs. $4.6{\times}10^{-4}$ at 2048, $\mu{=}0.9$): what returns at the quasi-deterministic end is not per-step gradient share but outcome sensitivity, and it returns only in the reduction's predicted harmful direction. Across the tested batch-size sweep, only SAM exhibits the lower-sharpness response; §5.3 provides a pathwise decomposition of the ResNet linSAM–SAM endpoint gap. Two boundary facts complete the phase picture (Fig. 11c–d): the near-$2/\eta$ anchor is a special interior point—at $\mu{=}0.9$ the batch sharpness leaves $2/\eta$ in both directions ($3.9{\pm}1.0$ at $B{=}32$; $94.8{\pm}42.6$ mid-window at 512, above *both* EoSS plateaus, before memorization collapses it to $\sim$1)—and at the tested $\mu{=}0$ operating points the measured batch sharpness does not concentrate near the deterministic $2/\eta$ reference: it runs 69–508 across B (three- to twenty-five-fold above it) with healthy optimization throughout (final accuracies 32–50%, consistent with each cell's epoch count). We therefore do not use the deterministic boundary as a descriptive law for these cells. These cells also caution against treating the deterministic $2/\eta$ line as a universal descriptive law for the measured stochastic diagnostic; the reference lines are contextual markers, not fitted constraints for every cell of the sweep.

operating point (B, $\mu{=}0.9$)	share s_1	paired $\Delta\lambda_{\mathrm{ref}}$ vs. SGD, 95% CI
128 (canonical, late window)	$2.1{\times}10^{-4}$	$+0.6$ $[-3.9, +5.0]$
512 (mid window)	$7.2{\times}10^{-4}$	$+12.1$ $[-28.6, +52.8]$
2048 (quasi-deterministic, mid window)	$4.6{\times}10^{-4}$	$+44.0$ $[+39.6, +48.4]$

Table 8: Exploratory batch-size sweep: one rank-1 damping edit at $\mu{=}0.9$ ($n{=}3$ per cell; Fig. 11). The gradient-energy share stays near 10^{-4} while the paired outcome shifts from null to sharply positive as the dynamics turn quasi-deterministic. This sweep is exploratory: the fixed step budget conflates batch size with epochs seen, and late windows are used for $B \leq 128$ but mid windows for $B \geq 512$, so it establishes an outcome-sensitivity contrast rather than a controlled regime bridge.

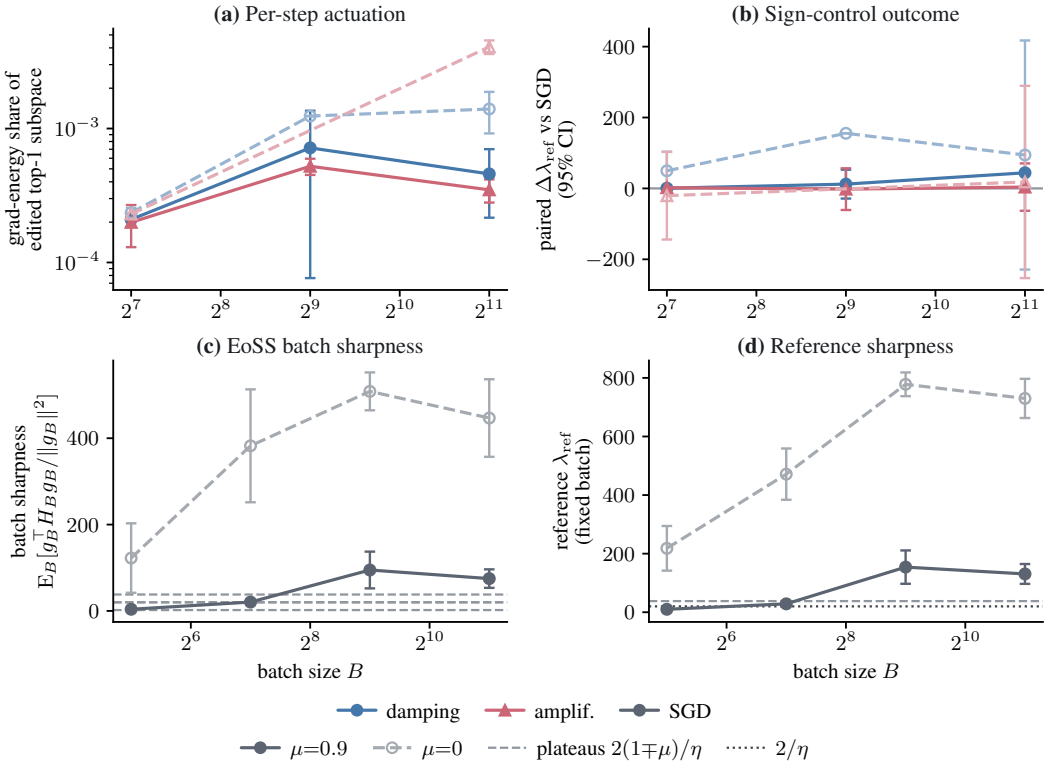

Figure 11: Regime-transition sweep ($B \in \{32,128,512,2048\} \times \mu \in \{0,0.9\}$, n=3, canonical protocol; the B=128/μ=0.9 anchor reuses the main runs). **(a)** gradient-energy share of the edited top-1 subspace (damping/amplifying arms): near-flat at 10^{-4}–10^{-3} across the whole range: no monotone return of per-step effect. **(b)** paired Δ reference λ_{ref} vs. SGD with 95% CIs: damping (circles) shows a statistically resolved shift toward higher sharpness at the quasi-deterministic $2048/\mu$=0.9 point (+44.0, 95% CI $[+39.6, +48.4]$); the $128/\mu$=0 point estimate is positive but not statistically resolved at the 95% level; amplification (triangles) is never statistically resolved toward lower sharpness. **(c)** EoSS batch sharpness vs. B with the momentum plateaus $2(1\mp\mu)/\eta$: the near-$2/\eta$ anchor is a special point, and the measured batch sharpness lies several-fold above the deterministic $2/\eta$ reference at μ=0. **(d)** reference λ_{ref} (fixed 512-batch ARPACK): decouples from $2/\eta$ at both extremes. Window rule per supplement: $B \leq 128$ late, $B \geq 512$ mid (training phase). In (a), (c), and (d), markers show across-seed means of per-seed window means and error bars 95% CIs (n=3).

D.6 BN-PROTOCOL CONTROLS: FULL RESULTS

BN-protocol controls: BatchNorm bookkeeping alone does not reproduce the SAM–linSAM separation. SAM updates the BN running statistics twice per step (forwards at θ and $\theta+\epsilon$), while the reference λ_{ref} is measured in eval mode through those statistics, so BN-state evolution is a potential contributor to the measured response. Three bit-verified protocol controls (n=5, identical protocol) test this explanation directly. Reproducing SAM's second forward on a pure-SGD update (`sgd_bn2x`) leaves late λ_{ref} without a statistically resolved difference from SGD (26.3 ± 0.8 vs. 28.6 ± 3.4; paired $\Delta=-2.1$, 95% CI $[-5.4, +1.2]$): the extra BN update alone does not produce a statistically resolved shift at this power. Supplying linSAM with SAM's exact BN evolution (`linsam_bnmatched`, buffer trajectories bit-identical to SAM's) produces a modest resolved downward shift relative to linSAM (31.7 ± 1.7 vs. 34.5 ± 2.1; paired $\Delta=-2.8$, 95% CI $[-5.3, -0.3]$) but remains far above SAM (22.9 ± 1.6). Thus matching BN evolution is insufficient to reproduce the SAM-level reduction, although these controls do not exclude a smaller BN-state contribution. Because its buffer trajectory is supplied from the matched SAM run, `linsam_bnmatched` is a causal diagnostic for the BN-state channel rather than a standalone deployable optimizer. Conversely, freezing the running statistics in the official `disable_running_stats` style (`sam_bnfreeze`) keeps the training dynamics but raises the measured λ_{ref} to 74.2 ± 4.1, confirming that the running-stat channel can materially affect the measured reference sharpness.

D.7 VIT CONTROL: FULL RESULTS

ViT control: the re-query requirement is not universal across the tested systems. The ResNet controls establish that the transition along the tested interpolation family cannot be reproduced without the finite re-query at the BatchNorm/CIFAR-100 operating point, but they do not imply that every stochastic system requires this component. We repeat the central four arms on ViT-tiny (LayerNorm, no running statistics) at two operating points. At the below-EoS point (η=0.01 so $2/\eta$=200; n=3), SAM (21.1 ± 0.5) and its linearization (21.5 ± 0.5) show no statistically resolved difference from each other, both far below SGD (40.4 ± 2.8), with rank-1 damping in between (36.3 ± 0.5, within the SGD seed envelope at this power). To obtain a matched high-curvature operating point with λ_{ref} near $2/\eta$ requires the batch-size knob rather than the learning rate: at bs 128, an lr sweep over $[0.02, 0.2]$ leaves the late $\lambda_{\mathrm{ref}}/(2/\eta)$ at 0.34–0.41 with $\eta\,\lambda_{\mathrm{ref}} \approx 0.7$–$0.8$ across a $10\times$ lr range (an lr-independent sub-EoS equilibrium), whereas at bs 512 and lr 0.05 the late median settles at $0.96 \times 2/\eta$ across seeds ($0.95/1.08/0.84/0.98/0.94$; still below the heavy-ball reference line $2(1+\mu)/\eta$). At that matched high-curvature operating point (n=5) the ordering that isolates this component is reproduced: SAM separates below SGD (23.3 vs. 38.4 late medians; paired $\Delta=-15.1$, 95% CI $[-19.9, -10.3]$). The linearized control retains a small statistically resolved difference from SAM ($\Delta(\mathrm{linSAM} - \mathrm{SAM})=+1.7$, 95% CI $[+1.1, +2.3]$) while satisfying the prespecified TOST equivalence criterion at n=5 and the ±3.8 margin (90% TOST CI $[+1.2, +2.1]$). Rank-1 damping, which shows no statistically resolved difference at the below-EoS point, has a positive point estimate at the matched high-curvature point (paired $\Delta=+4.7$, 95% CI $[-1.3, +10.6]$): four of five seeds are positive, with transient peaks up to $2.3\times$ the SGD late level, the first appearance of the boundary-moving response in the ViT control, but the effect is not statistically resolved at the 95% level, and the evidence should not be described as resolved. Across both operating points, the linearized control remains close to SAM; at the matched high-curvature point it satisfies the prespecified equivalence test at n=5 and the ±3.8 margin. Thus, the decomposition of SAM into first-order and nonlinear components is system-dependent across the two tested systems; we use the ViT results to delimit the SAM-account claim, not to extend the full-batch EoS theory to Transformers. SAM costs $\sim$2–3 accuracy points at both budgets (0.306 vs. 0.336 below EoS; 0.371 vs. 0.390 at the matched point).

D.8 THE 200-EPOCH RECIPE: FULL DETAILS

The ordering survives and separates under the full standard 200-epoch recipe. To provide an external ordering check beyond the 5000-step underfitting regime (§6), we re-ran the three-way under the full standard recipe—78,200 steps (200 epochs), cosine lr $0.1 \to 0$, pad/crop + horizontal-flip augmentation—with exact ARPACK curvature every 5 epochs (40 checkpoints) under the deterministic protocol, n=5 seeds per arm (a 100-epoch n=3 replication and an intermediate 12000-step n=2 run had already preserved the ordering in point estimate). In the steady window (epochs 30–120, where cosine lr runs $0.088 \to 0.023$ and the annealing boundary $2(1+\mu)/\eta(t)$

is still informative), the ordering is now separated, not merely preserved in point estimate: SGD 17.91 ± 0.92, rank-1 damping 19.51 ± 4.08, SAM 12.33 ± 0.78; SAM's CI overlaps neither arm, with paired per-seed $\Delta(\text{SGD}) = -5.58$, 95% CI $[-6.46, -4.71]$. The rank-1 arm remains statistically unresolved relative to SGD: its paired difference is $+1.60$, 95% CI $[-2.94, +6.14]$, *above* SGD with the point estimate on the same above-SGD side as every damping arm of the 5000-step protocol, and the arm does not satisfy the prespecified equivalence test against SGD at this sample size and margin, while its edit fires on 37% of direction-steps, concentrated in the sharpened phase and disengaging, by the self-gating design, as annealing carries λ_{ref} below the fixed $2/\eta$ target. This long-recipe arm retains the fixed target used by the original controller rather than tracking the time-varying momentum stability reference. It therefore tests the persistence of the three-way empirical ordering under a standard training recipe, not an annealed-boundary-tracking controller. As $\text{lr} \to 0$ the boundary inflates and every arm falls well below it (Fig. 12a), so late-epoch λ_{ref} ceases to discriminate. Final accuracies on the full test set show no statistically resolved difference (74.7 ± 0.5 / 74.9 ± 0.7 / 75.3 ± 0.2 for SGD / rank-1 / SAM; SAM's paired accuracy gain $+0.57 \pm 0.75$pp, CI containing zero, p=0.10): under a deployment-grade recipe the residual sharpness separation yields no statistically resolved accuracy gain at this power, consistent with the monitoring-vs-control split being a statement about sharpness, not about accuracy.

D.9 Additional figures

Figures 12–17 collect the supporting visualizations referenced in §D.8, §5.1 and the full-batch controls, and §D.11; they are grouped here to keep the appendix text sections compact.

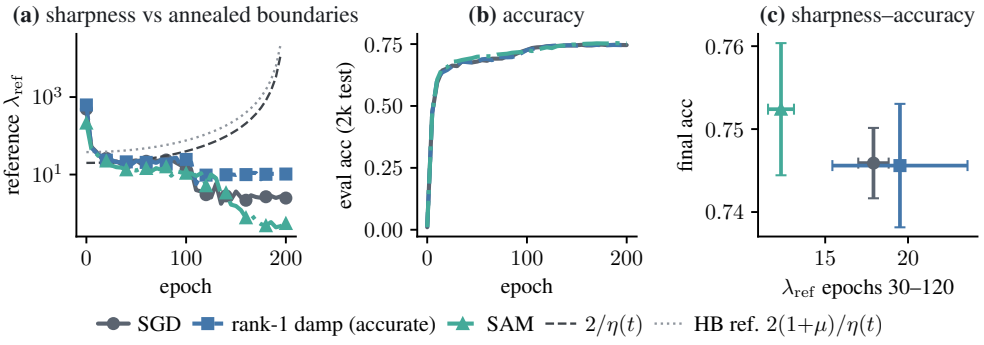

Figure 12: Standard long recipe, full budget (200 epochs, cosine $0.1 \to 0$, pad/crop+flip augmentation, n=5 seeds, exact ARPACK curvature every 5 epochs). **(a)** Reference λ_{ref} sits below the annealed reference curves $2/\eta(t)$ and $2(1+\mu)/\eta(t)$ throughout; the arms can separate, if at all, only in the steady window (epochs 30–120). **(b)** Test accuracy (fixed 2,000-example test subset). **(c)** Steady-window λ_{ref} vs. final accuracy (2k-subset values, mean±95% CI): rank-1 damping shows no statistically resolved difference from SGD (19.5 vs. 17.9; paired Δ=$+1.6$, 95% CI $[-2.9, +6.1]$); SAM separates below both (12.3 ± 0.8; paired Δ=-5.6, 95% CI $[-6.5, -4.7]$) and shows no statistically resolved accuracy gain on the full test set (SAM 75.3 ± 0.2 vs. 74.7 ± 0.5/74.9 ± 0.7; paired $+0.57 \pm 0.75$pp, p=0.10). Optimizer identity is encoded redundantly by color, line style, and marker.

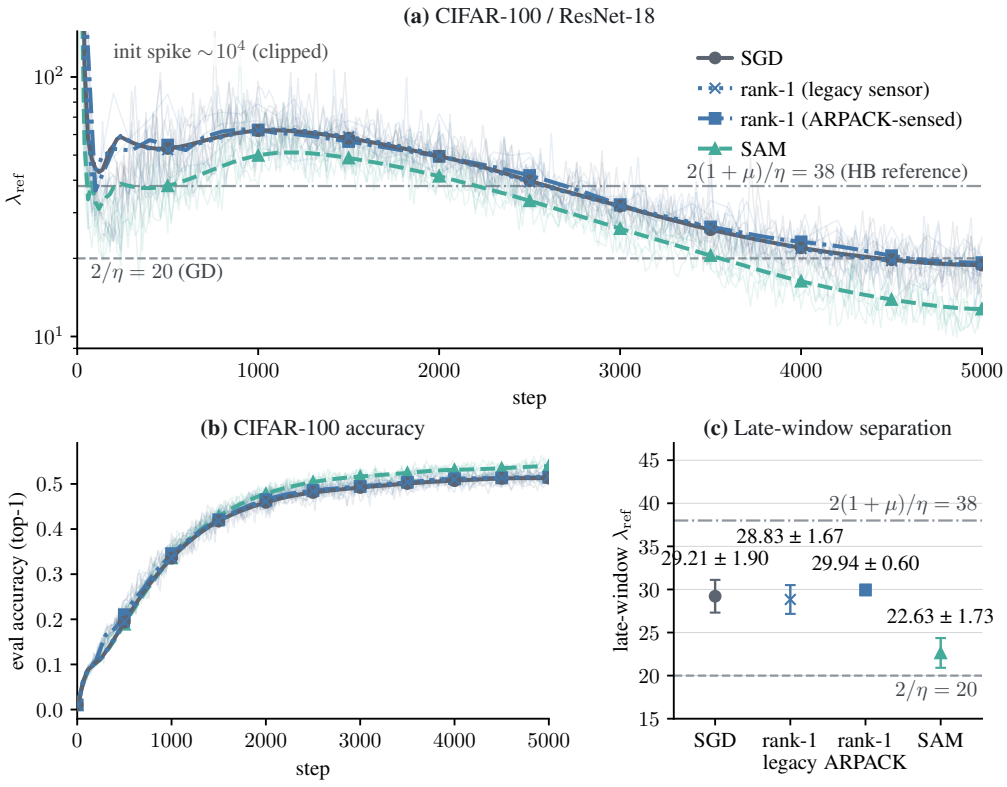

Figure 13: At-scale reproduction: ResNet-18 (11.2M) on CIFAR-100, deterministic CUDA, with λ_{ref} measured by ARPACK on a deterministic 512-sample evaluation batch. **(a)** λ_{ref} trajectories (faint per seed, bold seed mean; y-axis clipped at 150 because the initialization transient reaches $\sim 10^4$) for SGD, the legacy clamped C3 rank-1 arm with the original sensor, the corrected ARPACK-sensed rank-1 arm (the three canonical-protocol seeds), and SAM. The GD reference $2/\eta$=20 and the heavy-ball reference line $2(1+\mu)/\eta$=38 are shown for context. **(b)** Evaluation accuracy. **(c)** Late-window (step $\geq$ 2000) λ_{ref} mean$\pm$95% CI, per-seed statistics (SGD, legacy rank-1, and SAM at n=5 canonical seeds; corrected arm n=3): neither rank-1 arm shows a statistically resolved reduction relative to SGD, while SAM is lower. The legacy clamped arm fires on $\leq 4\%$ of steps because of the diagnosed sensing false-convergence; the corrected ARPACK-sensed arm engages throughout the substantive window (pooled fire rate 93.7% over steps 1–4000, n=3; five-seed engagement diagnostics in Appendix D.14) but still shows no statistically resolved sharpness reduction. λ_{ref} uses the deterministic+ARPACK protocol; under the earlier nondeterministic power-iteration measurement the ordering and CI separation are preserved, while the absolute λ_{ref} values shift upward.

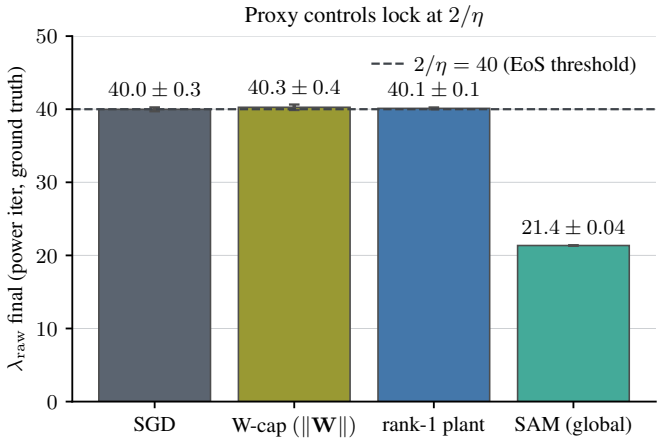

Figure 14: The W-cap and free-lagged rank-1 edit (regime i) both lock at $2/\eta$; SAM escapes below it throughout, the sign-flipped amplifying arm locally (Fig. 7). (The accurate-v_{top} rank-1 variant, regime ii, Fig. 15, instead escapes above $2/\eta$; the γ-cap control C2 is reported in the text.) Error bars are 95% CI (n=5 for all arms).

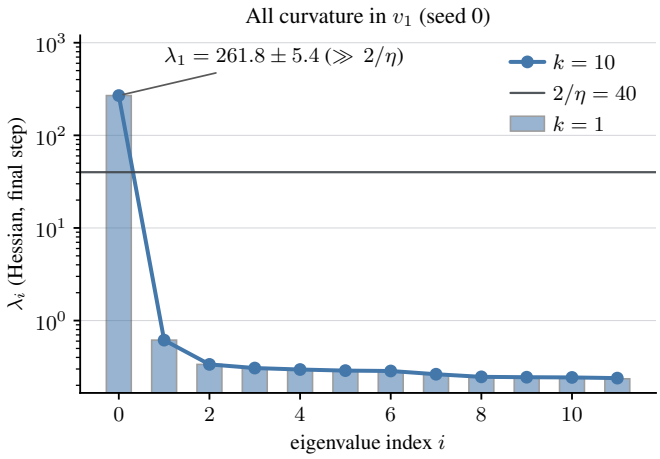

Figure 15: Top-12 Hessian spectrum of the 1-D-sharp teacher–student landscape after unconditional rank-k top-eigenspace damping with *accurate* power-iterated v_{top}. The displayed spectrum is seed 0. Across n=5 seeds, the top-eigenvalue summary is $\lambda_1 = 261.8 \pm 5.4$ (mean$\pm$95% CI), while $\lambda_2 \approx 0.6$; thus the spectrum is effectively one-dimensionally sharp. There is therefore nowhere for sharpness to migrate to in this testbed, consistent with the k-invariance (k=1, 2, 5, 10 bit-identical; text) predicted by cap removal and inconsistent with a migration account in which additional damping lowers λ_{raw}.

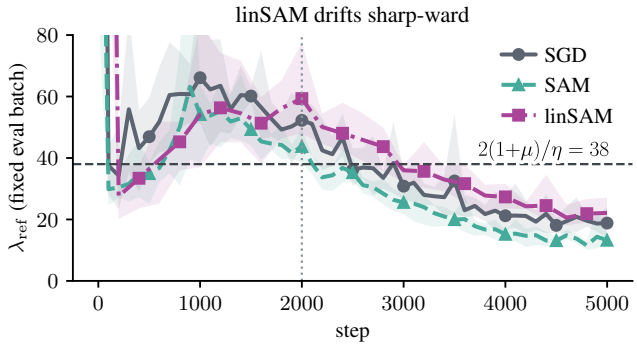

Figure 16: The full-rank linearized-SAM control ($\mathbf{g} + (\rho/\|\mathbf{g}\|)H_B\mathbf{g}$, ρ=0.05 matched to SAM; n=5 seeds, bands min–max; SGD/SAM references n=5; late window $\geq$ 2000). The first-order correction, which amplifies positive-curvature components (and attenuates negative-curvature ones), does *not* reproduce SAM: linSAM lands at 34.49 ± 2.05, *above* SGD (Δ=+5.27, 95% CI $[+1.94, +8.61]$) and $\sim$12 above SAM, the same drift toward higher sharpness as rank-k amplification. The remaining linSAM–SAM transition in this tested BatchNorm ResNet regime requires the finite re-query residual along the interpolation of §5.3; increasing the rank of the first-order correction alone does not reproduce it.

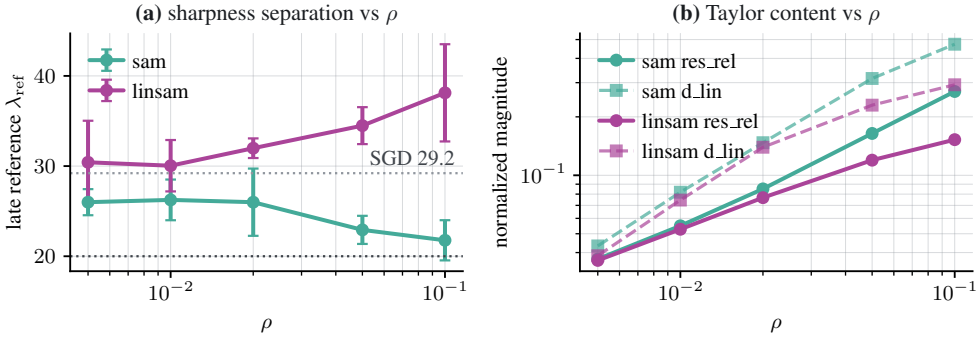

Figure 17: ρ-sweep, full SAM vs. linearized SAM ($\rho \in \{0.005, 0.01, 0.02, 0.05, 0.1\}$; n=3 per new ρ, n=5 at linSAM ρ=0.05; late window $\geq$ 2000). **(a)** late reference λ_{ref} vs. ρ (log axis; SGD 29.21 ± 1.90 dotted): SAM shows an overall decreasing trend toward lower sharpness; under the 95% paired-CI criterion, its first statistically resolved separation from SGD occurs at the n=5 canonical cell ρ=0.05, while the n=3 sweep cells at $\rho \in \{0.005, 0.01, 0.02\}$ remain unresolved at this level. linSAM shows an overall increasing trend, with the upward separation statistically resolved from ρ=0.02: the separation never inverts and generally widens with the radius. In (a), markers show across-seed late-window means and vertical bars 95% CIs; paired significance claims use the 95% CIs defined in §3. **(b)** the re-query's Taylor content on the current minibatch: nonlinear residual $\|\nabla L_B(\theta+\rho\hat{\mathbf{g}}) - (\mathbf{g} + \rho H_B\hat{\mathbf{g}})\|/\|\mathbf{g}\|$ vs. its first-order part $\rho\|H_B\hat{\mathbf{g}}\|/\|\mathbf{g}\|$ (log-log): the residual remains 0.5–0.95$\times$ the first-order term at every tested ρ: the first-order expansion is never parametrically close to the re-query it replaces.

D.10 STATISTICS DETAILS: TOST MARGINS AND DETRENDED CORRELATION

Margin sensitivity. Because the $1/4$ margin is a design choice, we recomputed every TOST test at $1/8$ and $1/2$ of the SAM effect. At the $1/8$ margin, no tested arm satisfies the equivalence criterion, including k=10. At the $1/2$ margin, the rank-1 damping arm, its current-minibatch-sensing variant, the k=5/k=10 damping arms, the mild amplification control (α=0.5), the α=1 single-direction amplification arm, and the true-$\lambda_{\max}$ sensing arm satisfy the criterion at their stated sample sizes. Rank-k amplification at k=5/20 and the wide-rank damping arms k=2/20 do not satisfy the criterion at any tested margin; the 200-epoch rank-1 damping arm likewise does not satisfy it at any of the three tested margins. Amplifying controls show predominantly upward point estimates, with statistically resolved increases under the 95% paired-CI criterion for linSAM from ρ=0.02 in the sweep and for rank-k amplification at k=20 (+4.88, 95% CI $[+0.79, +8.97]$); none produces the SAM-like downward shift. Thus the asymmetric qualitative picture is margin-robust, while the only damping arm satisfying the prespecified $1/4$-margin criterion at its stated sample size is k=10. Detrended Pearson correlation uses a rolling-mean (window 5) residual to remove the monotone global trend, the meaningful quantity when both proxy and $\lambda_{\max}$ drift.

D.11 THE ρ-SWEEP: PER-RADIUS RESULTS AND FIGURES

Per-radius late reference λ_{ref}: full SAM $26.0 / 26.3 / 26.0 / 22.4 / 21.8$ at ρ=0.005 / 0.01 / 0.02 / 0.05 / 0.1; linearized SAM $30.4 / 30.0 / 32.0 / 34.5 / 38.1$ over the same grid. Per-radius paired contrasts (95% CIs): full SAM's Δ(SAM − SGD) is -2.59 $[-5.80, +0.62]$ / -2.33 $[-7.32, +2.65]$ / -2.59 $[-9.37, +4.20]$ / -6.29 $[-8.73, -3.84]$ / -6.81 $[-9.85, -3.78]$; linearized SAM's paired Δ vs. SGD is $+1.84$ $[-5.32, +9.00]$ / $+1.46$ $[-2.26, +5.18]$ / $+3.40$ $[+0.22, +6.57]$ / $+5.27$ $[+1.94, +8.61]$ / $+9.54$ $[+2.60, +16.48]$ over the same grid.

Full-rank negative-sign control (linsamneg). The arm $\mathbf{g} \leftarrow \mathbf{g} - (\rho/\|\mathbf{g}\|)H_B g$ with ρ=0.05 matched (one HVP per step, no perturbed-point re-query, BN protocol as linSAM, n=5) applies the identical correction magnitude with the sign flipped. Late reference λ_{ref}: 31.29 ± 2.07; paired contrasts $+2.07$, 95% CI $[-1.55, +5.70]$ vs. SGD (this arm does not satisfy the prespecified equivalence test at the stated sample size and margin) and -3.20, 95% CI $[-5.96, -0.44]$ vs. linSAM. Batch sharpness: 20.85 ± 0.93, paired $+0.43$, 95% CI $[-2.73, +3.59]$ vs. the instrumented SGD rerun (n=3), at the SGD level. Realized-magnitude caveat: the arm is step-wise norm-matched by construction, but realized along-trajectory correction magnitudes differ (0.56 ± 0.16 vs. linSAM's 0.23 ± 0.03, late-window Taylor column) because the trajectories explore different curvature regions.

Update-norm diagnostic. Could the monotone interpolation response of §5.3 be an effective-step-size artifact? We reconstruct $\|g_\tau\|$ from the per-step instrumentation of the interpolation arms; the endpoints are bitwise identical to the independent `sam` and `linsam_bnmatched` runs (dedicated 60-step validation contracts, metrics and jsonl agreeing to machine precision). Under the pre-update semantics (the norm that defines the effective step $\eta\|g_\tau(\theta_t)\|$), the directly instrumented late-window means are $1.690/1.692/1.646$ at $\tau = 0/0.25/0.5$ (cross-seed sd 0.03–0.08, n=5), a 2.7% span relative to $\|g_{\mathrm{lin}}\|$. The τ=1 arm has no per-step jsonl, so its pre-update norm is not treated as a direct measurement. A cross-arm validation gives 1.657 with a matched spread of 6.4% at p50, consistent with the directly logged range but used only as a validation point. A secondary semantics that instead closes the reconstruction at the Taylor point θ_{t+1} shows a monotone -10.2% trend, but that semantics mixes in the post-descent contraction of the same-batch gradient (measured $G'/G \approx 0.58$ late), an outcome of the step's aggressiveness, not a cause. The available pre-update diagnostic therefore does not indicate a large effective-step-size change along the tested interpolation. The quantitative record behind Fig. 18: at $\tau = 0/0.25/0.5/1$ (mean±95% CI, n=5), late λ_{ref} is $31.70 \pm 1.70 \rightarrow 29.77 \pm 1.14 \rightarrow 27.45 \pm 2.56 \rightarrow 22.93 \pm 1.56$ and λ_{step} is $19.91 \pm 2.55 \rightarrow 18.94 \pm 0.92 \rightarrow 17.70 \pm 1.25 \rightarrow 16.24 \pm 0.50$; the paired λ_{step} contrast τ=1 vs. τ=0 is -3.67 with 95% CI $[-5.95, -1.39]$.

D.12 RANK INTERPOLATION: FULL TABLE AND FIGURE

Table 9 gives the full rank-interpolation statistics, and Figure 19 visualizes the same ARPACK-sensed rank-k sweep. In Table 9/Fig. 19, the legacy label "damping" denotes the no-floor raw-divisor signed rescaling for k>1; only the corrected rank-1 arm is a pure attenuation test in practice.

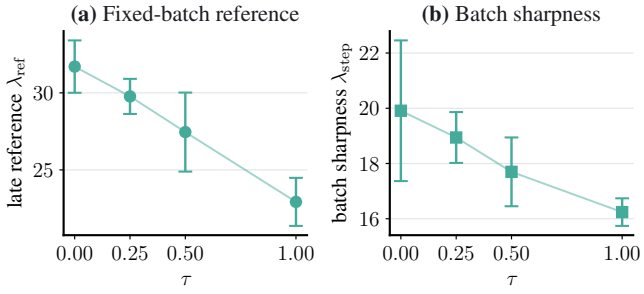

Figure 18: Finite re-query interpolation, on both sharpness quantities. **(a)** Late fixed-batch reference sharpness $\lambda_{\text{ref}}(\tau)$ ($t \geq 2000$, fixed-512 ARPACK estimate) and **(b)** EoSS batch sharpness $\lambda_{\text{step}}(\tau)$ ($t \geq 2000$, m=4 minibatches), mean±95% CI (n=5 at every τ), under $g_\tau = g_{\text{lin}} + \tau(g_{\text{SAM}} - g_{\text{lin}})$ at fixed ρ=0.05. The sampled arm means of both responses decrease across the four tested τ values; endpoints are BN-matched linSAM (τ=0) and full SAM (τ=1). Across the directly instrumented $\tau \in \{0, 0.25, 0.5\}$ arms, the late-window pre-update norm spans 2.7%; the τ=1 norm is cross-arm validated rather than directly logged (see the update-norm diagnostic in this subsection).

Table 9: Rank interpolation at ResNet-18/CIFAR-100 scale (deterministic protocol of §5.1; late reference λ_{ref}, step ≥ 2000, mean±95% CI; SGD 29.21 ± 1.90 and SAM 22.63 ± 1.73 at n=5). Δ(SGD) is the paired per-seed mean difference with its 95% CI for ordinary difference inference. Equivalence is evaluated separately by TOST at α=.05, using the corresponding 90% CI against the prespecified ± 1.645 margin; at this margin and the stated sample sizes, only the k=10 damping arm satisfies the equivalence criterion. fired is the per-direction fraction of edits with realized damping/amplification factor > 0.1; share is the edited subspace's fraction of the instantaneous gradient energy, $\sum_{i \leq k} \cos^2(\mathbf{g}, \hat{v}_i)$; amplification edits are applied unconditionally every step, so fired is not applicable to them (n.a.). n=3 seeds per damping arm; n=5 for the amplifying arms $k \in \{5, 20\}$.

arm	k	late λ_{ref}	Δ(SGD), 95% CI	fired	share
damping	1	29.94 ± 0.60	$+1.36$ $[-2.28, +5.00]$	0.80	2.0×10^{-4}
damping	2	29.13 ± 4.15	$+0.55$ $[-6.41, +7.51]$	0.74	4.2×10^{-4}
damping	5	28.82 ± 5.62	$+0.24$ $[-3.34, +3.82]$	0.60	9.3×10^{-4}
damping	10	29.47 ± 3.41	$+0.88$ $[-0.20, +1.97]$	0.38	1.8×10^{-3}
damping	20	30.23 ± 4.34	$+1.65$ $[-1.98, +5.27]$	0.20	2.9×10^{-3}
amplify α=1	1	28.73 ± 0.82	$+0.15$ $[-3.64, +3.93]$	n.a.	—
amplify α=1	5	31.22 ± 2.42	$+2.01$ $[-0.74, +4.76]$	n.a.	8.4×10^{-4}
amplify α=1	20	34.09 ± 2.28	$+4.88$ $[+0.79, +8.97]$	n.a.	2.6×10^{-3}

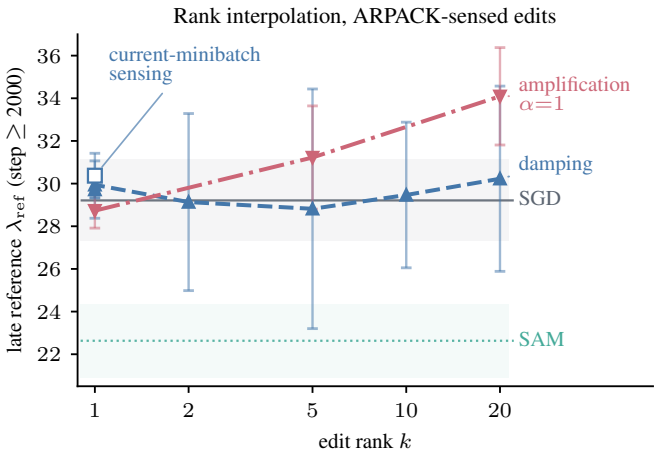

Figure 19: Rank interpolation (ARPACK-sensed rank-k edits, deterministic protocol; damping arms n=3, amplifying arms $k \in \{5, 20\}$ n=5; late reference λ_{ref} mean±95% CI; SGD/SAM bands shaded, n=5). Damping arms (blue) sit inside the SGD band at every $k \in \{1, 2, 5, 10, 20\}$: no descent below the SGD band at any tested rank; the current-minibatch-sensed control (open markers, §5.1) coincides. Rank-k amplification (rose, α=1) drifts *above* SGD as k grows: the direction of increasing sharpness, away from SAM. The edited subspace carries $\leq 0.3\%$ of the per-step gradient energy at every k (Table 9): single-subspace edits produce no statistically resolved λ_{ref} reduction at this operating point at any tested rank. Only k=10 satisfies the prespecified equivalence criterion at this sample size and the ± 1.645 margin (paired Δ=+0.88; 90% TOST CI $[+0.15, +1.62]$); the ordinary paired difference and its 95% CI are reported in Table 9 (full table: Appendix D).

D.13 REALIZED ACTUATION STRENGTH: FULL TABLE

The per-step realized relative edit $\|\mathbf{g}' - \mathbf{g}\|/\|\mathbf{g}\|$ is computed exactly from the per-step plant logs (orthonormal sensed basis): for eigenspace rescaling edits $\sqrt{\sum_i (f_i \cos_i)^2}$ with $d_i = \hat{\lambda}_i/\lambda_{\text{targ}}$ and $f_i = \text{clip}(1 - 1/d_i, -1, 1)$ on the raw divisor (no $\max(\cdot, 1)$ floor; directions with $d_i < 1$ thus receive bounded sign-reversed amplification rather than attenuation) and for amplification $f_i = -\alpha$ unconditionally; angles are exact $\arccos$, no small-edit approximation. The linSAM/SAM magnitudes are the Taylor columns of the instrumented reruns (measured every 200 steps on the current training minibatch, train-mode BN), so the comparison is same-order-of-magnitude rather than identically sampled (§5.1, Table 10).

Table 10: Realized actuation strength, all arms (exact plant statistics over all 5000 steps per seed; late window $t \geq 2000$; n per arm). "fired (steps)" is the fraction of steps on which at least one direction fires; "fired (dir.)" is the per-direction fired fraction (the "fired" column of Table 9); the current-minibatch rank-1 control fires every step. $\|\mathbf{g}' - \mathbf{g}\|/\|\mathbf{g}\|$ is the fired-step mean for the eigenspace edits (amplification is unconditional) and the late-window mean for the linSAM/SAM corrections (Taylor-column probes every 200 steps on the current training minibatch, train-mode BN), a same-order-of-magnitude, not identically sampled, comparison. Paired $\Delta\lambda_{\text{ref}}$ vs. SGD (95% CIs, from Table 9) are reported for ordinary difference inference; equivalence results use separate 90% TOST CIs as described in Table 9 (current-minibatch control: §5.1; linSAM: §5.3; SAM: its ρ=0.05 radius cell); the bnmatched arm is not separately contrasted. The strongest eigenspace edit (k=20) actuates 3.7× less per step than linSAM and 5.7× less than SAM, yet only SAM separates downward from SGD.

arm	n	fired (st.)	fired (dir.)	$\|\mathbf{g}' - \mathbf{g}\|/\|\mathbf{g}\|$	angle	late	paired $\Delta\lambda_{\text{ref}}$
damp k=1	3	0.80	0.80	0.011	0.6°	0.002	+1.36 [−2.28, +5.00]
damp k=2	3	0.80	0.74	0.014	0.8°	0.003	+0.55 [−6.41, +7.51]
damp k=5	3	0.81	0.60	0.018	1.0°	0.007	+0.24 [−3.34, +3.82]
damp k=10	3	0.83	0.38	0.031	1.7°	0.017	+0.88 [−0.20, +1.97]
damp k=20	3	0.83	0.20	0.061	3.4°	0.035	+1.65 [−1.98, +5.27]
damp k=1 (current-minibatch)	5	1.00	1.00	0.012	0.7°	0.008	+1.16 [−0.69, +3.01]
amp k=5 (α=1)	5	1.00	—	0.041	2.3°	0.025	+2.01 [−0.74, +4.76]
amp k=20 (α=1)	5	1.00	—	0.081	4.4°	0.045	+4.88 [+0.79, +8.97]
linSAM (ρ=0.05)	5	—	—	0.18	—	0.23	+5.27 [+1.94, +8.61]
linSAM (bnmatched, τ=0)	5	—	—	0.18	—	0.24	—
SAM (ρ=0.05)	3	—	—	0.28	—	0.35	−6.29 [−8.73, −3.84]

D.14 ENGAGEMENT AND ELIGIBILITY DIAGNOSTICS: ARPACK-SENSED RANK-1 DAMPING

Sensing is the first suspect: an optimizer gating on mini-batch power iteration's $\hat{\lambda}$ reads $\approx 19 < 2/\eta$ and fires on $\leq 4\%$ of steps (≈ SGD), but direct ARPACK on a deterministic 512-sample batch gives $\lambda_{\text{ref}} \approx 36$ (median 35.7, 88% of steps above $2/\eta$): the ≈19 was a power-iteration false convergence (+45% mean underestimate): the sensing proxy was broken, not the regime. The legacy clamped arm using this sensor has late λ_{ref} 28.83 ± 1.67 (n=5; Fig. 13). *With the proxy fixed, the edit fires, and sharpness still shows no resolved shift.* Warm-started ARPACK sensing (k=2, every N=20 steps) without the $\max(\cdot, 1)$ floor gives damping divisor ≈ 1.8 and damping factor ≈ 0.44 every step; pre-registered eligibility holds on all n=5 seeds (divisor median $1.86 \geq 1.5$; window-mean damping $+0.39 > 0.3$; Ritz effective-force rate 100%), and the auxiliary fire-rate criterion over the substantive window too (98.5%, steps 1–4000; 83% over the full 5000, the late dip being engagement by design). Yet steady λ_{ref} medians $[37.2, 36.9, 39.7, 35.6, 36.3]$ overlap vanilla SGD (33.8 ± 3.7).

D.15 THE STOCHASTIC-REGIME AMPLIFICATION ARM AND CURRENT-MINIBATCH SENSING: FULL RESULTS

The amplification arm and the current-minibatch-sensing control: no statistically resolved effect in either direction at this power. Rank-1 amplification under the identical deterministic +

ARPACK protocol: α=0.5 (n=5) and α=1.0 (n=3) both have late λ_{ref} estimates near the SGD level (29.18 ± 1.44 and 28.73 ± 0.82 vs. SGD 29.21 ± 1.90), and neither yields a statistically resolved downward paired contrast at the stated sample size (accuracies unmoved, 0.513–0.514 vs. SGD's 0.512), with SAM's 22.63 ± 1.73 the only arm below. The α=1.0 arm is diagnostic: the momentum-regime analogue of the 2-D prediction, $2(1+\mu)/((1+\alpha)\eta)\approx19$, sits $\sim$10 units below the SGD band, far larger than the $\sim$1.9-unit CI, yet no separation occurs (intermittent dips to $\sim$15, no sustained shift; 0 NaN steps). The fixed-batch confound is excluded by intervention: perhaps the fixed-batch $\hat{v}_1$ (overlap 0.26) is simply the *wrong* direction, so we re-ran the accurate damping optimizer with sensing on the *current training minibatch itself* (ARPACK k=2 every N=20 steps; n=5). The edit engages even harder (sensed divisor median 2.35 vs. $\sim$1.5 fixed-batch, realized damping $+0.58$, firing on 100% of steps) and the outcome shows no resolved change: late reference $\lambda_{\text{ref}} = 30.37 \pm 1.05$, no closer to SAM, accuracies unmoved, the paired point estimate $+1.16$, 95% CI $[-0.69, +3.01]$, *above* SGD with no reduction in any seed. Single-direction action thus shows no statistically resolved effect in *either* sign: under bs-128 noise the training gradient aligns with *any* top-eigenvector estimate at $|\cos| \approx 0.008$, even the current minibatch's own, so a rank-1 edit removes only $\sim10^{-4}$ of the per-step gradient energy; only full SAM separates below SGD (whether its first-order expansion alone reproduces the shift is tested below; it does not), and the rank-k interpolation remains consistent with weak per-step actuation through k=20; as §5.2 shows, however, gradient share alone does not determine the long-horizon response.

E MATCHED-STEP X-SAM PROJECTOR ABLATION

We test whether adding the rank-1 top-eigenvector correction used by X-SAM (Duan et al., 2026) on top of SAM produces any further λ_{ref} reduction. We compare four optimizers on CIFAR-10/ResNet-18 in *deterministic* CUDA mode (n=10 seeds; default conv-backward is non-deterministic: two identical SAM runs diverge to a $\sim10^{-1}$ terminal param difference, producing $\sim$6-unit spurious λ_{ref} gaps that deterministic mode eliminates): SGD, SAM (ρ=0.05), *matched-step* X-SAM projector (Duan et al. Eq. 11: fixed α=0.2, no clamp, sign-gated, unconditional; fires every step), and a *clamped* variant ($d = \max(\hat{\lambda}/(2/\eta), 1)$, a conservative ablation). We measure λ_{ref} by ARPACK on a deterministic 512-sample batch (k=3, tol=10^{-6}); a mini-batch power-iteration estimate false-converges on this cross-entropy Hessian, a $+45\%$ mean underestimate (measured directly), so we report the ARPACK value.

Late-training (step $\geq$ 3000) reference λ_{ref}, n=10 mean$\pm$95% CI: SGD 31.1 ± 2.3, SAM 23.3 ± 1.6, matched-step X-SAM projector 23.9 ± 1.6, and the clamped projector variant 23.4 ± 1.6 (the three SAM-family CIs overlap; Fig. 20a). The paired difference $\Delta\lambda_{\text{ref}}$(X-SAM projector $-$ SAM) is $+0.59 \pm 2.55$ mid-training and $+0.62$ with 95% CI $[-0.61, +1.85]$ late (Fig. 20b). *Thus no rank-1 effect beyond SAM is resolved at this power; the interval excludes reductions larger than approximately* 0.61 *reference-sharpness units, while increases up to* 1.85 *remain compatible with the data.* We detect none of the substantial λ_{ref} reduction that Duan et al.'s "explicitly constrains λ_{ref}" framing would imply, and no generalization gain either: final accuracies (n=10) are SGD 0.802, SAM 0.817, matched-step X-SAM projector 0.814, clamped 0.815 (all within $\pm$0.02 CIs), so the X-SAM projector is not above SAM (it is 0.3 points below). Escape is *overwhelmingly* the amplifying SAM component: SAM descends $\sim$8 units below SGD in this CIFAR-10 setting (the CIFAR-100 three-way of §5.1 separates by $\sim$7) whether or not the rank-1 correction rides on top. (A rank-1-only ablation, no SAM base, was run on the MLP testbed of §4.3: it escapes past $2/\eta$ to that section's escape value, 261.8 ± 5.4.)

Fidelity to Duan et al.'s protocol. Our ablation is a controlled λ_{ref}-dynamics comparison, not a replication of their accuracy benchmark (their CIFAR-10/ResNet-18 setting: SGD base, cosine lr 0.1, weight decay 5×10^{-5}, batch 256, 200 epochs, flip/pad/crop augmentation, ρ=0.05, α=0.2, 3 runs; they report X-SAM 94.71 ± 0.20 vs. SAM 93.64 ± 0.30 there, and $+1.21\%$ average / $+2.08\%$ maximum across their CIFAR-10 backbones). Three deviations matter. (i) *Gradient normalization*: their Alg. 2 updates along the unit-normalized corrected gradient, which decouples step magnitude from the loss scale; our matched-step arm applies the sign-gated projector $P_t = I - \alpha\,\text{sign}(\langle g, v\rangle)\, vv^\top$ (their Eq. 11 correction) to the *raw* adversarial gradient so the effective step size η is matched across SGD/SAM/X-SAM projector; with normalized updates the correction is confounded with a per-step learning-rate change. (ii) *Schedule*: 5000 steps ($\sim$13 epochs) at constant lr 0.1, batch 128, weight decay 5×10^{-4}, no augmentation, deterministic CUDA, chosen for seed-level reproducibility (n=10)

rather than peak accuracy. (iii) *Eigenvector refresh*: warm-started power iteration (5 HVPs/step, BN statistics frozen during HVPs) rather than their intermittent estimation, giving a fresher v and hence if anything a stronger correction than their protocol. Our null result therefore bounds the rank-1 correction under matched-η dynamics; a full-protocol replication (cosine schedule, augmentation, normalized updates) remains the gap between our setting and theirs.

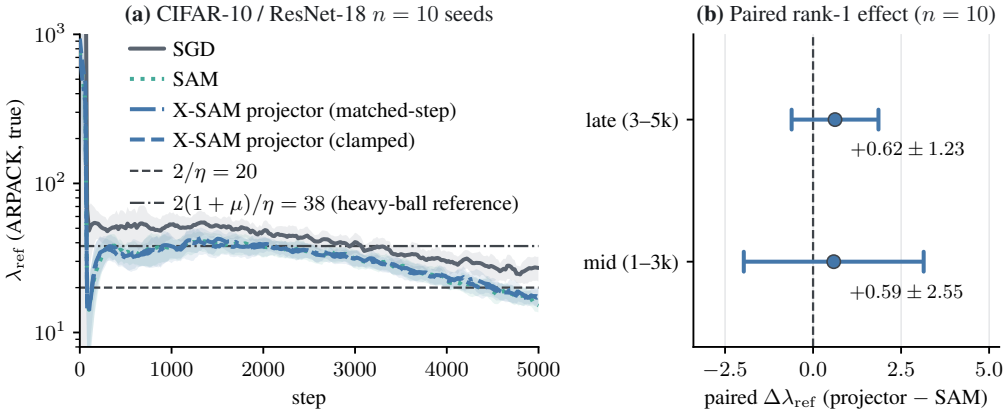

Figure 20: Matched-step X-SAM projector ablation on CIFAR-10/ResNet-18 (n=10 deterministic seeds, ARPACK reference λ_{ref}). **(a)** λ_{ref} trajectories (bold seed-mean, shaded 95% CI); SAM, the matched-step X-SAM projector, and the clamped projector variant all descend below SGD, with the two projector variants close to SAM at this power. **(b)** Paired $\Delta\lambda_{\text{ref}}$(X-SAM projector − SAM) at mid and late training: both CIs cross zero (late $+0.62$, 95% CI $[-0.61, +1.85]$; reductions beyond $\sim$0.61 units excluded), so no rank-1 effect on λ_{ref} beyond SAM is detected at this power. The rank-1-only ablation (no SAM base; MLP testbed, §4.3) escapes upward to $\lambda_{\text{raw}} = 261.8 \pm 5.4$.

F Hyperparameters and Compute Cost

Tables 11 and 12 summarize the hyperparameters of the main experiment families; all values are read from the run manifests and configs.

Run-family labels in Tables 11–12 preserve manifest names; interpret k>1 "damping" using the signed-rescaling semantics of §5.2.

Compute cost. All GPU experiments ran on a single RTX 5080 (16 GB); the unified MLP fixed-divisor grid was executed as a separate CPU round. The MLP families complete in seconds per run ($\sim$11–17 s, from manifests), while the unified fixed-divisor grid contains 18 cells $\times$ 5 model seeds; Table 11 reports its per-cell CPU cost. The ResNet-18 families dominate the GPU budget: deterministic ARPACK three-way (15 runs) $\sim$5.2 h; CIFAR-10 X-SAM projector ablation (40 runs) $\sim$33 h; ARPACK-sensing edit (5 runs) $\sim$4.6 h; clamp-free bridge $\sim$1.9 h; the cross-batch/rank-interpolation/long-recipe queue (38 runs) $\sim$40 h; the 200-epoch external-validity queue (15 runs) $\sim$31.8 h; batch-sharpness + linearized-SAM controls (6 runs) $\sim$2.2 h; the batch-sharpness-across-arms + ρ-sweep + n=5 top-up queue (44 runs) $\sim$24.5 h; and the regime-transition sweep (46 runs incl. pilot corners and one GPU-timeout re-run) $\sim$55 h, roughly 199 GPU-hours in total for the ResNet-18 runs reported here (an order-of-magnitude estimate from run manifests; per-run times vary with GPU sharing).

Table 11: Hyperparameters of the main experiment families (from run manifests). All ResNet-18 arms share momentum μ=0.9, weight decay 5×10^{-4}, batch size 128, 5000 steps, lr 0.1, and (unless noted) the deterministic-CUDA + ARPACK measurement protocol of §5.1.

family	configuration	per-run cost
MLP teacher–student (main)	n=1024, dim 20, difficulty (noise std) 0.3, teacher [128, 128], student [256, 256] tanh, full-batch GD, lr 0.05 ($2/\eta$=40), 3000 steps, data seed 1, 5 seeds	∼11 s
Rank-1 amplification control	sign-flip arms $\alpha \in \{0.25, 0.5, 1\}$ ($g' = g + \alpha(g^\top \hat{v})\hat{v}$, warm-started block power iteration, ∼3 HVPs/step, fires every step) + sgd/sam/rankk_k1 references, 5 seeds (§4.3)	∼21 s all arms
Unified fixed-divisor round	fixed-d damping arms $\eta \in \{0.025, 0.05, 0.1\} \times d \in \{1.25, 1.5, 2, 2.5, 3, 6.5\}$ ($g' = g - (1 - 1/d)(g^\top \hat{v})\hat{v}$, warm-started block-power-iteration sensing, fires every step, *no* $\hat{\lambda}$ feedback), 3000 steps, single CPU round (float32, torch 2.9.1; 18 cells × 5 model seeds 0–4, data seed 1) (§4.2, Fig. 1)	∼50 s per cell
Delayed-onset control	adaptive arm to its per-seed gate-opening step $T_{\text{gate}} \in \{188, \ldots, 206\}$ (mean 196.6), then fixed $d \in \{2.5, 3\}$; adaptive reference 5/5 stable (§4.2; Appendix C.5)	negligible (CPU)
MLP sweeps	base as main, 4000 steps; sharpness measured every 50 steps (warm-started power iteration, 150 iters, tol 10^{-4}; λ_{raw} and λ_c), snapshots every 10; grids: lr 0.01–0.2 (20× range), difficulty 0.1–1.2 (12×), width 32–256 (8×)	∼17 s
ResNet-18 CIFAR-100 three-way	SGD / rank-1 edit / SAM, n=5; λ_{ref} by ARPACK (k=3, tol 10^{-6}) on a deterministic 512-sample eval batch every 100 steps	15–31 min
X-SAM projector ablation (CIFAR-10)	SGD / SAM (ρ=0.05) / matched-step X-SAM projector (ρ=0.05, α=0.2, sign-gated, no clamp) / clamped variant (d= $\max(\hat{\lambda}/(2/\eta), 1)$), n=10; λ_{ref} by ARPACK (k=1, tol 10^{-6}, every 20 steps; every-100 k=3 backups retained)	29–75 min
Clamp-free bridge (CIFAR-100)	rank-1 clamp-free (edit factor clipped to $[-1, 1)$, fires every step), n=5; late-window $\hat{\lambda}_{\mathbf{g}}$ on the deterministic 512-sample eval batch (§C.4)	10–45 min
True-$\lambda_{\max}$ sensing arm	rank-1 with ARPACK sensing (k=2, tol 10^{-2}, every N=20 steps on batch 512), no clamp floor, n=5 (§5.1)	51–63 min
At-scale amplification	rank-1 *amplifying* arm with fixed edit factor α along the sensed direction, $g+\alpha(g^\top \hat{v})\hat{v}$; ARPACK sensing as above; α=0.5 n=5, α=1.0 n=3 (§5.1)	56–57 min

Table 12: Additional hyperparameters of the main experiment families (from run manifests).

family	configuration	per-run cost
Cross-batch sensing study	SGD instrumented with ARPACK top-3 on two extra fixed 512-batches + the current bs-128 minibatch, n=3 (§5.1)	26–27 min
Batch-sharpness measurement	SGD rerun of the cross-batch protocol, additionally logging the current-minibatch gradient Rayleigh quotient and an m=4-minibatch average $\mathbb{E}_B[\mathbf{g}_B^\top H_B \mathbf{g}_B / \|\mathbf{g}_B\|^2]$ at every outcome step, n=3 (trajectories bit-identical to the cross-batch study; §5.1)	24 min
Full-rank linearized SAM	$g \leftarrow g + (\rho/\|g\|)H_B g$ with ρ=0.05 matched to the SAM arm, H_B the current-minibatch Hessian (one HVP/step, no perturbed-point re-query), n=5 (instrumented rerun with the batch-sharpness and Taylor-residual logging; per-step trajectory verified bit-identical to the uninstrumented runs; §5.1)	30 min
Negative-sign control	$g \leftarrow g - (\rho/\|g\|)H_B g$ with ρ=0.05 matched (sign-flipped linSAM correction, one HVP/step, no re-query, BN protocol as linSAM, step-wise norm-matched), n=5 (§5.3)	55–113 min
Batch sharpness across arms	the central arms rerun with the cross-batch instrumentation and the m=4 batch-sharpness estimator: sam (n=3), linsam (n=5), amplify α=0.5 (n=3), rank-1 damping (n=3); sam/linsam additionally log the Taylor residual of the re-query (§5.1, Figs. 2, 17)	26–65 min
ρ-sweep, SAM vs. linSAM	$\rho \in \{0.005, 0.01, 0.02, 0.1\}$ per arm (ρ=0.05 from the rows above), n=3 per new ρ, with the Taylor-residual logging (§5.3)	15–20 min
Current-minibatch-sensing control	rank-1 damping; ARPACK requests the top 2 eigenpairs on the current minibatch every N=20 steps for sensing, while the edit itself is applied only along the leading eigenvector $\hat{v}_1$; n=5 (§5)	34 min
Rank interpolation	ARPACK-sensed rank-k edits: damping $k \in \{1, 2, 5, 10, 20\}$, amplification α=1 at $k \in \{5, 20\}$ (N=20; N=40 for k=20), n=3 per damping arm and n=5 for the amplifying arms; k=1 damping and the SGD baseline carry top-10 outcome-spectrum instrumentation (the amplify k=1 row of Table 9 is the stochastic-regime amplification arm above)	22–99 min
Long-training validation	three-way (sgd/rank1/sam) rerun, 12000 steps, pad/crop + flip augmentation, deterministic + ARPACK, n=2 seeds; the standard 100-epoch recipe (39,100 steps, cosine $0.1 \rightarrow 0$, augmentation, measurement every 5 epochs, n=3); and the full standard 200-epoch recipe (78,200 steps, same schedule/augmentation/measurement, n=5) (§5.1)	15–297 min
Regime-transition sweep	$B \in \{2048, 512, 128, 32\} \times \mu \in \{0, 0.9\} \times$ {SGD-sensing reference, rank-1 damping, rank-1 amplification α=0.5}, n=3; canonical protocol otherwise (lr 0.1 not rescaled with B; the damping arm's λ_{targ}=$2/\eta$ at every μ); the $(128, 0.9)$ anchor cell reuses the runs above; batch sharpness on m=4 fresh training-batch-size batches (§5.1)	17–119 min
ViT high-curvature control	lr grid $\{0.02, 0.05, 0.1, 0.2\}$ @ bs 128 (pilot, sgd, seed 0) + bs-512/lr-0.05 matched high-curvature point: 4 arms $\times$ 5 seeds (ARPACK-sensed rank-1 edit N=20; `linsam` via a chunked HVP, 128/chunk, LayerNorm-only path), deterministic protocol of §5.1	14–115 min

G DETAILED SCOPE AND LIMITATIONS

This appendix collects the full intervention- and measurement-specific scope statements summarized in Section 6. Our claims are deliberately intervention- and measurement-specific. Lemma 1 is an orbit-location statement conditional on the 2-D reduced recurrence (assumptions and gaps G1–G5: Appendix A); the corresponding period-2 orbit is repelling under the full 2-D Jacobian, so the theory does not explain attraction, basin size, or network-level convergence. The fixed-divisor grid therefore tests only the location prediction on trajectories that exhibit the corresponding sign-flip behavior; survival is non-monotone and numerically sensitive near the boundary. The full-batch experiments use a synthetic tanh teacher–student system with one data realization and multiple model seeds, so they establish a controlled mechanistic testbed rather than broad architectural generality.

In the stochastic experiments, λ_{ref} is a fixed-batch, eval-mode, BatchNorm-state-dependent diagnostic, and the λ_{step} estimator freezes BatchNorm statistics in eval mode; neither quantity is asserted to be a parameterization-invariant or exact train-mode stability variable for every optimizer arm. The $2/\eta$ target of the stochastic rescaling controller is an intervention convention inherited from the full-batch construction, not a theoretical momentum boundary, and the periodically refreshed current-minibatch-sourced control does not provide per-step same-minibatch oracle sensing. The rank-1 attenuation null and the wider raw-divisor rescaling results cover top-Hessian-eigenspace edits through $k \leq 20$ at one ResNet-18/CIFAR-100 operating point and one ViT-tiny cross-system control that does not isolate architecture, normalization, dataset, or protocol; together with the controlled full-batch MLP, the tested systems leave broader architectural and dataset generality open. For $k > 1$ the raw-divisor rescaling has no $\max(d_i, 1)$ floor (above-target directions are attenuated while below-target directions receive bounded sign-reversed amplification), and we provide neither a controlled fixed-strength attenuation sweep nor a norm-matched random-subspace baseline, so the null cannot be attributed uniquely to the use of Hessian eigendirections or to the realized strength and geometry of the tested edits. Several supporting cells use $n = 3$ seeds; a statistically unresolved difference is not evidence of exact equivalence, and equivalence conclusions are sample-size- and margin-specific. The batch-size sweep changes examples seen and summary windows and is therefore an exploratory operating-point map, not a causal estimate of a stochastic-to-deterministic transition. The canonical ResNet protocol is short, constant-lr, and augmentation-free; the 200-epoch experiment validates only the tested SGD/rank-1/SAM ordering under the standard recipe. The tested intervention families rescale gradient components or compare against SAM-derived corrections; we do not test direct gradient-based optimization of $\lambda_{\max}$ or other higher-order curvature controllers, so the conclusions concern the tested directional-gradient interventions, not the controllability of sharpness in general. Finally, the linSAM–SAM interpolation varies the entire finite re-query residual bundle and therefore does not isolate which nonlinear constituent causes the ResNet response; the ViT control further shows that the decomposition is system-dependent. The paper is diagnostic rather than algorithm-design work: we use successful controls such as SAM and full-batch amplification to delimit failure modes, but do not claim a new optimizer that improves accuracy or universally reduces sharpness.