OPTIMIZER SWITCHING AT THE EDGE OF STABILITY: THE ONSET OF LATE-PHASE SAM DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Late switching from gradient descent (GD) to sharpness-aware minimization (SAM) can recover much of the benefit of training with SAM throughout. Existing work has separately characterized SAM near the edge of stability, its oscillatory dynamics around sharp minima, and the empirical effectiveness of activating SAM late in training. What remains unclear is the finite-step mechanism that connects these phenomena: what changes immediately when SAM is activated at the *same parameter state* reached by late-stage GD?

We identify a switch-induced onset mechanism near the edge of stability. In an aligned quadratic model, we derive an exact same-state switching window in which GD remains loss-decreasing while a normalized-SAM step is loss-increasing. In a nonlinear sharp-valley model, we show that normalization turns the optimizer change into a finite $O(\eta\rho)$ normal displacement even as the incoming residual vanishes, with leading direction

$$-\frac{H^2 y}{\|Hy\|},$$

which preferentially filters the displacement toward high-curvature modes. We then establish an explicit invariant basin in the frozen quadratic normal model within which transverse components contract geometrically and the phase-corrected dynamics converge to the known principal SAM two-cycle.

Experiments recover the predicted local signatures under observable geometry. On ResNet-18/CIFAR-10, same-state interventions reveal finite SAM-specific forcing and a structured post-switch transition between sharp-side exit and flat-side stabilization. Near this transition, short phase-alternating transients recur across independently trained networks, although they do not quantitatively track the frozen quadratic shell. Together, these findings support a finite-step account of how late optimizer switching can initiate SAM's oscillatory dynamics near the edge of stability.

1 INTRODUCTION

Sharpness-aware minimization (SAM) evaluates its update gradient at a fixed-radius perturbation and often converges to flatter solutions than standard training (Foret et al., 2021). Its late-stage dynamics can be qualitatively different: normalized SAM may oscillate along a dominant-curvature direction and, in nonlinear valleys, induce motion toward lower curvature (Bartlett et al., 2023; Wen et al., 2023). Normalization is central because the perturbation retains length ρ even as the gradient becomes small (Dai et al., 2023).

This behavior is especially relevant near the edge of stability. GD often operates near $\eta\lambda_{\max} \approx 2$ (Damian et al., 2023), whereas the SAM stability boundary also depends on the gradient scale and ρ (Long & Bartlett, 2024). Meanwhile, switching from SGD to SAM only late in training can recover much of the benefit of full-SAM training (Zhou et al., 2025). What remains unclear is the immediate finite-step mechanism: *what changes when normalized SAM is activated at the same state reached by late-stage GD?*

We identify a three-stage onset mechanism. First, in an aligned quadratic model, the same state can be loss-decreasing under GD but loss-increasing under normalized SAM. Second, near a nonlinear

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.

sharp valley, the switch produces a finite $O(\eta\rho)$ normal displacement with leading term $-\eta\rho\,\frac{H^2 y}{\|Hy\|}$, which preferentially filters the displacement toward high-curvature modes. Third, the frozen quadratic normal dynamics admit an explicit invariant basin with geometric transverse contraction toward the known phase-alternating SAM two-cycle. This last step is conditional: the finite kick is not asserted to enter the basin from every initial state. Our analysis therefore characterizes *oscillatory onset*, rather than longer-time shell tracking or flatness selection.

Contributions. We make four contributions. First, we derive an exact *same-state switching* window in which GD decreases the loss while normalized SAM increases it, with a sufficient extension under partial spectral alignment. Second, we establish a *finite normalized kick*: as the incoming residual becomes $o(\rho)$, the first post-switch normal displacement remains $O(\eta\rho)$ with leading H^2 spectral filtering. Third, for the frozen quadratic map, we give an explicit invariant basin with geometric transverse contraction toward the known principal SAM two-cycle. Finally, controlled objectives recover the predicted local signatures, while ResNet-18/CIFAR-10 interventions reveal same-state reversal, finite SAM-specific forcing, a structured sharp/flat transition, and short phase-alternating transients across independent runs, but not quantitative frozen-shell tracking.

2 Related Work

SAM and its dynamics. SAM seeks parameters with low loss in a fixed-radius neighborhood (Foret et al., 2021) and has motivated variants that modify its geometry or computational cost (Kwon et al., 2021; Kim et al., 2022; Zhuang et al., 2022; Du et al., 2022; Liu et al., 2022). Its behavior has also been connected to implicit regularization, feature balancing, and robustness (Andriushchenko & Flammarion, 2022; Baek et al., 2024; Springer et al., 2024; Li et al., 2024). Our focus is narrower: the local dynamics created when normalized SAM is activated at a fixed late-stage state.

For normalized SAM, Bartlett et al. (2023) show that quadratic dynamics approach a leading-eigendirection two-cycle and that nonlinear effects can drive motion toward lower curvature. Related analyses connect SAM to leading-eigenvector alignment (Wen et al., 2023), emphasize the dynamical role of normalization (Dai et al., 2023), and study eigenvalue regulation, stochastic effects, and convergence of practical SAM (Agarwala & Dauphin, 2023; Si & Yun, 2023; Oikonomou & Loizou, 2025). We use the known quadratic two-cycle only as a frozen local reference; our question concerns the finite-step dynamics that arise when SAM is first switched on.

Edge of stability and optimizer switching. GD in neural networks often operates near the classical quadratic edge $\eta\lambda_{\max} \approx 2$ (Cohen et al., 2021; Arora et al., 2022; Wang et al., 2022; Damian et al., 2023). For SAM, Long & Bartlett (2024) derive a different local stability boundary that also depends on the gradient norm and perturbation radius ρ. We exploit the mismatch between these boundaries rather than proposing a new SAM edge: the same late-stage state can remain loss-decreasing under GD while becoming loss-increasing under normalized SAM.

The closest empirical motivation is Zhou et al. (2025), who show that switching from SGD to SAM late in training can recover much of the benefit of full-SAM training and describe longer-time escape toward flatter solutions. Their experiments use practical normalized SAM, whereas their theoretical dynamics simplify its gradient-dependent normalization. Our analysis addresses the complementary short-time question: what happens *at the switch itself*? We derive an exact same-state switching window, a finite $O(\eta\rho)$ curvature-filtered displacement, and a conditional frozen-model contraction toward the known principal two-cycle. Together, these results identify a finite-step mechanism for the onset of late-phase SAM dynamics. More broadly, recent work has explored state-dependent optimizer switching as an algorithmic strategy (Pandey et al., 2026); our focus instead is the local dynamics induced specifically by switching from GD to normalized SAM.

3 Setup

Let $L : \mathbb{R}^d \to \mathbb{R}$ be a smooth loss and $g(w) = \nabla L(w)$. For $g(w) \neq 0$, gradient descent (GD) and normalized sharpness-aware minimization (SAM) update

$$w^+_{\mathrm{GD}} = w - \eta g(w), \qquad w^+_{\mathrm{SAM}} = w - \eta \nabla L\left(w + \rho \frac{g(w)}{\|g(w)\|}\right),$$

where $\eta > 0$ is the outer step size and $\rho > 0$ is the SAM perturbation radius. We study a trajectory generated by GD up to a switching time and by normalized SAM thereafter, with both the parameter state and η unchanged at the switch. Since normalized SAM is undefined at $g(w) = 0$, all post-switch statements concern nonstationary states, including limits in which the incoming residual approaches zero.

Local sharp-valley model. For the nonlinear analysis, write $w = (x, y)$, where $x \in U \subset \mathbb{R}^p$ denotes tangential coordinates and $y \in \mathbb{R}^k$ normal coordinates. Locally, we assume

$$L(x, y) = \frac{1}{2} y^\top H(x) y + R(x, y), \tag{1}$$

where $H(x)$ is symmetric positive definite and $mI \preceq H(x) \preceq \bar{\lambda} I$. Its leading eigenvalue $\lambda_1(x) := \lambda_{\max}(H(x))$ is assumed simple and uniformly separated from the remaining spectrum; we write $v_1(x)$ for a corresponding unit eigenvector. We further assume that H has uniformly bounded first derivatives and that the remainder is locally cubic in the normal displacement, with

$$\|\nabla_x R(x, y)\| = O(\|y\|^3), \qquad \|\nabla_y R(x, y)\| = O(\|y\|^2),$$

uniformly in the local neighborhood. Precise constants and neighborhood conditions are given in the appendices.

We measure proximity to the classical GD stability threshold by $\delta(x) := 2 - \eta\lambda_1(x)$. Thus, $\delta(x) > 0$ places the dominant quadratic mode below the GD stability threshold, while $0 < \delta(x) \ll 1$ denotes the near-edge regime. When the base point is clear, we write simply λ_1 and δ.

4 Switching Changes the Stability Boundary

We first isolate the effect of changing the optimizer while holding the parameter state fixed. Consider

$$L(w) = \frac{1}{2}(w - w^\star)^\top H (w - w^\star), \qquad H \succeq 0,$$

and let v_1 be a unit eigenvector associated with $\lambda_1 = \lambda_{\max}(H)$. At an aligned state,

$$w - w^\star = x v_1, \qquad x > 0,$$

GD and normalized SAM induce

$$x^+_{\mathrm{GD}} = (1 - \eta\lambda_1)x, \qquad x^+_{\mathrm{SAM}} = (1 - \eta\lambda_1)x - \eta\rho\lambda_1.$$

Theorem 1 (Exact same-state switching window)**.** *At the same aligned quadratic state, GD strictly decreases the loss while normalized SAM strictly increases it if and only if*

$$\frac{2x}{x + \rho} < \eta\lambda_1 < 2. \tag{2}$$

The upper boundary is the classical GD threshold. The lower boundary is specific to normalized SAM and depends on the residual-to-perturbation ratio x/ρ. In particular,

$$\frac{2x}{x + \rho} \to 0 \qquad \text{as} \qquad \frac{x}{\rho} \to 0,$$

while the GD threshold remains 2. Thus the optimizer switch alone can reverse the sign of the one-step loss change, without any change in the parameter state.

Near the GD edge, write $\eta\lambda_1 = 2 - \delta$ and $0 < \delta < 2$. The normalized-SAM loss-increase condition becomes

$$x < \rho \, \frac{2 - \delta}{\delta}. \tag{3}$$

Hence, at fixed edge distance δ, the transition is controlled by the scale ratio x/ρ: normalization retains a perturbation of radius ρ even as the incoming residual shrinks.

Beyond exact alignment. The same mechanism persists without perfect leading-mode alignment, though the resulting condition is sufficient rather than exact. Write

$$g = \sum_i a_i v_i, \qquad \beta = \frac{|a_1|}{\|g\|}, \qquad r = \frac{\rho}{\|g\|}.$$

For a quadratic loss, the gradient at the SAM perturbation is exactly $g_{\mathrm{SAM}} = (I + rH)g$, so its relative leading-mode component satisfies

$$\beta_{\mathrm{SAM}}^2 = \frac{a_1^2(1 + r\lambda_1)^2}{\sum_i a_i^2(1 + r\lambda_i)^2} \geq \beta^2.$$

Combining this spectral reweighting with the exact quadratic loss-change identity gives the sufficient same-state window

$$\frac{2}{\beta^2(1 + \rho\lambda_1/\|g\|)} < \eta\lambda_1 < 2. \tag{4}$$

This window is nonempty whenever

$$\frac{\rho\lambda_1}{\|g\|} > \frac{1 - \beta^2}{\beta^2}.$$

Thus, a sufficiently large normalized perturbation relative to the incoming gradient scale can compensate for imperfect initial alignment. The exact quadratic identity, proofs, and a local nonlinear extension are given in Appendix B.

Relation to the SAM edge. The gradient-dependent SAM edge of Long & Bartlett (2024) characterizes the local stability boundary of SAM itself. Our question is different: we hold the parameter state fixed and compare the update maps before and after an optimizer switch. Theorem 1 therefore identifies an exact region in which the same state is loss-decreasing under GD but loss-increasing under practical normalized SAM. This same-state mismatch is the first step of the onset mechanism studied here.

The boundary comparison does not determine the subsequent trajectory. We next show that, near a nonlinear sharp valley, the optimizer switch also creates a finite curvature-weighted displacement even when the incoming normal residual is vanishingly small relative to ρ.

5 Normalization Produces a Finite Spectral Kick

We next ask what motion is produced immediately after the switching boundary of Section 4 is crossed. Return to the sharp-valley model (1), let $H_0 = H(x_0)$, and consider a nonstationary state (x_0, y_0) close to the valley floor. Because the tangential gradient is $O(\|y_0\|^2)$ whereas the normal gradient is $H_0 y_0 + O(\|y_0\|^2)$, the normalized gradient becomes asymptotically normal as $\|y_0\| \to 0$. The fixed perturbation radius ρ then produces a qualitatively different scale of motion from an ordinary gradient step.

Theorem 2 (Finite post-switch displacement)**.** *Under the local assumptions of Section 3, there exist uniform local constants such that, for sufficiently small $\rho > 0$ and nonzero y_0,*

$$y_1 = -\eta\rho\frac{H_0^2 y_0}{\|H_0 y_0\|} + (I - \eta H_0)y_0 + O(\eta\rho\|y_0\|) + O(\eta\rho^2) + O(\eta\|y_0\|^2). \tag{5}$$

In particular, for fixed sufficiently small ρ and $\|y_0\|/\rho \to 0$,

$$y_1 = -\eta\rho\frac{H_0^2 y_0}{\|H_0 y_0\|} + O(\eta\rho^2) + o(\eta\rho). \tag{6}$$

Moreover, if ρ and $\|y_0\|/\rho$ are uniformly sufficiently small, then there exist constants $c_-, c_+ > 0$, independent of y_0, such that

$$c_-\eta\rho \leq \|y_1\| \leq c_+\eta\rho. \tag{7}$$

Theorem 2 exposes the finite-step effect of normalization. As the incoming normal residual becomes negligible relative to ρ, the ordinary GD contribution $(I - \eta H_0)y_0$ vanishes at the $o(\eta\rho)$ scale, while

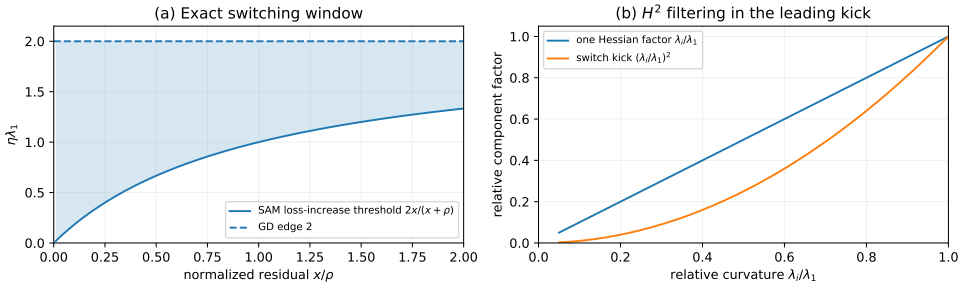

Figure 1: **Analytical predictions for the switching boundary and spectral kick.** (a) In the aligned quadratic model, the GD-decrease/SAM-increase window widens as the residual ratio x/ρ decreases. (b) Near the valley floor, the leading post-switch displacement applies an H^2 spectral filter, attenuating relative nonprincipal components by $(\lambda_i/\lambda_1)^2$.

the normalized-SAM contribution remains of order $\eta\rho$. Thus, the first post-switch displacement need not vanish with the residual that produced it.

The direction of this displacement also has a simple spectral structure. Let

$$H_0 v_i = \lambda_i v_i, \qquad y_0 = \sum_i y_{0,i} v_i.$$

The leading term in (6) multiplies the ith eigendirection by λ_i^2. Consequently, for initial directions lying in a fixed cone around the principal eigendirection, with $P_1 y_0 \neq 0$,

$$\frac{\|P_i y_1\|}{\|P_1 y_1\|} = \left(\frac{\lambda_i}{\lambda_1}\right)^2 \frac{\|P_i y_0\|}{\|P_1 y_0\|} + O(\rho) + o(1), \qquad i \geq 2. \tag{8}$$

Hence, the switch applies, to leading order, an H_0^2 spectral filter: components associated with smaller eigenvalues are attenuated relative to the principal component by the squared eigenvalue ratio. This is a local filtering statement, not a claim of instantaneous or global alignment of a neural-network trajectory.

Appendix C gives the complete uniform assumptions, remainder bounds, amplitude thresholds, and fixed-cone condition. It also records the joint limit $\rho \to 0$ and $\frac{\|y_0\|}{\rho} \to 0$, under which

$$y_1 = -\eta\rho \frac{H_0^2 y_0}{\|H_0 y_0\|} + o(\eta\rho). \tag{9}$$

Together, Theorems 1 and 2 establish the finite-step bridge at the optimizer switch. The first identifies states whose one-step loss behavior changes when GD is replaced by normalized SAM; the second shows that the same switch generates an $O(\eta\rho)$, curvature-weighted normal displacement. What remains is to understand whether this displacement can organize into an oscillatory local dynamics.

6 From the Kick to Oscillatory Onset

The finite displacement of Section 5 characterizes the immediate effect of activating normalized SAM, but not the subsequent trajectory. To obtain a tractable local closure, we freeze the tangential coordinate x, neglect the higher-order remainder R, and study the exact normalized-SAM map for the resulting quadratic normal model:

$$F(y) = (I - \eta H)y - \eta\rho \frac{H^2 y}{\|Hy\|}. \tag{10}$$

Let $\eta\lambda_1 = 2 - \delta$ $(0 < \delta < 1)$, where λ_1 is the simple leading eigenvalue of H. Along its eigenvector v_1, the map has the exact phase-alternating two-cycle $\{cv_1, -cv_1\}$ with

$$c = \frac{\eta\lambda_1\rho}{2 - \eta\lambda_1} = \frac{(2-\delta)\rho}{\delta}. \tag{11}$$

This two-cycle and its radius are known (Bartlett et al., 2023). We use it as the frozen oscillatory reference state and establish an explicit invariant basin in which the full normal map contracts toward it. To state the result, choose $\sigma_0 \in \{-1, 1\}$ so that the principal component of $\sigma_0 y_0$ is positive, and define the phase-corrected iterates

$$\widehat{y}_t = (-1)^t \sigma_0 y_t = \alpha_t v_1 + z_t, \qquad z_t \perp v_1, \qquad \alpha_t > 0.$$

Let

$$g_t = 1 - \frac{\alpha_t}{c}, \qquad r_t = \frac{\|z_t\|}{\alpha_t}.$$

Writing $q_i = \lambda_i/\lambda_1$, define

$$q_* = \frac{\lambda_2}{\lambda_1}, \qquad \kappa = \frac{q_*^2}{2},$$

and

$$q_{\mathrm{fr}} = \max_{i \geq 2} \max\left\{q_i^2, \left|1 - (2-\delta)q_i - \delta q_i^2\right|\right\} < 1. \tag{12}$$

Theorem 3 (Frozen principal-shell attraction)**.** *Suppose the initial state has nonzero principal projection and*

$$g_0 \geq \kappa r_0^2, \qquad q_{\mathrm{fr}}^2 \leq 1 - \delta. \tag{13}$$

Then, the basin condition is forward invariant and, for every $t \geq 0$,

$$r_t \leq q_{\mathrm{fr}}^t r_0, \qquad g_t \geq \kappa r_t^2 \geq 0. \tag{14}$$

Moreover,

$$\widehat{y}_t \longrightarrow c v_1,$$

or equivalently,

$$y_t \longrightarrow (-1)^t \sigma_0 c v_1.$$

Therefore, the frozen iterates converge to the known principal two-cycle.

The theorem provides a conditional local closure for the post-switch dynamics. Transverse components contract geometrically relative to the principal mode, while the radial slack that defines the basin is preserved. Appendix D gives the exact component recurrences, projective contraction argument, invariant-basin proof, and explicit convergence bounds.

The edge distance δ also exposes a limitation of the frozen model. From (11),

$$c = \frac{(2-\delta)\rho}{\delta} \sim \frac{2\rho}{\delta} \qquad (\delta \downarrow 0),$$

while the radial relaxation time scales as $1/\delta$. Approaching the exact GD edge therefore enlarges the frozen shell while slowing radial relaxation; the limit $\delta \downarrow 0$ is not uniformly benign.

Theorem 3 provides the local oscillatory closure of the switching mechanism. Within the explicit basin (13), the frozen dynamics contract toward the principal SAM two-cycle, showing how the finite normalized displacement of Section 5 can organize into phase-alternating motion. Basin capture and tracking under evolving curvature constitute the subsequent dynamical regime beyond this frozen analysis.

7 Experiments

The theory makes predictions at different levels of geometric specificity. Quantities such as the local normal coordinates and the frozen principal shell are directly observable only in controlled models, whereas same-state loss reversal, finite SAM-specific forcing, and phase-alternating motion admit neural-network proxies. We therefore use controlled objectives for quantitative validation and neural networks to test which *onset signatures* remain visible beyond the idealized local geometry.

Table 1: **Controlled validation of the local mechanism.** The frozen quadratic case provides the reference geometry; subsequent settings relax its assumptions. Shell error measures deviation from the predicted phase-alternating principal shell.

Controlled regime	Shell error
Frozen quadratic	5.7×10^{-6}
Moving curvature	8.0×10^{-3}
Fast eigendirection rotation	2.4×10^{-2}
Small eigengap	5.2×10^{-2}
Nonlinear remainder	1.38×10^{-1}

Controlled models recover the local mechanism. We first use synthetic sharp-valley objectives with known normal coordinates, Hessian spectrum, and local geometry. The suite progressively departs from the frozen quadratic reference through curvature evolution, eigendirection rotation, reduced spectral separation, and nonlinear remainder terms.

Table 1 summarizes the resulting hierarchy. Same-state interventions recover the GD-decrease/SAM-increase regime of Theorem 1, with switching events observed in 75% of the tested controlled configurations. The first post-switch displacement exhibits the predicted dependence on ρ and H^2 spectral filtering. Most sharply, the frozen quadratic model tracks the predicted phase-alternating shell with relative error 5.7×10^{-6}. Agreement remains close under slowly moving curvature (8.0×10^{-3}), while larger deviations appear when the frozen assumptions are stressed through fast eigendirection rotation (2.4×10^{-2}), a small eigengap (5.2×10^{-2}), or a nonlinear remainder (1.38×10^{-1}).

These experiments serve two purposes: they quantitatively recover the switching, finite-displacement, and frozen-shell predictions when the corresponding geometry is observable, and they show how agreement changes as the dynamics depart from the frozen reference.

Same-state neural interventions. We next train ResNet-18 on CIFAR-10 with SGD and intervene at late checkpoints by branching from the *same parameter state* into a GD step and a normalized-SAM step with the same outer step size. Local measurements use a fixed 2,048-example diagnostic subset, and leading Hessian eigenpairs are retained only when predeclared residual and eigengap criteria are satisfied.

The most direct theoretical signature remains visible in this setting. We repeatedly observe states whose diagnostic loss decreases under GD but increases under normalized SAM. In the seed-0 sweep, such crossings occur 13, 5, and 13 times at epochs 160, 180, and 190, respectively. Since both updates start from identical parameters, these interventions isolate the effect of changing the optimizer before the two branches have visited different regions of the landscape.

We also isolate the SAM-specific component of the update,

$$\Delta w_{\text{excess}} = \Delta w_{\text{SAM}} - \Delta w_{\text{GD}}. \tag{15}$$

Its magnitude varies systematically with ρ: log–log fits of $\|\Delta w_{\text{excess}}\|$ against ρ give $R^2 > 0.98$ at all three checkpoints. The fitted slopes vary across checkpoints, so we do not interpret them as evidence for a universal unit exponent. Rather, the experiment provides a network-level counterpart of the finite post-switch forcing predicted by Theorem 2.

A structured post-switch regime boundary. The theory does not predict that every finite kick enters the frozen principal-shell basin. We therefore examine the subsequent neural transient without assuming shell capture. At each post-switch state, we track $\delta_t = 2 - \eta\lambda_{1,t}$ and classify $\delta_t \le 0$ as a *sharp-side exit*, $\delta_t \ge 1$ as a *flat-side escape*, and $0 < \delta_t < 1$ as the *controlled-edge region*.

Figure 2 summarizes 1,725 interventions from three late checkpoints over a dense grid of 23 initial edge distances and 25 SAM radii. The dominant empirical feature is a structured transition between sharp-side and flat-side outcomes rather than a single critical radius. Of the 1,725 interventions, 1,140 resolve toward the flat side, 580 through the sharp side, and five fail the spectral validity gate. Sixty-three remain in the controlled-edge region for at least two consecutive states. The transition shifts substantially across checkpoints, arguing against a universal absolute-ρ threshold.

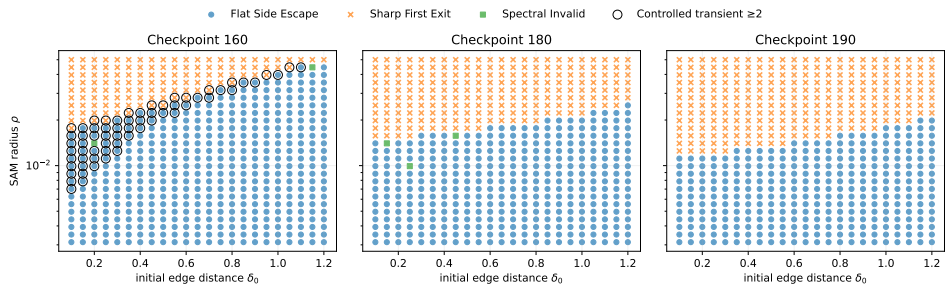

Figure 2: **Post-switch outcomes organize around a structured regime boundary.** ResNet-18/CIFAR-10, seed 0. Circles denote flat-side escape, crosses denote sharp-side exit, and squares denote cells that fail the spectral validity gate; open rings mark interventions with at least two consecutive controlled-edge states. The transition shifts across checkpoints rather than concentrating at a universal absolute value of ρ.

Table 2: **The post-switch transition recurs across independently trained networks.** F and S denote flat-side escape and sharp-side exit as ρ increases. Shown are seed/checkpoint groups that bracket a transition; other sampled groups remain on one side of the tested radius window.

Seed	Epoch	Outcomes	Max. controlled	Phase alt.
1	190	F F S S	2	1.00
2	190	F S S S	2	1.00
3	160	F S S S	4	1.00

Oscillatory onset near the regime boundary. Trajectories near the empirical transition exhibit a sharper dynamical signature. At epoch 160 with $\delta_0 = 0.15$, the intervention $\rho = 0.01768$ remains in the controlled-edge region for five states with persistent phase alternation before resolving toward the flat side, whereas the nearby choice $\rho = 0.01984$ exits through the sharp side after two updates. Thus small changes in ρ can send the same late-training state to opposite sides of the post-switch transition.

The phase-alternating transient captures the oscillatory onset predicted by the local theory, while also marking the quantitative limit of the frozen approximation: along the flat-resolving trajectory, the shell-proxy error is 0.72–0.96. The neural dynamics therefore reproduce the oscillatory signature near the transition without quantitatively tracking the frozen principal shell.

The transition recurs across independent training runs. We finally repeat a preselected intervention grid on four independently trained networks. All 48 interventions yield valid outcome summaries: 40 resolve through the sharp side and eight through the flat side. In 25 cases, subsequent spectral estimation terminates only after the predeclared outcome has already been resolved, leaving the recorded outcome unchanged.

The transition appears across multiple independently trained networks, although its absolute location varies. As ρ increases, seed 1 at epoch 190 exhibits F–F–S–S, while seed 2 at epoch 190 and seed 3 at epoch 160 exhibit F–S–S–S. Controlled-edge transients remain short, with the longest lasting five states. Thus the recurring feature is the *transition itself*, rather than a seed-independent critical radius.

From local theory to neural dynamics. The experiments reveal a clear hierarchy of transfer. The most local predictions are the most robust: at the same parameter state, switching optimizers can reverse the one-step loss behavior and generate a finite SAM-specific response. Post-switch trajectories further exhibit a structured sharp/flat transition, with short phase-alternating transients near its boundary. By contrast, quantitative principal-shell tracking is specific to the frozen local geometry and is not observed in the neural benchmark. This hierarchy is consistent with the theoretical structure: the switching and finite-kick results govern the immediate dynamics, whereas shell attraction additionally requires frozen curvature and entry into an explicit invariant basin. Thus the neural results capture the predicted signatures of oscillatory onset while delineating where the frozen approximation ceases to be quantitative.

Experimental protocol and reproducibility. We use fixed, predeclared experimental configurations throughout. Neural experiments train ResNet-18 on CIFAR-10 for 200 epochs with SGD, batch size 128, initial learning rate 0.1 with cosine decay, momentum 0.9, and weight decay 5×10^{-4}. Same-state interventions use checkpoints at epochs 160, 180, and 190 and a fixed 2,048-example diagnostic subset. Leading Hessian eigenpairs are estimated by Lanczos iterations with predefined residual and eigengap validity gates. Appendix 7 specifies the intervention grids, classification rules, seeds, spectral settings, and metric definitions; the supplementary material provides the code and configurations used to reproduce the reported results.

8 Discussion and Limitations

Our results provide a local finite-step bridge from optimizer switching to oscillatory onset. At the same state, normalized SAM can reverse the one-step loss behavior of GD; near a sharp valley, normalization produces an $O(\eta\rho)$ displacement with an H^2 spectral bias; and, within an explicit invariant basin, the frozen quadratic dynamics contract toward the known phase-alternating principal two-cycle. Together, these results characterize the onset of late-phase SAM dynamics rather than the full subsequent trajectory.

The scope of this bridge is deliberately local. The exact switching window assumes quadratic alignment, with only a sufficient extension beyond alignment. The finite-kick expansion is local and does not imply globally exact spectral alignment. The oscillatory result freezes curvature and eigendirections and, crucially, does not establish either automatic capture into its invariant basin or tracking of the shell under evolving geometry.

The neural experiments reflect this distinction. Across checkpoints and independent training runs, we observe same-state loss reversals, finite SAM-specific responses, and a structured sharp/flat boundary with short phase-alternating transients. Quantitative frozen-shell tracking, however, is not observed: shell-proxy errors remain large and the transition location varies across checkpoints and seeds. The neural evidence therefore supports the *onset signatures* of the mechanism, not sustained principal-shell tracking or a universal critical perturbation radius. A natural next step is to analyze capture and tracking when curvature and the leading eigendirection evolve with the SAM oscillation, and to incorporate stochastic gradients. Such extensions are needed to connect the local onset mechanism to longer-time sharpness reduction, flatness selection, and generalization.

9 Conclusion

We studied what changes immediately when normalized SAM is activated at a parameter state produced by late-stage GD. Even without moving the parameters, changing the update rule can reverse the one-step loss behavior: a state that remains below the GD stability threshold can lie beyond the corresponding normalized-SAM loss-increase boundary. Near a sharp valley, normalization further converts a small incoming residual into a finite $O(\eta\rho)$, curvature-weighted displacement with a leading H^2 spectral bias. In the frozen quadratic model, an explicit invariant basin then provides a regime in which these post-switch dynamics organize toward the known phase-alternating principal two-cycle. These results connect three dynamical scales that are usually considered separately: the optimizer-dependent stability boundary at the switching state, the finite displacement generated by normalized SAM, and the emergence of phase-alternating motion around a dominant-curvature direction. In this view, late optimizer switching is itself a dynamical intervention: it changes not only the update magnitude, but also the local stability and spectral structure governing the next stage of the trajectory.

Controlled experiments recover these predictions when the relevant local geometry is directly observable. Neural-network interventions reveal the corresponding onset signatures, same-state loss reversal, finite SAM-specific forcing, and short phase-alternating transients near a structured post-switch regime boundary, while delineating the quantitative limit of the frozen approximation through the absence of principal-shell tracking. Together, the theory and experiments identify a finite-step mechanism by which switching optimizers can initiate late-phase SAM dynamics. Extending this mechanism from onset to sustained dynamics requires understanding basin capture, evolving curvature and eigendirections, and stochastic gradients. Establishing that bridge would connect the local switching mechanism developed here to longer-time sharpness reduction and flatness selection.

Reproducibility Statement

Complete proofs of the switching, finite-kick, and frozen-shell results are provided in the appendices, including explicit constants and invariant-basin conditions. The experimental protocol specifies the diagnostic objective, spectral validity gates, intervention quantities, and the distinction between controlled-coordinate measurements and neural observable proxies. Frozen experiment configurations and anonymized code are provided in the supplementary material to reproduce the reported synthetic and mechanism experiments. An anonymous reproducibility repository is available at `https://anonymous.4open.science/r/late-sam-switching-E550/`.

AI Use Statement

Generative AI tools were used as research-assistance tools for proof checking, code implementation and refactoring, grammar checking, and manuscript editing. All claims, mathematical proofs, citations, code, and numerical results were reviewed and verified by the authors, who take responsibility for the final submission and all AI-assisted artifacts.

References

Atish Agarwala and Yann Dauphin. SAM operates far from home: Eigenvalue regularization as a dynamical phenomenon. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 152–168, 2023.

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, 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.

Christina Baek, Zico Kolter, and Aditi Raghunathan. Why is SAM robust to label noise? In *International Conference on Learning Representations*, 2024.

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.

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*, 2021.

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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*, 2023.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Y. F. Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022.

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

Minyoung Kim, Da Li, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Bingcong Li, Liang Zhang, and Niao He. Implicit regularization of sharpness-aware minimization for scale-invariant problems. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

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.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *International Conference on Learning Representations*, 2025.

Alok Kumar Pandey, Umang Chaturvedi, Aatish Rana, and Gopi Krishna Nedanuri. AOS: Adaptive optimizer switching via training-state signals for faster convergence and better generalization, 2026.

Dongkuk Si and Chulhee Yun. Practical sharpness-aware minimization cannot converge all the way to optima. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jacob Springer, Vaishnavh Nagarajan, and Aditi Raghunathan. Sharpness-aware minimization enhances feature quality via balanced learning. In *International Conference on Learning Representations*, 2024.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *International Conference on Learning Representations*, 2023.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, 2025.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022.

A Additional Preliminaries and Scope

Normalized SAM is defined at nonstationary points, $\nabla L(w) \neq 0$, since its perturbation uses the direction $\nabla L(w)/\|\nabla L(w)\|$. Accordingly, all switching results below assume a nonzero incoming gradient, and limits toward the valley floor are taken through nonstationary states.

The nonlinear analysis is local. We work in a neighborhood admitting the sharp-valley representation

$$L(x, y) = \frac{1}{2} y^\top H(x) y + R(x, y),$$

with all Taylor segments contained in this neighborhood and all remainder constants uniform on the corresponding local set.

The appendices establish the three components of the analysis: the same-state switching boundary, the finite normalized post-switch displacement, and the frozen phase-alternating dynamics. The first two characterize the immediate effect of switching to normalized SAM; the third gives an explicit invariant regime in which the frozen dynamics contract toward the principal two-cycle. Capture into this regime and tracking under evolving curvature belong to the subsequent dynamical regime beyond the frozen analysis.

B Same-State Switching: Detailed Results and Proofs

This appendix develops the same-state switching results of Section 4. We first give the exact aligned quadratic calculation, then extend the analysis to nonaligned positive-semidefinite quadratics.

A *same-state switch* compares GD and normalized-SAM updates initialized at the identical parameter vector. We characterize when the GD step strictly decreases the loss while the normalized-SAM step strictly increases it. This is a local one-step property that identifies the immediate dynamical change induced by switching optimizers, the subsequent SAM trajectory is analyzed separately.

B.1 Exact Aligned Quadratic Transition

Proof of Theorem 1. Since

$$w - w^\star = x v_1,$$

we have

$$\nabla L(w) = H(x v_1) = \lambda_1 x v_1.$$

A GD step therefore gives

$$w^+_{\mathrm{GD}} - w^\star = (1 - \eta\lambda_1) x v_1,$$

and hence,

$$L(w^+_{\mathrm{GD}}) - L(w^\star) = \frac{\lambda_1}{2}(1 - \eta\lambda_1)^2 x^2.$$

Since

$$L(w) - L(w^\star) = \frac{\lambda_1}{2} x^2,$$

strict one-step GD descent is equivalent to

$$|1 - \eta\lambda_1| < 1,$$

or

$$0 < \eta\lambda_1 < 2.$$

For normalized SAM,

$$\frac{\nabla L(w)}{\|\nabla L(w)\|} = \operatorname{sgn}(x) v_1.$$

The inner point is therefore

$$w + \rho \operatorname{sgn}(x) v_1,$$

at which

$$\nabla L(w + \rho \operatorname{sgn}(x) v_1) = \lambda_1 \big(x + \rho \operatorname{sgn}(x)\big) v_1.$$

Thus,

$$x^+_{\mathrm{SAM}} = (1 - \eta\lambda_1)x - \eta\rho\lambda_1 \operatorname{sgn}(x).$$

Set

$$a := \eta\lambda_1, \qquad r := |x|.$$

Multiplying the last display by $\operatorname{sgn}(x)$ gives

$$\operatorname{sgn}(x)x^+_{\mathrm{SAM}} = r - a(r + \rho).$$

Since the quadratic loss depends only on the squared amplitude, normalized SAM strictly increases the loss if and only if

$$|r - a(r + \rho)| > r.$$

The branch $r - a(r + \rho) > r$ is impossible because $a, \rho > 0$. Hence, loss increase is equivalent to

$$r - a(r + \rho) < -r,$$

or

$$a(r + \rho) > 2r.$$

Therefore, normalized SAM is one-step loss-increasing exactly when

$$\eta\lambda_1 > \frac{2r}{r + \rho}.$$

Intersecting this condition with the exact GD descent condition $\eta\lambda_1 < 2$ proves the result. □

B.2 Partial Alignment in a Positive-Semidefinite Quadratic

We next remove exact alignment while retaining an exact quadratic model. The resulting loss-change identity also makes explicit how normalization reweights the spectral components of the incoming gradient.

Lemma 1 (Exact quadratic loss-change formula)**.** *Suppose*

$$L(w + u) = L(w) + g^\top u + \frac{1}{2}u^\top Hu, \qquad H \succeq 0, \qquad g \neq 0.$$

Let

$$Hv_i = \lambda_i v_i, \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq 0,$$

and expand

$$g = \sum_i a_i v_i.$$

Define

$$s := \|g\|, \qquad r := \frac{\rho}{s}.$$

For the normalized-SAM update

$$w^+ = w - \eta\nabla L\Big(w + \rho\frac{g}{s}\Big),$$

the exact one-step loss change is

$$L(w^+) - L(w) = -\eta \sum_i a_i^2(1 + r\lambda_i)\left[1 - \frac{\eta\lambda_i(1 + r\lambda_i)}{2}\right].$$

Proof. The gradient at the SAM inner point is

$$q := \nabla L\Big(w + \rho\frac{g}{s}\Big) = g + H\left(\rho\frac{g}{s}\right) = (I + rH)g.$$

Hence,

$$q = \sum_i a_i(1 + r\lambda_i)v_i.$$

The SAM displacement is

$$d = -\eta q.$$

Because the objective is exactly quadratic,

$$L(w+d) - L(w) = g^\top d + \frac{1}{2} d^\top H d.$$

Therefore,

$$L(w^+) - L(w) = -\eta g^\top q + \frac{\eta^2}{2} q^\top H q.$$

Using the eigen-expansion,

$$g^\top q = \sum_i a_i^2 (1 + r\lambda_i)$$

and

$$q^\top H q = \sum_i \lambda_i a_i^2 (1 + r\lambda_i)^2.$$

Factoring the resulting expression proves the formula. □

Lemma 2 (Spectral amplification of the perturbed gradient)**.** *Under the assumptions of Lemma 1, choose a leading eigenvector v_1 and define*

$$\beta := \frac{|a_1|}{s}.$$

Let

$$\beta_{\mathrm{SAM}} := \frac{|\langle \nabla L(w + \rho g/s), v_1 \rangle|}{\|\nabla L(w + \rho g/s)\|}.$$

Then,

$$\beta_{\mathrm{SAM}} \geq \beta.$$

More precisely,

$$\beta_{\mathrm{SAM}}^2 \geq \frac{\beta^2 (1 + r\lambda_1)^2}{\beta^2 (1 + r\lambda_1)^2 + (1 - \beta^2)(1 + r\lambda_2)^2}.$$

The inequality $\beta_{\mathrm{SAM}} \geq \beta$ does not require simplicity of the leading eigenvalue. Strict amplification relative to the chosen v_1 occurs whenever $0 < \beta < 1$, $\rho > 0$, and some nonzero component of g lies in an eigenspace with eigenvalue strictly smaller than λ_1.

Proof. The perturbed gradient is

$$q = \sum_i a_i (1 + r\lambda_i) v_i.$$

Hence,

$$\beta_{\mathrm{SAM}}^2 = \frac{a_1^2 (1 + r\lambda_1)^2}{\sum_i a_i^2 (1 + r\lambda_i)^2}.$$

Since

$$\lambda_i \leq \lambda_1,$$

we have

$$(1 + r\lambda_i)^2 \leq (1 + r\lambda_1)^2.$$

Therefore,

$$\sum_i a_i^2 (1 + r\lambda_i)^2 \leq s^2 (1 + r\lambda_1)^2,$$

which gives

$$\beta_{\mathrm{SAM}}^2 \geq \frac{a_1^2}{s^2} = \beta^2.$$

For the sharper estimate, use

$$\lambda_i \leq \lambda_2, \qquad i \geq 2.$$

Then,

$$\sum_{i \geq 2} a_i^2 (1 + r\lambda_i)^2 \leq (1 + r\lambda_2)^2 \sum_{i \geq 2} a_i^2 = s^2 (1 - \beta^2)(1 + r\lambda_2)^2.$$

Substitution into the exact expression for β_{SAM}^2 proves the stated bound. □

Proof of the partial-alignment claim in Section 4. Under the assumptions of Lemma 1, if

$$\beta^2 \eta \lambda_1 \left(1 + \frac{\rho \lambda_1}{\|g\|}\right) > 2,$$

then one normalized-SAM step strictly increases the quadratic loss.

Consequently, a sufficient same-state GD-decreasing/SAM-increasing window is

$$\frac{2}{\beta^2 \left(1 + \rho \lambda_1 / \|g\|\right)} < \eta \lambda_1 < 2.$$

Let

$$B := \sum_i a_i^2 (1 + r\lambda_i) \left[\frac{\eta \lambda_i (1 + r\lambda_i)}{2} - 1\right].$$

By Lemma 1,

$$L(w^+) - L(w) = \eta B.$$

For every i,

$$\frac{\eta \lambda_i (1 + r\lambda_i)}{2} - 1 \geq -1.$$

Therefore,

$$\begin{aligned} B &\geq a_1^2 (1 + r\lambda_1) \frac{\eta \lambda_1 (1 + r\lambda_1)}{2} - \sum_i a_i^2 (1 + r\lambda_i) \\ &\geq a_1^2 (1 + r\lambda_1) \frac{\eta \lambda_1 (1 + r\lambda_1)}{2} - s^2 (1 + r\lambda_1) \\ &= s^2 (1 + r\lambda_1) \left[\frac{\beta^2 \eta \lambda_1 (1 + r\lambda_1)}{2} - 1\right]. \end{aligned}$$

Hence, $B > 0$ under the stated sufficient condition. For a positive-semidefinite quadratic, the ordinary GD update satisfies strict one-step descent whenever

$$0 < \eta \lambda_1 < 2.$$

Intersecting the two sufficient conditions gives the stated switching window.

Suppose $\beta > 0$. The sufficient window is nonempty whenever

$$\frac{\rho \lambda_1}{\|g\|} > \frac{1 - \beta^2}{\beta^2}.$$

Thus, for fixed nonzero principal alignment and fixed $\rho \lambda_1$, decreasing the residual gradient scale can open a sufficient GD-decreasing/SAM-increasing window.

The sufficient lower endpoint lies below the GD upper endpoint 2 exactly when

$$\frac{2}{\beta^2 (1 + \rho \lambda_1 / \|g\|)} < 2.$$

This is equivalent to

$$\beta^2 \left(1 + \frac{\rho \lambda_1}{\|g\|}\right) > 1,$$

or

$$\frac{\rho \lambda_1}{\|g\|} > \frac{1 - \beta^2}{\beta^2}.$$

□

C Finite Post-Switch Displacement and Spectral Filtering

This appendix develops the finite post-switch displacement analysis of Section 5. In a local multidimensional sharp-valley model, we first characterize the normalized SAM perturbation direction, then derive the resulting finite-amplitude normal displacement and its leading spectral filtering effect.

The finite-amplitude result holds uniformly over nonzero incoming normal directions and requires no prior alignment with the leading eigendirection. A principal-cone condition enters only in the finer modal comparison that quantifies how the post-switch displacement is filtered toward the leading curvature mode.

C.1 Local Sharp-Valley Geometry

Assumption 4 (Local sharp valley)**.** *Let*

$$L(x,y) = \frac{1}{2} y^\top H(x) y + R(x,y), \qquad x \in U \subset \mathbb{R}^p, \qquad y \in \mathbb{R}^k.$$

Fix $x_0 \in U$. Assume that on a neighborhood of $(x_0, 0)$:

(i)
$$mI \preceq H(x) \preceq \bar{\lambda} I;$$

(ii) H is differentiable and
$$\|DH(x)\|_{\mathrm{bil}} \le A_H.$$
Consequently, for any x, x' whose connecting segment remains inside this neighborhood,
$$\|H(x) - H(x')\|_{\mathrm{op}} \le A_H \|x - x'\|;$$

(iii)
$$\|\nabla_x R(x,y)\| \le C_x \|y\|^3, \qquad \|\nabla_y R(x,y)\| \le C_y \|y\|^2.$$

Let $H_0 := H(x_0)$ have eigenvalues
$$\lambda_1 > \lambda_2 \ge \cdots \ge \lambda_k > 0$$
with corresponding orthonormal eigenvectors $v_1, \ldots, v_k$.

C.2 Direction of the Normalized Inner Perturbation

Lemma 3 (Pre-switch normalized-gradient direction)**.** *Under Assumption 4, let*
$$g_0 := \nabla L(x_0, y_0) = (g_{x,0}, g_{y,0}), \qquad y_0 \ne 0.$$
Then, as $y_0 \to 0$,
$$g_{x,0} = O(\|y_0\|^2),$$
$$g_{y,0} = H_0 y_0 + O(\|y_0\|^2),$$
and
$$\|g_0\| = \|H_0 y_0\| + O(\|y_0\|^2).$$
Consequently,
$$\frac{g_0}{\|g_0\|} = \begin{pmatrix} 0 \\ H_0 y_0 / \|H_0 y_0\| \end{pmatrix} + O(\|y_0\|),$$
uniformly over all nonzero directions y_0 sufficiently close to the origin.

Proof. For any $u \in \mathbb{R}^p$ with $\|u\| = 1$,
$$\left| \left\langle u, \nabla_x \left(\frac{1}{2} y_0^\top H(x_0) y_0 \right) \right\rangle \right| = \frac{1}{2} \left| y_0^\top DH(x_0)[u] y_0 \right| \le \frac{A_H}{2} \|y_0\|^2.$$

Taking the supremum over unit u and using the remainder bound gives

$$\|g_{x,0}\| \le \frac{A_H}{2}\|y_0\|^2 + C_x\|y_0\|^3 = O(\|y_0\|^2).$$

Similarly,

$$g_{y,0} = H_0 y_0 + \nabla_y R(x_0, y_0) = H_0 y_0 + O(\|y_0\|^2).$$

Since

$$H_0 \succeq mI,$$

we have

$$\|H_0 y_0\| \ge m\|y_0\|.$$

Hence,

$$\|g_0 - (0, H_0 y_0)\| = O(\|y_0\|^2) = o(\|H_0 y_0\|).$$

It follows that

$$\|g_0\| = \|H_0 y_0\| + O(\|y_0\|^2).$$

Finally, for nonzero vectors a, b satisfying $\|a - b\| \le \|b\|/2$,

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \le \frac{2\|a - b\|}{\|b\|}.$$

Applying this with

$$a = g_0, \qquad b = (0, H_0 y_0)$$

yields

$$\frac{g_0}{\|g_0\|} = \begin{pmatrix} 0 \\ H_0 y_0 / \|H_0 y_0\| \end{pmatrix} + O(\|y_0\|).$$

Uniformity follows from the direction-independent lower bound $\|H_0 y_0\| \ge m\|y_0\|$. □

Lemma 4 (Uniform size of the leading displacement)**.** *For every nonzero $y \in \mathbb{R}^k$,*

$$m \le \frac{\|H_0^2 y\|}{\|H_0 y\|} \le \lambda_1.$$

Proof. Write

$$y = \sum_i a_i v_i.$$

Then,

$$\frac{\|H_0^2 y\|^2}{\|H_0 y\|^2} = \frac{\sum_i \lambda_i^4 a_i^2}{\sum_i \lambda_i^2 a_i^2}.$$

The right-hand side is a weighted average of the values λ_i^2. Since

$$m \le \lambda_i \le \lambda_1,$$

the ratio lies between m^2 and λ_1^2. Taking square roots proves the claim. □

C.3 Finite Post-Switch Displacement

Proof of Theorem 2. Under Assumption 4, fix $\eta > 0$ in a bounded range and let $\rho > 0$ be sufficiently small that the normalized-SAM inner point and the one-step outer update remain inside the local sharp-valley neighborhood.

For any sufficiently small $y_0 \ne 0$, consider one normalized-SAM step from (x_0, y_0):

$$(\widetilde{x}_0, \widetilde{y}_0) = (x_0, y_0) + \rho \frac{\nabla L(x_0, y_0)}{\|\nabla L(x_0, y_0)\|},$$

followed by

$$y_1 = y_0 - \eta \nabla_y L(\widetilde{x}_0, \widetilde{y}_0).$$

Then,

$$y_1 = -\eta\rho\frac{H_0^2 y_0}{\|H_0 y_0\|} + (I - \eta H_0)y_0 + \mathcal{E}_{\mathrm{dir}} + \mathcal{E}_{\mathrm{geom}} + \mathcal{E}_R, \tag{16}$$

where, uniformly over all nonzero directions y_0,

$$\|\mathcal{E}_{\mathrm{dir}}\| \le C_{\mathrm{dir}}\eta\rho\|y_0\|,$$
$$\|\mathcal{E}_{\mathrm{geom}}\| \le C_{\mathrm{geom}}\eta\big(\rho\|y_0\|^2 + \rho^2\|y_0\|\big),$$

and

$$\|\mathcal{E}_R\| \le C_R\eta\big(\|y_0\|^2 + \rho\|y_0\| + \rho^2\big).$$

In particular,

$$y_1 = -\eta\rho\frac{H_0^2 y_0}{\|H_0 y_0\|} + (I - \eta H_0)y_0 + O(\eta\rho\|y_0\|) + O(\eta\rho^2) + O(\eta\|y_0\|^2). \tag{17}$$

Consequently, if

$$\frac{\|y_0\|}{\rho} \to 0$$

with fixed $\eta > 0$ and ρ restricted to a sufficiently small bounded interval, then

$$y_1 = -\eta\rho\frac{H_0^2 y_0}{\|H_0 y_0\|} + O(\eta\rho^2) + o_{\|y_0\|/\rho\to 0}(\eta\rho).$$

If, in addition,

$$\rho \to 0, \qquad \frac{\|y_0\|}{\rho} \to 0,$$

then

$$y_1 = -\eta\rho\frac{H_0^2 y_0}{\|H_0 y_0\|} + o(\eta\rho).$$

Define

$$u_0 := \frac{\nabla L(x_0, y_0)}{\|\nabla L(x_0, y_0)\|} = (u_{x,0}, u_{y,0}).$$

By Lemma 3,

$$u_{x,0} = O(\|y_0\|),$$

and

$$u_{y,0} = \frac{H_0 y_0}{\|H_0 y_0\|} + O(\|y_0\|).$$

Hence,

$$\widetilde{x}_0 - x_0 = O(\rho\|y_0\|)$$

and

$$\widetilde{y}_0 = y_0 + \rho\frac{H_0 y_0}{\|H_0 y_0\|} + O(\rho\|y_0\|).$$

The outer normal gradient is

$$\nabla_y L(\widetilde{x}_0, \widetilde{y}_0) = H(\widetilde{x}_0)\widetilde{y}_0 + \nabla_y R(\widetilde{x}_0, \widetilde{y}_0).$$

Insert and subtract H_0:

$$\begin{aligned} H(\widetilde{x}_0)\widetilde{y}_0 = H_0 y_0 &+ \rho\frac{H_0^2 y_0}{\|H_0 y_0\|} \\ &+ H_0 O(\rho\|y_0\|) \\ &+ \big[H(\widetilde{x}_0) - H_0\big]\widetilde{y}_0. \end{aligned}$$

The direction-approximation term satisfies

$$\|H_0 O(\rho\|y_0\|)\| = O(\rho\|y_0\|).$$

For the geometric variation term, the local derivative bound gives

$$\|H(\widetilde{x}_0) - H_0\| \le A_H \|\widetilde{x}_0 - x_0\| = O(\rho\|y_0\|).$$

Moreover,

$$\|\widetilde{y}_0\| \le \|y_0\| + \rho + O(\rho\|y_0\|).$$

Therefore,

$$\|[H(\widetilde{x}_0) - H_0]\widetilde{y}_0\| = O(\rho\|y_0\|^2 + \rho^2\|y_0\|).$$

Finally,

$$\|\nabla_y R(\widetilde{x}_0, \widetilde{y}_0)\| \le C_y \|\widetilde{y}_0\|^2,$$

so

$$\|\nabla_y R(\widetilde{x}_0, \widetilde{y}_0)\| = O(\|y_0\|^2 + \rho\|y_0\| + \rho^2).$$

Combining these estimates,

$$\begin{aligned}\nabla_y L(\widetilde{x}_0, \widetilde{y}_0) = H_0 y_0 + \rho \frac{H_0^2 y_0}{\|H_0 y_0\|} \\ + O(\rho\|y_0\|) \\ + O(\rho\|y_0\|^2 + \rho^2\|y_0\|) \\ + O(\|y_0\|^2 + \rho\|y_0\| + \rho^2).\end{aligned}$$

Substituting into

$$y_1 = y_0 - \eta \nabla_y L(\widetilde{x}_0, \widetilde{y}_0)$$

gives (16) and (17).

If

$$\|y_0\|/\rho \to 0,$$

then, for fixed $\eta > 0$,

$$(I - \eta H_0) y_0 = o(\eta\rho),$$

and all mixed terms involving $\rho\|y_0\|$ or $\|y_0\|^2$ are also $o(\eta\rho)$. The remainder evaluated at the SAM inner point retains the $O(\eta\rho^2)$ contribution. This proves the fixed-small-ρ late-switch expansion. Under the additional limit $\rho \to 0$, that term is itself $o(\eta\rho)$.

Moreover, there exist constants $c_-, c_+ > 0$ and thresholds $\rho_0, \varepsilon_0 > 0$ such that, whenever

$$0 < \rho \le \rho_0, \qquad 0 < \|y_0\| \le \varepsilon_0 \rho,$$

one has

$$c_- \eta\rho \le \|y_1\| \le c_+ \eta\rho.$$

The constants may be chosen uniformly over all directions of y_0. Thus, an arbitrarily small but nonzero pre-switch normal state can produce a finite $O(\eta\rho)$ post-switch normal displacement. This conclusion does not require prior principal alignment.

By Lemma 4, the leading displacement term satisfies

$$m\eta\rho \le \eta\rho \frac{\|H_0^2 y_0\|}{\|H_0 y_0\|} \le \lambda_1 \eta\rho.$$

The error terms in (17) are $o(\eta\rho)$ when $\|y_0\|/\rho$ and ρ are sufficiently small. Choosing ε_0 and ρ_0 so that the total error is, for example, at most $(m/2)\eta\rho$ gives the lower bound, while the upper bound follows similarly after enlarging the constant. □

C.4 PRINCIPAL-MODE SPECTRAL FILTERING

The finite-amplitude conclusion is uniform over the incoming direction, but the modal composition of the leading displacement depends on the spectrum of H_0. We now impose a fixed principal-cone condition to quantify this spectral effect.

When the principal component is nonzero, write

$$y_0 = \sum_{i=1}^{k} a_i v_i$$

and define the initial projective ratio

$$r_0 := \frac{\left(\sum_{i\geq 2} a_i^2\right)^{1/2}}{|a_1|}.$$

The finite-displacement result below does not require $a_1 \neq 0$ or a bound on r_0. These conditions are introduced only for the subsequent spectral-filtering statement.

Proof of the spectral-filtering claim in Section 5. Under the assumptions of Theorem 2, suppose in addition that

$$a_1 \neq 0, \qquad r_0 \leq R_0 < \infty.$$

Let

$$P_i := v_i v_i^\top, \qquad P_\perp := I - P_1,$$

and define the post-switch frozen-basis projective ratio

$$r_1 := \frac{\|P_\perp y_1\|}{\|P_1 y_1\|},$$

whenever the denominator is nonzero. Then, uniformly over the fixed cone $r_0 \leq R_0$,

$$\frac{\|P_i y_1\|}{\|P_1 y_1\|} = \left(\frac{\lambda_i}{\lambda_1}\right)^2 \frac{\|P_i y_0\|}{\|P_1 y_0\|} + O(\rho) + o_{\|y_0\|/\rho\to 0}(1), \qquad i \geq 2.$$

Consequently,

$$r_1 \leq \left(\frac{\lambda_2}{\lambda_1}\right)^2 r_0 + O(\rho) + o_{\|y_0\|/\rho\to 0}(1).$$

Writing

$$y_0 = \sum_i a_i v_i,$$

the leading displacement in Theorem 2 is

$$-\eta\rho \frac{H_0^2 y_0}{\|H_0 y_0\|} = -\eta\rho \sum_i \frac{\lambda_i^2 a_i}{\|H_0 y_0\|} v_i.$$

Therefore, the i-to-1 modal ratio of the leading term is exactly

$$\left(\frac{\lambda_i}{\lambda_1}\right)^2 \frac{|a_i|}{|a_1|}.$$

Since

$$r_0 \leq R_0,$$

we have

$$|a_1| \geq \frac{\|y_0\|}{\sqrt{1+R_0^2}}.$$

Also,

$$\|H_0 y_0\| \leq \lambda_1 \|y_0\|.$$

Hence, the magnitude of the principal component of the leading displacement is bounded below by

$$\eta\rho\frac{\lambda_1^2|a_1|}{\|H_0y_0\|} \geq \frac{\eta\rho\lambda_1}{\sqrt{1+R_0^2}}.$$

Thus, the leading principal amplitude is uniformly of order $\eta\rho$ on the fixed cone. The remainder in the late-switch expansion is

$$O(\eta\rho^2) + o(\eta\rho).$$

Dividing the numerator and the denominator of each modal ratio by the leading principal amplitude therefore perturbs the ratio by

$$O(\rho) + o(1).$$

This proves the componentwise statement.

For the transverse ratio, the leading transverse component satisfies

$$\left\|\sum_{i\geq 2}\lambda_i^2 a_i v_i\right\| \leq \lambda_2^2\left(\sum_{i\geq 2}a_i^2\right)^{1/2}.$$

Dividing by the corresponding principal leading component $\lambda_1^2|a_1|$ and applying the same perturbation estimate yields

$$r_1 \leq \left(\frac{\lambda_2}{\lambda_1}\right)^2 r_0 + O(\rho) + o(1).$$

□

D Frozen Principal-Shell Dynamics

This appendix develops the frozen quadratic dynamics underlying the local oscillatory analysis of Section 6. Fixing the tangential coordinate and suppressing the nonlinear remainder yields an exact normalized-SAM normal map with fixed curvature and eigendirections. This model isolates the mechanism governing phase alternation and transverse contraction.

The principal two-cycle and its radius are known for normalized SAM on quadratic objectives (Bartlett et al., 2023). Building on this reference dynamics, we identify an explicit forward-invariant basin and prove geometric contraction of transverse components within it, yielding convergence of the phase-corrected iterates to the principal two-cycle. The basin conditions precisely specify the local regime in which this oscillatory organization is guaranteed.

D.1 Exact Frozen Quadratic Map

Let

$$H \succ 0, \qquad Hv_i = \lambda_i v_i, \qquad \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_k > 0,$$

with an orthonormal eigenbasis. Assume

$$1 < \eta\lambda_1 < 2.$$

Define

$$a := \eta\lambda_1, \qquad \delta := 2 - a \in (0,1),$$

and the frozen quadratic normal map

$$F(y) := (I - \eta H)y - \eta\rho\frac{H^2y}{\|Hy\|}, \qquad y \neq 0. \tag{18}$$

This is the normalized-SAM normal update for the quadratic objective $\frac{1}{2}y^\top Hy$.

Define the principal shell radius

$$c := \frac{\eta\lambda_1\rho}{2-\eta\lambda_1} = \frac{(2-\delta)\rho}{\delta}. \tag{19}$$

With this choice of c, for every $\sigma \in \{-1, +1\}$,

$$F(\sigma c v_1) = -\sigma c v_1.$$

Consequently,

$$\mathcal{S} := \{-cv_1, +cv_1\}$$

is an exact two-cycle of the frozen map (Bartlett et al., 2023).

D.2 Principal and Transverse Coordinates

For a state with nonzero principal projection, write

$$y = \sigma \alpha v_1 + z, \qquad \alpha > 0, \qquad z \perp v_1, \qquad \sigma \in \{-1, +1\}.$$

Define the projective ratio

$$r := \frac{\|z\|}{\alpha}, \tag{20}$$

the normalized eigenvalues

$$q_i := \frac{\lambda_i}{\lambda_1}, \qquad q_* := q_2 = \frac{\lambda_2}{\lambda_1} < 1,$$

and the weighted angular factor

$$\chi := \frac{\|Hy\|}{\lambda_1 \alpha}. \tag{21}$$

Lemma 5 (Exact weighted-angle identity)**.** *If*

$$z = \sum_{i \geq 2} z_i v_i,$$

then

$$\chi^2 = 1 + \sum_{i \geq 2} q_i^2 \frac{z_i^2}{\alpha^2}.$$

Consequently,

$$1 \leq \chi \leq \sqrt{1 + q_*^2 r^2},$$

and

$$0 \leq 1 - \frac{1}{\chi} \leq \frac{q_*^2}{2} r^2. \tag{22}$$

Proof. Since

$$Hy = \sigma \lambda_1 \alpha v_1 + \sum_{i \geq 2} \lambda_i z_i v_i,$$

orthogonality gives

$$\|Hy\|^2 = \lambda_1^2 \alpha^2 + \sum_{i \geq 2} \lambda_i^2 z_i^2.$$

Dividing by $\lambda_1^2 \alpha^2$ gives the exact identity and hence the upper bound on χ.

For $u \geq 0$,

$$1 - \frac{1}{\sqrt{1+u}} = \frac{u}{\sqrt{1+u}(1+\sqrt{1+u})} \leq \frac{u}{2}.$$

Taking

$$u = \sum_{i \geq 2} q_i^2 \frac{z_i^2}{\alpha^2} \leq q_*^2 r^2$$

proves (22). □

D.3 Exact Component Dynamics

Define

$$\tau := \frac{\eta\rho\lambda_1^2}{\|Hy\|}. \tag{23}$$

Lemma 6 (Exact component multipliers and principal sign flip)**.** *Under the frozen map* (18)*, the principal component multiplier is*

$$\mu_1 = 1 - \eta\lambda_1 - \tau = -(1 - \delta + \tau) < 0. \tag{24}$$

For every $i \geq 2$, the corresponding transverse multiplier is

$$\mu_i = 1 - (2 - \delta)q_i - \tau q_i^2. \tag{25}$$

Hence, a nonzero principal projection changes sign at every frozen iteration and remains nonzero.

Proof. The v_i component of the normalized-SAM correction equals

$$-\eta\rho\frac{\lambda_i^2 y_i}{\|Hy\|}.$$

For $i = 1$, this is $-\tau y_1$, and hence,

$$\mu_1 = 1 - \eta\lambda_1 - \tau = -(1 - \delta + \tau).$$

Since

$$0 < \delta < 1, \qquad \tau > 0,$$

the multiplier is strictly negative and nonzero. For $i \geq 2$,

$$\eta\lambda_i = (2 - \delta)q_i$$

and

$$\eta\rho\frac{\lambda_i^2}{\|Hy\|} = \tau q_i^2,$$

which proves (25). □

Lemma 7 (Exact principal-magnitude recurrence)**.** *Let α^+ denote the magnitude of the principal projection after one frozen step. Then*

$$\alpha^+ = (1 - \delta)\alpha + \frac{(2 - \delta)\rho}{\chi}. \tag{26}$$

Define the relative radial deficit

$$g := 1 - \frac{\alpha}{c}, \qquad g^+ := 1 - \frac{\alpha^+}{c}.$$

Then,

$$g^+ = (1 - \delta)g + \delta\left(1 - \frac{1}{\chi}\right). \tag{27}$$

Proof. By Lemma 6,

$$\alpha^+ = (1 - \delta + \tau)\alpha.$$

Since

$$\|Hy\| = \lambda_1\alpha\chi,$$

we have

$$\tau\alpha = \frac{\eta\rho\lambda_1}{\chi} = \frac{(2 - \delta)\rho}{\chi}.$$

This proves (26).

Moreover,

$$\delta c = (2 - \delta)\rho.$$

Dividing (26) by c gives

$$\frac{\alpha^+}{c} = (1 - \delta)\frac{\alpha}{c} + \frac{\delta}{\chi}.$$

Subtracting from one gives (27). □

D.4 Projective Contraction in the Sub-Shell Region

Lemma 8 (Equivalent sub-shell conditions)**.** *The following three conditions are equivalent:*

$$\tau \geq \delta,$$

$$\|Hy\| \leq \lambda_1 c,$$

and

$$\alpha\chi \leq c.$$

Proof. Using

$$\tau = \frac{\eta\rho\lambda_1^2}{\|Hy\|}$$

and

$$\delta c = \eta\rho\lambda_1,$$

the inequality $\tau \geq \delta$ is equivalent to

$$\lambda_1 c \geq \|Hy\|.$$

Since

$$\|Hy\| = \lambda_1\alpha\chi,$$

this is equivalent to

$$\alpha\chi \leq c.$$

□

Lemma 9 (Exact projective contraction in the sub-shell region)**.** *For $q \in (0,1)$ define*

$$R_q(\tau) := \frac{|1-(2-\delta)q-\tau q^2|}{1-\delta+\tau}.$$

For every $\tau \geq \delta$,

$$R_q(\tau) \leq \max\left\{q^2,\, |1-(2-\delta)q-\delta q^2|\right\}. \tag{28}$$

Define

$$q_{\mathrm{fr}} := \max_{i\geq 2} \max\left\{q_i^2,\, |1-(2-\delta)q_i-\delta q_i^2|\right\}. \tag{29}$$

Then,

$$q_{\mathrm{fr}} < 1,$$

and every sub-shell iterate satisfies

$$r^+ \leq q_{\mathrm{fr}} r. \tag{30}$$

Proof. Fix $q \in (0,1)$ and set

$$N(\tau) := 1-(2-\delta)q-\tau q^2, \qquad D(\tau) := 1-\delta+\tau.$$

Since $D(\tau) > 0$, on every interval on which N has fixed sign, R_q equals either N/D or $-N/D$. Moreover,

$$\frac{d}{d\tau}\frac{N(\tau)}{D(\tau)} = \frac{-q^2(1-\delta)-1+(2-\delta)q}{D(\tau)^2},$$

whose numerator is independent of τ. Thus each branch is monotone. An interior zero of N gives $R_q = 0$ and cannot be a maximizer. Consequently, over $[\delta,\infty)$ the supremum occurs either at $\tau = \delta$ or in the limit $\tau \to \infty$. These two values are

$$R_q(\delta) = |1-(2-\delta)q-\delta q^2|$$

and

$$\lim_{\tau\to\infty} R_q(\tau) = q^2,$$

which proves (28).

We now show that both endpoint values are strictly below one. Clearly

$$q^2 < 1.$$

Let

$$m(q) := 1 - (2 - \delta)q - \delta q^2.$$

Since $q > 0$,

$$m(q) < 1.$$

Also,

$$1 + m(q) = (1 - q)(2 + \delta q) > 0,$$

so

$$m(q) > -1.$$

Hence,

$$|m(q)| < 1,$$

and therefore, $q_{\text{fr}} < 1$.

Finally, Lemma 6 gives

$$\frac{|z_i^+|}{\alpha^+} = R_{q_i}(\tau)\frac{|z_i|}{\alpha}.$$

Squaring and summing over $i \geq 2$ yields

$$(r^+)^2 \leq q_{\text{fr}}^2 r^2,$$

which proves (30). □

D.5 A Forward-Invariant Frozen Basin

Define

$$\kappa := \frac{q_*^2}{2}. \tag{31}$$

Lemma 10 (A sufficient sub-shell slack condition)**.** *If*

$$g \geq \kappa r^2, \tag{32}$$

then

$$\alpha\chi \leq c.$$

Thus, the state lies in the sub-shell region.

Proof. By Lemma 5,

$$1 - \frac{1}{\chi} \leq \kappa r^2.$$

Therefore,

$$g \geq \kappa r^2 \geq 1 - \frac{1}{\chi}.$$

Since

$$g = 1 - \frac{\alpha}{c},$$

we obtain

$$\frac{\alpha}{c} \leq \frac{1}{\chi},$$

or equivalently

$$\alpha\chi \leq c.$$

□

Proof of Theorem 3. Assume the rate-separation condition

$$q_{\text{fr}}^2 \leq 1 - \delta. \tag{33}$$

Suppose the initial state has nonzero principal projection and satisfies

$$g_0 \geq \kappa r_0^2. \tag{34}$$

We prove

$$g_t \geq \kappa r_t^2$$

by induction. It holds initially by (34), and Lemma 10 therefore places the initial state in the sub-shell region.

Suppose the invariant inequality holds at time t. Since the state is sub-shell, Lemma 9 gives

$$r_{t+1} \leq q_{\mathrm{fr}} r_t.$$

The exact deficit recurrence (27) and $1 - \chi_t^{-1} \geq 0$ imply

$$g_{t+1} \geq (1-\delta) g_t.$$

Hence,

$$g_{t+1} \geq (1-\delta)\kappa r_t^2 \geq q_{\mathrm{fr}}^2 \kappa r_t^2 \geq \kappa r_{t+1}^2,$$

where the middle inequality uses (33). The induction closes. Thus all iterates remain sub-shell and

$$r_t \leq q_{\mathrm{fr}}^t r_0.$$

By Lemma 6, the principal multiplier is strictly negative at every step. Hence a nonzero principal projection remains nonzero and alternates sign.

To obtain radial convergence, use Lemma 5:

$$1 - \frac{1}{\chi_t} \leq \kappa r_t^2 \leq \kappa q_{\mathrm{fr}}^{2t} r_0^2.$$

Therefore,

$$g_{t+1} \leq (1-\delta) g_t + \delta \kappa q_{\mathrm{fr}}^{2t} r_0^2.$$

Set

$$A := 1 - \delta, \qquad B := q_{\mathrm{fr}}^2.$$

Iterating the preceding recurrence yields

$$g_t \leq A^t g_0 + \delta \kappa r_0^2 \sum_{j=0}^{t-1} A^{t-1-j} B^j.$$

If $A \neq B$, the finite geometric convolution is

$$\sum_{j=0}^{t-1} A^{t-1-j} B^j = \frac{A^t - B^t}{A - B},$$

whereas if $A = B$ it equals

$$t A^{t-1}.$$

Since

$$0 < A < 1, \qquad 0 < B < 1,$$

both cases imply

$$g_t \to 0.$$

Thus,

$$\alpha_t = c(1 - g_t) \to c$$

and

$$r_t \to 0.$$

After fixing $\sigma_0 \in \{-1, +1\}$ so that the principal component of $\sigma_0 y_0$ is positive, define

$$\sigma_t := (-1)^t \sigma_0, \qquad \widehat{y}_t := \sigma_t y_t.$$

By construction of σ_t, the phase-corrected principal component is positive. Writing

$$\widehat{y}_t = \alpha_t v_1 + \widehat{z}_t, \qquad \|\widehat{z}_t\| = \alpha_t r_t,$$

gives

$$\|\widehat{y}_t - c v_1\| \leq |\alpha_t - c| + \alpha_t r_t.$$

Since

$$|\alpha_t - c| = cg_t$$

and

$$\alpha_t = c(1 - g_t),$$

we obtain

$$\frac{\|\widehat{y}_t - cv_1\|}{c} \leq g_t + (1 - g_t)r_t \leq g_t + r_t. \tag{35}$$

Both terms tend to zero, proving

$$\widehat{y}_t \longrightarrow cv_1.$$

□

E Experimental Details and Reproducibility

This appendix specifies the experimental protocol underlying Section 7. The experiments are divided into controlled sharp-valley tests, where the geometric quantities appearing in the theory are directly observable, and neural-network interventions, where we test observable post-switch signatures without assuming access to a global sharp-valley coordinate system. Unless stated otherwise, experimental settings, intervention grids, validity thresholds, and outcome definitions are fixed before inspecting individual trajectories.

E.1 Controlled Experiments

Objectives and regimes. The controlled suite is designed to separate the assumptions entering the local analysis. Each objective has the sharp-valley form

$$L(x, y) = \frac{1}{2} y^\top H(x) y + R(x, y), \tag{36}$$

with explicitly available normal coordinates, Hessian spectrum, and leading eigendirection. We consider six regimes:

1. a *frozen quadratic* reference, with constant H and $R \equiv 0$;
2. *moving curvature*, in which the eigenvalues of $H(x)$ vary with the tangential coordinate;
3. *rotating eigendirections*, in which the eigenbasis of $H(x)$ varies with x;
4. a *small-eigengap* regime that weakens separation of the leading normal mode;
5. a *fast-rotation* regime that increases variation of the leading eigendirection; and
6. a *nonlinear-remainder* regime with $R \neq 0$.

The frozen quadratic case is the exact reference model for Section 6; the remaining cases progressively stress the assumptions used by that approximation. Exact objective coefficients, initial states, and parameter grids are provided in the released configuration files so that the reported instances can be reconstructed without retuning.

Switching experiment. For each controlled configuration, we select a nonstationary state and evaluate one GD step and one normalized-SAM step from exactly the same parameter vector and with the same outer step size η. A *same-state switching event* is recorded when

$$L(w^+_{\mathrm{GD}}) < L(w) \qquad \text{and} \qquad L(w^+_{\mathrm{SAM}}) > L(w). \tag{37}$$

For aligned quadratic instances, the observed event is compared directly with the exact theoretical condition

$$\frac{2x}{x + \rho} < \eta\lambda_1 < 2. \tag{38}$$

Across the controlled suite used in the main text, the empirical same-state switching-event rate is 0.75. This rate is a summary over the tested configurations, not an estimate of a population probability; the purpose of the experiment is to verify that the predicted switching region is realized under controlled geometry.

Finite-displacement experiment. To test Theorem 2, we initialize states close to the valley floor and vary the SAM perturbation radius while keeping the local objective fixed. For each intervention we record the first post-switch normal displacement and compare it with the leading prediction

$$d_{\mathrm{pred}} = -\eta\rho \frac{H_0^2 y_0}{\|H_0 y_0\|}. \tag{39}$$

Radius dependence is assessed from the measured displacement magnitude as ρ varies. Spectral filtering is evaluated in the eigenbasis of H_0: for $i \geq 2$, we compare the post-switch projective ratio

$$\frac{\|P_i y_1\|}{\|P_1 y_1\|} \tag{40}$$

with the leading-order prediction

$$\left(\frac{\lambda_i}{\lambda_1}\right)^2 \frac{\|P_i y_0\|}{\|P_1 y_0\|}. \tag{41}$$

Only configurations satisfying the fixed-cone condition used in the theoretical filtering statement are included in this comparison.

Frozen-shell experiment. For the frozen quadratic reference, we iterate the exact normal map

$$y_{t+1} = (I - \eta H) y_t - \eta\rho \frac{H^2 y_t}{\|H y_t\|} \tag{42}$$

from an initial condition satisfying the invariant-basin assumptions of Theorem 3. The reference shell radius is

$$c = \frac{\eta\lambda_1\rho}{2 - \eta\lambda_1}. \tag{43}$$

After correcting for the alternating phase, the trajectory is compared with cv_1. The same diagnostic is then applied to the moving-curvature, rotating-eigendirection, small-eigengap, fast-rotation, and nonlinear stress tests, using the corresponding instantaneous local spectral quantities.

For the configurations reported in the main text, the resulting shell errors are

Regime	Shell error
Frozen quadratic	5.68×10^{-6}
Moving curvature	8.0×10^{-3}
Fast eigendirection rotation	2.4×10^{-2}
Small eigengap	5.2×10^{-2}
Nonlinear remainder	1.38×10^{-1}

These stress tests are not used to extend the frozen-shell theorem beyond its assumptions. They quantify how the frozen reference deteriorates when those assumptions are progressively relaxed.

E.2 Neural-Network Experiments

Training protocol. We use ResNet-18 on CIFAR-10. Each network is trained for 200 epochs using SGD with batch size 128, momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.1 followed by cosine decay. The primary same-state interventions use checkpoints from epochs 160, 180, and 190. Independent training runs are used for the cross-seed replication study.

Diagnostic objective. Local post-switch measurements are performed on a fixed diagnostic subset of 2,048 CIFAR-10 examples, with diagnostic-subset seed 271828. The same subset is used for the local loss, gradient, Hessian-vector products, and spectral measurements associated with an intervention. Holding this objective fixed prevents minibatch variation from being confounded with the optimizer switch.

The diagnostic objective is used to probe local dynamics; it is not intended as an independent estimate of test performance or generalization.

Spectral estimation. Leading Hessian eigenpairs are estimated by Lanczos iterations using Hessian-vector products on the fixed diagnostic objective. Checkpoint-level spectral estimates use 30 Lanczos iterations, whereas trajectory-level estimates use 10 iterations. Eigenvector signs are aligned across successive states before signed principal-mode quantities are compared.

We use predeclared numerical validity gates. A residual below 10^{-3} is classified as *strongly valid*, while a residual below 10^{-2} is accepted as *valid*. We additionally require a resolved relative leading eigengap exceeding 10^{-3} whenever the interpretation depends on a distinguished principal eigendirection. Measurements failing the required gate are marked spectrally invalid rather than assigned an outcome based on an unreliable eigenpair.

Same-state interventions. At each selected checkpoint, interventions branch from exactly the same parameter vector. One branch performs a GD update and the other performs a normalized-SAM update with the same outer step size. For the primary same-state sweep, the initial edge-distance grid is

$$\delta_0 \in \{0.1, 0.2, 0.4, 0.8, 1.2\},$$

and the perturbation-radius grid is

$$\rho \in \{0.0125, 0.025, 0.05, 0.1\}.$$

The effective outer step size for an intervention is calibrated through the desired initial edge distance using

$$\eta = \frac{2 - \delta_0}{\lambda_{1,0}}, \tag{44}$$

where $\lambda_{1,0}$ is the accepted leading Hessian eigenvalue at the intervention state.

A same-state crossing is recorded when the diagnostic loss decreases under the GD branch but increases under the normalized-SAM branch. In the seed-0 sweep, the resulting numbers of calibrated crossings are 13, 5, and 13 at epochs 160, 180, and 190, respectively.

SAM-specific displacement. To separate the effect of changing the optimizer from the displacement that would have occurred under GD, we define

$$\Delta w_{\text{excess}} = \Delta w_{\text{SAM}} - \Delta w_{\text{GD}}. \tag{45}$$

For each checkpoint, we regress $\log \|\Delta w_{\text{excess}}\|$ against $\log \rho$ over the tested radii. The fitted slopes at epochs 160, 180, and 190 are respectively 1.558, 0.819, and 1.428, with $R^2 > 0.98$ in all three cases. These regressions are used only as finite-radius diagnostics. In particular, the fitted slopes are not interpreted as estimates of a universal asymptotic exponent.

Dense regime sweep. To resolve the post-switch transition more finely, we perform a dense sweep at epochs 160, 180, and 190 over 23 initial edge distances and 25 perturbation radii, for a total of

$$3 \times 23 \times 25 = 1{,}725$$

interventions. The exact ordered grid values are fixed in the released experiment configuration and are used unchanged to generate Figure 2. Each trajectory is followed through the predeclared finite post-switch horizon or until a terminal outcome is resolved.

Across this sweep, 1,140 interventions resolve through the flat side, 580 through the sharp side, and 5 fail the spectral validity gate. Among the valid trajectories, 63 contain at least two consecutive controlled-edge states. The longest controlled transient in this dense seed-0 sweep contains four states.

Outcome definitions. At every spectrally valid state we define

$$\delta_t = 2 - \eta \lambda_{1,t}. \tag{46}$$

The three trajectory states used throughout the paper are

$$\delta_t \le 0 \quad \Longrightarrow \quad \textit{sharp-side exit},$$
$$0 < \delta_t < 1 \quad \Longrightarrow \quad \textit{controlled-edge state},$$

and

$$\delta_t \geq 1 \quad \Longrightarrow \quad \textit{flat-side escape}.$$

A trajectory outcome is determined by the first terminal sharp-side or flat-side event. A *controlled transient length* is the number of consecutive spectrally valid states satisfying $0 < \delta_t < 1$ before the terminal outcome.

These labels describe the local spectral diagnostic only. In particular, "flat-side escape" denotes motion to $\delta_t \geq 1$ and is not used as a claim that the network has reached a globally flat minimum.

Phase alternation. Let $v_{1,t}$ denote the phase-aligned leading Hessian eigenvector and let s_t denote the signed projection of the diagnostic gradient onto that direction. For a controlled transient, phase alternation records whether successive nonzero signed principal components reverse sign. We report the fraction of eligible successive pairs satisfying

$$s_t s_{t+1} < 0. \tag{47}$$

A value of 1 therefore denotes alternation at every eligible transition in the reported transient. This statistic tests the oscillatory signature only; it does not measure agreement with the predicted shell amplitude.

Shell-proxy error. For trajectories on which the required spectral quantities remain valid, we compare the observed phase-corrected principal amplitude with the instantaneous frozen-shell reference

$$c_t = \frac{\eta \lambda_{1,t} \rho}{2 - \eta \lambda_{1,t}}. \tag{48}$$

The shell-proxy error is the relative discrepancy between the observed principal-mode amplitude and this local frozen reference, as implemented in the released evaluation script. It is used as a diagnostic rather than as a neural-network theorem test: the frozen analysis assumes constant curvature and eigendirections, whereas both can evolve substantially along a network trajectory.

For the representative flat-resolving trajectory discussed in Section **??**, the shell-proxy error remains approximately 0.72–0.96. This large discrepancy is why the neural experiments are interpreted as evidence for oscillatory onset rather than quantitative principal-shell tracking.

Cross-seed replication. We repeat a compact, preselected intervention grid on four independently trained networks, using the same diagnostic definitions and validity criteria. Across the resulting 48 interventions, 40 resolve through the sharp side and 8 through the flat side. The absolute transition location varies across seeds and checkpoints. Representative ordered outcome sequences as ρ increases include

F–F–S–S

for seed 1 at epoch 190, and

F–S–S–S

for seed 2 at epoch 190 and seed 3 at epoch 160. Here F denotes flat-side escape and S denotes sharp-side exit. The longest controlled-edge transient observed in the replication study contains five states.

The purpose of this experiment is to test recurrence of the qualitative post-switch transition across independently trained networks. It does not support a seed-independent critical value of ρ.

Numerical failures and spectral stopping. Every one of the 48 cross-seed interventions produces a valid terminal outcome summary. Of these, 23 complete the full requested diagnostic trajectory and 25 terminate subsequent spectral estimation only after the predeclared sharp-side or flat-side outcome has already been resolved. Among these 25 post-resolution stops, 17 are caused by non-finite Hessian-vector products and 8 by a non-finite Lanczos coefficient. Because these events occur after the terminal outcome has been determined, they do not alter the reported outcome classification.

Across all 230 valid spectral evaluations in this replication study, the spectral computation is performed in double precision on CPU and requires no numerical jitter. No intervention is assigned a terminal outcome from an unresolved non-finite spectral state.

Reported metrics. For clarity, the principal reported quantities are:

- *same-state crossing*: GD decreases and normalized SAM increases the fixed diagnostic loss from the same incoming state;
- *SAM-specific displacement*: $\|\Delta w_{\mathrm{SAM}} - \Delta w_{\mathrm{GD}}\|$;
- *edge distance*: $\delta_t = 2 - \eta\lambda_{1,t}$;
- *controlled transient length*: the number of consecutive valid states satisfying $0 < \delta_t < 1$ before terminal resolution;
- *phase alternation*: the fraction of eligible consecutive principal-mode projections whose signs alternate; and
- *shell-proxy error*: relative discrepancy from the instantaneous frozen-shell reference.