GRADIENT CLIPPING AT AND BEYOND THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

When training deep neural networks with full-batch gradient descent, the curvature of the loss stays at the threshold of instability, hovering around $2/\eta$ where η is the learning rate. We study how clipping changes this Edge-of-Stability (EoS) phenomenon in first-order (FO) and zeroth-order (ZO) optimization under batch and per-sample clipping. Our central finding is that clipping stabilizes FO and ZO training in qualitatively different ways. For FO, clipping shifts the relevant stability edge, and the trajectory remains near this clipping-aware edge; stronger clipping permits stability at larger curvature. For batch-clipped ZO, training remains bounded at curvatures well beyond the mean-square stability threshold, even though standard stability analysis predicts instability in this regime. For both batch- and per-sample clipped ZO, we characterize nonzero fixed points through the covariance of the parameter displacement and show local attraction within the corresponding closed covariance dynamics under explicit conditions. Experiments show that FO trajectories track clipping-aware stability edges, while batch-clipped ZO approaches its nonzero covariance balance beyond the predicted stability limit and per-sample ZO moves toward the predicted matrix-valued covariance balance along the leading-curvature directions.

1 INTRODUCTION

During full-batch gradient descent (GD) with step size $\eta > 0$, the largest Hessian eigenvalue often rises to and remains near the linear-instability threshold, $\eta\lambda_1(\boldsymbol{H}) \approx 2$. This Edge-of-Stability (EoS) behavior appears across GD and its first-order (FO) variants (Cohen et al., 2021; Damian et al., 2023). Zeroth-order (ZO) methods instead use random search directions and only function evaluations (Malladi et al., 2023; Zhang et al., 2025), so they require a stochastic stability criterion. Song et al. (2026) derive a mean-square boundary that depends on the full Hessian spectrum, becomes trace-dominated at high effective rank, and is approached by unclipped ZO trajectories.

Gradient clipping is widely used to control large updates and stabilize training (Pascanu et al., 2013; Zhang et al., 2020; Marshall et al., 2024). In practice, however, clipping can do more than bound individual updates: it can change which curvature regimes training reaches. Viewing clipping only as a safeguard or adaptive step-size reduction therefore leaves open what dynamical principle governs the trajectory once clipping becomes active.

This distinction matters most when training remains bounded where standard linear analysis predicts instability. Does clipping shift the stability edge, so that the trajectory tracks a modified boundary, or sustain a bounded fluctuating state beyond that edge through a nonlinear balance? Is the mechanism the same for deterministic FO and stochastic ZO, or when clipping is applied before rather than after sample aggregation? We answer these questions by identifying the rule-specific stability boundary or steady-state balance that selects each trajectory's operating point.

Contributions. Our results reveal a sharp contrast between FO and ZO optimization under clipping:

- **FO clipping shifts the edge.** For batch and per-sample clipping, we derive rule-specific update Jacobians and tractable stability proxies. Trajectories track the resulting edges, with stronger clipping supporting larger raw curvature.

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.

- **Batch-clipped ZO operates beyond the edge.** We first compute the mean-square linear stability boundary from the covariance operator and find that clipped trajectories remain bounded far beyond it. Under gradient–Hessian eigenvector alignment, we obtain a scalar characterization that explains why the leading stability threshold depends only weakly on the clipping strength. Beyond this linear edge, we derive a nonlinear covariance balance, give sufficient conditions for local attraction, and validate the balance experimentally.
- **Per-sample-clipped ZO has a matrix-valued covariance balance.** Sample-specific clipping precludes a scalar threshold. We derive a computable directional diagnostic, give sufficient conditions for existence and local attraction, and validate the balance empirically.

Related work. Prior work studies clipping as adaptive rescaling and its bias, convergence, and high-dimensional dynamics (Zhang et al., 2020; Qian et al., 2021; Koloskova et al., 2023; Marshall et al., 2024). Active batch-clipped GD equals normalized GD, whose long-run EoS dynamics were studied by Arora et al. (2022); we derive clipping-aware FO conditions, including per-sample clipping. Deterministic GD EoS is defined by linear stability; mini-batch SGD instead motivates interaction-aware or batch-dependent sharpness measures (Lee & Jang, 2023; Andreyev & Beneventano, 2024). Separately, second-moment stability around fixed minima is studied for SGD (Wu et al., 2018; Velikanov et al., 2023; Mulayoff & Michaeli, 2024) and linked to ZO EoS by Song et al. (2026). Our ZO result directly complements Mulayoff & Stich (2026): both show that linearized stability can miss bounded nonlinear behavior. Here clipping is the concrete mechanism that keeps trajectories bounded beyond the predicted boundary.

2 Preliminaries

We consider $\min_{\boldsymbol{x}\in\mathbb{R}^d} L(\boldsymbol{x})$, where $L(\boldsymbol{x}) := n^{-1}\sum_{i=1}^n \ell_i(\boldsymbol{x})$ and $i \in [n]$ indexes samples. We assume each $\ell_i : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, and L is bounded below. Let $\boldsymbol{g}_i(\boldsymbol{x}) := \nabla\ell_i(\boldsymbol{x})$ and $\boldsymbol{H}_i(\boldsymbol{x}) := \nabla^2\ell_i(\boldsymbol{x})$, with averages $\boldsymbol{g}(\boldsymbol{x}) := n^{-1}\sum_i \boldsymbol{g}_i(\boldsymbol{x})$ and $\boldsymbol{H}(\boldsymbol{x}) := n^{-1}\sum_i \boldsymbol{H}_i(\boldsymbol{x})$. At $\boldsymbol{x}_t$, write $\boldsymbol{g}_{i,t} := \boldsymbol{g}_i(\boldsymbol{x}_t)$, $\boldsymbol{H}_{i,t} := \boldsymbol{H}_i(\boldsymbol{x}_t)$, $\boldsymbol{g}_t := \boldsymbol{g}(\boldsymbol{x}_t)$, and $\boldsymbol{H}_t := \boldsymbol{H}(\boldsymbol{x}_t)$. We suppress arguments and time subscripts when clear. Full-batch FO gradient descent (FO-GD) updates $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\boldsymbol{g}_t$. Full-batch ZO gradient descent (ZO-GD) independently draws $\boldsymbol{u}_t \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and forms

$$d_{i,t}^{\mu} := \frac{\ell_i(\boldsymbol{x}_t + \mu\boldsymbol{u}_t) - \ell_i(\boldsymbol{x}_t - \mu\boldsymbol{u}_t)}{2\mu}, \qquad d_t^{\mu} := \frac{1}{n}\sum_{i=1}^n d_{i,t}^{\mu}$$

to make updates $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta d_t^{\mu}\boldsymbol{u}_t$. We study the following clipped gradient updates:

$$\begin{aligned}
\text{FO–batch:} \quad & \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\ \mathrm{vclip}_C(\boldsymbol{g}_t), \\
\text{FO–per-sample:} \quad & \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\eta}{n}\sum_{i=1}^n \mathrm{vclip}_C(\boldsymbol{g}_{i,t}), \\
\text{ZO–batch:} \quad & \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\ \mathrm{clip}_C(d_t^{\mu})\boldsymbol{u}_t, \qquad (1) \\
\text{ZO–per-sample:} \quad & \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{\eta}{n}\sum_{i=1}^n \mathrm{clip}_C(d_{i,t}^{\mu})\boldsymbol{u}_t, \qquad (2)
\end{aligned}$$

with step size $\eta > 0$ and threshold $C > 0$. Here $\mathrm{vclip}_C(\boldsymbol{v}) := \min\{1, C/\|\boldsymbol{v}\|\}\boldsymbol{v}$ and $\mathrm{clip}_C(z) := \mathrm{sign}(z)\min\{|z|, C\}$, with $\mathrm{vclip}_C(\boldsymbol{0}) = \boldsymbol{0}$. *Batch* and *per-sample* denote clipping after and before averaging. FO clips vector norms, whereas ZO clips scalar derivatives before multiplying by $\boldsymbol{u}_t$; hence its vector update is not norm-clipped.

We collect other symbols here. Let $\widehat{\boldsymbol{v}} := \boldsymbol{v}/\|\boldsymbol{v}\|$ for $\boldsymbol{v} \neq \boldsymbol{0}$. For a linear operator $\mathcal{A}$, $\mathrm{spec}(\mathcal{A})$ denotes its spectrum and $\rho(\mathcal{A}) := \max_{\lambda\in\mathrm{spec}(\mathcal{A})}|\lambda|$ its spectral radius. For a complex scalar λ, $\Re\lambda$ and $\Im\lambda$ denote its real and imaginary parts.

2.1 Linear stability around a stationary point

Standard EoS analysis asks whether an infinitesimal displacement from a stationary fixed point remains bounded under the optimization dynamics (Cohen et al., 2021; Song et al., 2026). Write

the update as $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta F_t(\boldsymbol{x}_t)$, let $\boldsymbol{x}^\star$ be a stationary fixed point, and define $\boldsymbol{\delta}_t := \boldsymbol{x}_t - \boldsymbol{x}^\star$. Linearizing the update map to first order gives

$$\boldsymbol{\delta}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{J}_t)\boldsymbol{\delta}_t, \qquad \boldsymbol{J}_t := DF_t(\boldsymbol{x}^\star). \tag{3}$$

For a deterministic update, an update rule is *linearly stable* if $\sup_{t\geq 0}\|\boldsymbol{\delta}_t\| < \infty$ for every $\boldsymbol{\delta}_0$. If $\boldsymbol{J}_t = \boldsymbol{J} \succeq 0$ is symmetric and time invariant, this is equivalent to $|1 - \eta\lambda_i(\boldsymbol{J})| \leq 1$ for all i.

Most works evaluate these stability notions on a local quadratic model (Cohen et al., 2021; Muehlebach & Jordan, 2021; Wu et al., 2018; Song et al., 2026). For GD, $\boldsymbol{H}^\star := \nabla^2 L(\boldsymbol{x}^\star) \succeq 0$ is the Jacobian $\boldsymbol{J}$, so $\eta\lambda_1(\boldsymbol{H}^\star) = 2$ separates linear stability from instability. Neural-network training often approaches and remains near this "edge," hence the Edge-of-Stability (Cohen et al., 2021).

For a stochastic update, a bounded expected displacement can mask stochastically growing displacements. Following Song et al. (2026), we instead require *mean-square linear stability*: $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{\delta}_t\|^2 < \infty$ for every $\boldsymbol{\delta}_0$. For deterministic optimizers the expectation is redundant, and the definition reduces to linear stability. For ZO-GD, $\boldsymbol{u}_t$ is resampled at every iteration, so even the full-batch update and its Jacobian are random. Applying mean-square stability to this random Jacobian yields a boundary that depends on the full Hessian spectrum (see Eq. (22)), becomes trace-dominated at high effective rank, and is empirically approached by ZO trajectories (Song et al., 2026).

2.2 Linear perturbation stability away from stationarity

Active clipping is observed away from stationarity. At a stationary point the aggregate gradient vanishes, so batch clipping is inactive and has the unclipped local derivative; the same holds for per-sample clipping when all sample gradients are inactive. To study how active clipping changes local perturbation growth, we therefore move to nonstationary training checkpoints, following the usual pointwise EoS convention of evaluating local stability using $\boldsymbol{H}(\boldsymbol{x}_t)$ along a nonstationary trajectory. Specifically, fix a checkpoint $\bar{\boldsymbol{x}}$, introduce an infinitesimal perturbation $\boldsymbol{e}_0$, and let $\boldsymbol{e}_t$ denote its evolution under the local linearized update, with the checkpoint-dependent quantities evaluated at $\bar{\boldsymbol{x}}$. This gives

$$\boldsymbol{e}_{t+1} = \big(\boldsymbol{I} - \eta \boldsymbol{J}_t\big)\boldsymbol{e}_t, \qquad \boldsymbol{J}_t := DF_t(\bar{\boldsymbol{x}}). \tag{4}$$

Equation (4) has the same Jacobian recursion as stationary linear stability in Eq. (3). The difference is that $\boldsymbol{J}_t$ is evaluated at a nonstationary point, where clipping may be active and its derivative therefore enters the Jacobian. For FO this gives a fixed deterministic matrix; for ZO, fresh directions $\boldsymbol{u}_t$ are drawn independently while $\bar{\boldsymbol{x}}$ remains fixed. We call boundedness of Eq. (4) *linear perturbation stability*, or *mean-square linear perturbation stability* for a random update. Setting $\bar{\boldsymbol{x}} = \boldsymbol{x}^\star$ recovers the corresponding stationary definitions. At this fixed $\bar{\boldsymbol{x}}$, the deterministic FO and stochastic ZO critical step sizes for update rule m are

$$\eta_m^\star := \sup\left\{\eta > 0 : \sup_{t\geq 0}\|\boldsymbol{e}_t\| < \infty \text{ for every } \boldsymbol{e}_0\right\}, \text{ and} \tag{5}$$

$$\eta_{m,\mathrm{ms}}^\star := \sup\left\{\eta > 0 : \sup_{t\geq 0}\mathbb{E}\|\boldsymbol{e}_t\|^2 < \infty \text{ for every } \boldsymbol{e}_0\right\}. \tag{6}$$

We use this notion of stability in Section 3 to study the FO thresholds: $\eta_{\mathrm{batch}}^\star$ and $\eta_{\mathrm{ps}}^\star$. On the other hand, EoS for clipped ZO updates occur beyond this linear stability and requires studying a finite displacement for the nonlinear ZO balance in Section 4.1.2.

3 First-order clipping shifts the stability edge

We apply the framework of Section 2.2 to analyze the stability edge of clipped FO updates. At a nonstationary checkpoint $\bar{\boldsymbol{x}}$, the edge is governed by the update Jacobian $\boldsymbol{J}$ in Eq. (4), rather than the Hessian $\boldsymbol{H}$ of unclipped GD. If $\boldsymbol{J}$ is symmetric positive semidefinite, the first unstable mode gives $\eta^\star = 2/\lambda_1(\boldsymbol{J})$. Thus, FO clipping moves, rather than replaces, the linear perturbation stability edge. For batch clipping, the nonzero spectrum of $\boldsymbol{J}$ admits a symmetric representation; for per-sample clipping, it generally does not.

3.1 FO-GD WITH BATCH CLIPPING

For $\boldsymbol{g} \neq 0$, define $\alpha = \min\{1, C/\|\boldsymbol{g}\|\}$, $\chi = \mathbf{1}\{\|\boldsymbol{g}\| > C\}$, and $\boldsymbol{Q} = \boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$. Under active clipping, the direction is $C\widehat{\boldsymbol{g}} = \alpha\boldsymbol{g}$: normalization removes the response along $\widehat{\boldsymbol{g}}$ and scales orthogonal changes by α. When $\boldsymbol{g} = 0$, set $\alpha = 1$ and interpret $\chi\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$ as zero. Away from the clipping boundary, differentiating the update direction gives

$$\boldsymbol{J}_{\text{batch}} = \alpha(\boldsymbol{I} - \chi\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top)\boldsymbol{H} = \begin{cases} \boldsymbol{H}, & \|\boldsymbol{g}\| < C, \\ \alpha\boldsymbol{Q}\boldsymbol{H}, & \|\boldsymbol{g}\| > C. \end{cases} \tag{7}$$

Although $\alpha\boldsymbol{Q}\boldsymbol{H}$ is nonsymmetric, its nonzero spectrum equals that of $\alpha\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}$ (Lemma C.7); hence the edge is determined by a real curvature eigenvalue.

Theorem 3.1 (Batch FO stability edge)**.** *Assume $\boldsymbol{H} \succeq 0$ and $\|\boldsymbol{g}\| \neq C$. The nonzero spectrum of the Jacobian is real and nonnegative. Define*

$$s_{\text{batch}} := \begin{cases} \lambda_1(\boldsymbol{H}) & \|\boldsymbol{g}\| < C \quad (\textit{clip inactive}), \\ \alpha\lambda_1(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}) & \|\boldsymbol{g}\| > C \quad (\textit{clip active}), \end{cases} \qquad \eta^\star_{\text{batch}} = \frac{2}{s_{\text{batch}}}. \tag{8}$$

Under active clipping, Cauchy interlacing gives, with $1/0 = +\infty$,

$$\frac{2}{\alpha\lambda_1(\boldsymbol{H})} \;\le\; \eta^\star_{\text{batch}} \;\le\; \frac{2}{\alpha\lambda_2(\boldsymbol{H})}. \tag{9}$$

Under active clipping, the exact object is $s_{\text{batch}} = \alpha\lambda_1(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q})$ and the easier proxy is $\alpha\lambda_1(\boldsymbol{H})$. Interlacing places the former between $\alpha\lambda_2(\boldsymbol{H})$ and $\alpha\lambda_1(\boldsymbol{H})$, so its error relative to the proxy is bounded by the relative top eigengap. This gap is at most 1.11% in our runs (Appendix C.16), making the proxy accurate in practice (e.g., Figure 1).

Under active clipping, substituting $\alpha = C/\|\boldsymbol{g}\|$ into the stability condition gives

$$\lambda_1(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}) < \frac{2\|\boldsymbol{g}\|}{\eta C}. \tag{10}$$

Thus clipping does not remove the stability edge but shifts it upward: decreasing C increases the maximum unscaled projected curvature compatible with stability. The block extension and further cluster discussion appear in Theorem C.8 and Appendix B.2.

3.2 FO-GD WITH PER-SAMPLE CLIPPING

Unlike batch clipping, per-sample clipping is applied to each sample before averaging. Let $\alpha_i = \min\{1, C/\|\boldsymbol{g}_i\|\}$ and $\chi_i = \mathbf{1}\{\|\boldsymbol{g}_i\| > C\}$, with $\alpha_i = 1$ and $\chi_i\widehat{\boldsymbol{g}}_i\widehat{\boldsymbol{g}}_i^\top = 0$ when $\boldsymbol{g}_i = 0$. Then

$$\boldsymbol{J}_{\text{ps}} = \frac{1}{n}\sum_i \alpha_i(\boldsymbol{I} - \chi_i\widehat{\boldsymbol{g}}_i\widehat{\boldsymbol{g}}_i^\top)\boldsymbol{H}_i = \boldsymbol{A} - \boldsymbol{R}, \quad \boldsymbol{A} = \frac{1}{n}\sum_i \alpha_i\boldsymbol{H}_i, \quad \boldsymbol{R} = \frac{1}{n}\sum_i \alpha_i\chi_i\widehat{\boldsymbol{g}}_i\widehat{\boldsymbol{g}}_i^\top\boldsymbol{H}_i. \tag{11}$$

Here $\boldsymbol{A}$ is weighted curvature and $\boldsymbol{R}$ is the sample-gradient responses to be removed. Writing $\boldsymbol{R}_s = (\boldsymbol{R} + \boldsymbol{R}^\top)/2$ and $\boldsymbol{R}_a = (\boldsymbol{R} - \boldsymbol{R}^\top)/2$, we have $\boldsymbol{R} = \boldsymbol{R}_s + \boldsymbol{R}_a$, its symmetric—skew-symmetric decomposition. Each sample has its own gradient direction and Hessian, so these terms do not combine into one symmetric compression as in batch clipping; hence $\boldsymbol{J}_{\text{ps}}$ is generally nonsymmetric and may have complex eigenvalues.

For any clipped or unclipped linearization of the form in Eq. (4), a mode with eigenvalue λ is multiplied by $1 - \eta\lambda$. Since $|1 - \eta\lambda|^2 = 1 - 2\eta\Re\lambda + \eta^2|\lambda|^2$, it is contractive exactly when $\Re\lambda > 0$ and $\eta < 2\Re\lambda/|\lambda|^2$. If any nonzero eigenvalue has $\Re\lambda \le 0$, no positive step is stable and we set $\eta^\star(\boldsymbol{J}) = 0$; otherwise, we define

$$\eta^\star(\boldsymbol{J}) := \min_{\lambda \in \text{spec}(\boldsymbol{J})\setminus\{0\}} \frac{2\Re\lambda}{|\lambda|^2}. \tag{12}$$

When $\boldsymbol{J}$ is symmetric positive semidefinite, this reduces to $\eta^\star(\boldsymbol{J}) = 2/\lambda_1(\boldsymbol{J})$. For the per-sample update, $\boldsymbol{J} = \boldsymbol{A} - \boldsymbol{R}$.

A symmetric proxy. Since Eq. (12) is computationally intractable, we let $\boldsymbol{R} = \boldsymbol{R}_s + \boldsymbol{R}_a$ to define the tractable symmetric proxy $\lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$, computable by Hessian–vector products. The identity $|\lambda|^2/\Re\lambda = \Re\lambda + (\Im\lambda)^2/\Re\lambda$ motivates the second-order error quantified below.

Proposition 3.2 (Symmetric proxy for the FO per-sample stability edge)**.** *Let $\boldsymbol{A}$, $\boldsymbol{R}_s$, and $\boldsymbol{R}_a$ be as in Eq. (11), and set $S := \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$. Suppose $S > 0$ and every nonzero eigenvalue of $\boldsymbol{A} - \boldsymbol{R}$ has positive real part. Then the per-sample critical step in Eq. (5) is $\eta^\star_{\mathrm{ps}} = \eta^\star(\boldsymbol{A} - \boldsymbol{R})$. If $\lambda^\star$ attains the minimum in Eq. (12), then*

$$\frac{2}{\eta^\star_{\mathrm{ps}}} - S \;\le\; \frac{\|\boldsymbol{R}_a\|_2^2}{\Re\lambda^\star}.$$

If, in addition, the leading cluster of $\boldsymbol{A} - \boldsymbol{R}_s$ has width ϵ and is separated from the remaining spectrum by a gap Γ_c, with $4\|\boldsymbol{R}_a\|_2 \le \Gamma_c$ as in Proposition C.28, set $\Delta_c := \max\{\|\boldsymbol{R}_a\|_2^2/\Re\lambda^\star,\ \epsilon + 4\|\boldsymbol{R}_a\|_2^2/\Gamma_c\}$. Whenever $\Delta_c < S$,

$$\left|\frac{2}{\eta^\star_{\mathrm{ps}}} - S\right| \le \Delta_c, \qquad \frac{2}{S + \Delta_c} \le \eta^\star_{\mathrm{ps}} \le \frac{2}{S - \Delta_c}. \tag{13}$$

The proxy error is therefore quadratic in the skew component: the first bound is one-sided without a spectral gap, while the cluster condition gives Eq. (13). In Section 5, $\|\boldsymbol{R}_a\|_2$ is about 1% of S and the relative proxy error $|2/\eta^\star_{\mathrm{ps}} - S|/S$ is at most 0.015%. Further details appear in Appendices C.15 and B.2.

FO clipping changes the local operator but retains the linear-edge mechanism: batch uses $\alpha\boldsymbol{QHQ}$, while per-sample uses a nonsymmetric Jacobian with proxy $\lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$. Section 5 shows trajectories approach both clipping-aware edges.

4 Zeroth-order clipping: from instability to a nonlinear balance

Unlike the deterministic FO Jacobian, the ZO Jacobian remains random at a fixed checkpoint because a new search direction $\boldsymbol{u}_t$ is drawn at every step. We therefore track perturbation covariance rather than only its mean. For batch clipping, we first compute the mean-square linear stability boundary from the covariance operator and show that trajectories remain bounded beyond it. Under gradient–Hessian eigenvector alignment, we obtain a scalar characterization whose leading term is independent of the clipping strength. We further show that the boundary is only second-order sensitive to alignment error, explaining why this scalar remains accurate at our non-aligned checkpoints. We then retain the full clipping operation in a local quadratic model and derive a locally attracting nonzero covariance balance that explains bounded behavior beyond the linear edge. Finally, for per-sample clipping, distinct sample gates prevent the batch scalar reduction and lead instead to a matrix-valued balance.

For fixed $\boldsymbol{u}$ and sufficiently smooth losses, Taylor expansion gives

$$d_i^\mu(\boldsymbol{x}; \boldsymbol{u}) = \boldsymbol{u}^\top \boldsymbol{g}_i + O(\mu^2), \qquad \nabla_{\boldsymbol{x}} d_i^\mu(\boldsymbol{x}; \boldsymbol{u}) = \boldsymbol{H}_i \boldsymbol{u} + O(\mu^2).$$

Thus, in the leading-order $\mu \to 0$ model, differentiating the ZO update with respect to $\boldsymbol{x}$ evaluates clip_C' at $\boldsymbol{u}^\top \boldsymbol{g}_i$. It is one below the threshold and zero after saturation, since a capped output no longer responds to an infinitesimal input change. Unlike the continuous rescaling in FO, this binary derivative gives the gates

$$a(\boldsymbol{u}) = \mathbf{1}\{|\boldsymbol{u}^\top \boldsymbol{g}| < C\}, \qquad a_i(\boldsymbol{u}) = \mathbf{1}\{|\boldsymbol{u}^\top \boldsymbol{g}_i| < C\}, \tag{14}$$

and, for one sample,

$$D_{\boldsymbol{x}}\big[\mathrm{clip}_C(\boldsymbol{u}^\top \boldsymbol{g}_i(\boldsymbol{x}))\boldsymbol{u}\big] = a_i(\boldsymbol{u})\, \boldsymbol{u}(\boldsymbol{H}_i \boldsymbol{u})^\top = a_i(\boldsymbol{u}) \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{H}_i. \tag{15}$$

Exact finite-μ Jacobians and their convergence to this model are given in Appendix D.

4.1 ZO-GD with batch clipping

4.1.1 Mean-square linear perturbation stability

Fix a nonstationary checkpoint $\bar{\boldsymbol{x}}$, hold its gradient and Hessian fixed, and resample $\boldsymbol{u}_t$, recomputing the gate $a(\boldsymbol{u}_t)$ in Eq. (14). As in Section 2.2, an infinitesimal perturbation $\boldsymbol{e}_t$ obeys

$$\boldsymbol{e}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{J}_t)\boldsymbol{e}_t, \qquad \boldsymbol{J}_t = DF(\bar{\boldsymbol{x}}; \boldsymbol{u}_t),$$

where F is the clipped ZO direction. Since the newly sampled $\boldsymbol{J}_t$ is independent of $\boldsymbol{e}_t$, the covariance $\boldsymbol{V}_t = \mathbb{E}[\boldsymbol{e}_t \boldsymbol{e}_t^\top]$ follows the closed recursion

$$\boldsymbol{V}_{t+1} = \mathcal{T}_\eta(\boldsymbol{V}_t), \qquad \mathcal{T}_\eta(\boldsymbol{V}) = \mathbb{E}\big[(\boldsymbol{I} - \eta \boldsymbol{J}_t)\boldsymbol{V}(\boldsymbol{I} - \eta \boldsymbol{J}_t)^\top\big].$$

Mean-square linear perturbation stability holds when $\rho(\mathcal{T}_\eta) < 1$; the first crossing $\rho(\mathcal{T}_\eta) = 1$ defines the alignment-free boundary $\eta^\star_{\mathrm{batch,ms}}$.

For batch clipping, Eq. (15) gives $\boldsymbol{J}_t = \boldsymbol{J}_{\mathrm{batch}}(\boldsymbol{u}_t) = a(\boldsymbol{u}_t)\boldsymbol{u}_t\boldsymbol{u}_t^\top \boldsymbol{H}$ and

$$\mathcal{T}_\eta(\boldsymbol{V}) = \boldsymbol{V} - \eta(\boldsymbol{K}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{K}^\top) + \eta^2 \mathbb{E}[\boldsymbol{J}_{\mathrm{batch}} \boldsymbol{V} \boldsymbol{J}_{\mathrm{batch}}^\top], \quad \boldsymbol{K} = \mathbb{E}[\boldsymbol{J}_{\mathrm{batch}}].$$

We numerically compute the boundary $\eta^\star_{\mathrm{batch,ms}}$ directly from the full covariance operator taken from 24 practical training checkpoints. At late training phase, we find that the step sizes used in 17 checkpoints lie beyond their $\eta^\star_{\mathrm{batch,ms}}$, while their trajectories remain bounded, with the actual step size reaching about 2.13 times the boundary. To understand why the mean-square linear perturbation stability fails to capture the effect of clipping, we next impose gradient–Hessian eigenvector alignment. Because $a^2 = a$, the gate probability $p = \mathbb{E}[a]$ then enters contraction as ηp and diffusion as $\eta^2 p$, cancelling from their leading ratio. The following theorem makes this mechanism explicit through a scalar characterization.

Theorem 4.1 (Scalar threshold under gradient–Hessian eigenvector alignment)**.** *Let $\boldsymbol{H} \succeq 0$ have eigenpairs $(\lambda_j, \boldsymbol{v}_j)$, and suppose $\widehat{\boldsymbol{g}} = \boldsymbol{v}_k$ with $\lambda_k > 0$. Define $\tau_\perp = \sum_{j \neq k} \lambda_j = \mathrm{Tr}(\boldsymbol{H}) - \widehat{\boldsymbol{g}}^\top \boldsymbol{H} \widehat{\boldsymbol{g}} > 0$ and $\lambda^\perp_{\max} = \max_{j \neq k} \lambda_j$. Let $r(\eta) \geq 0$ denote the radial correction from the gated gradient direction as in Eq. (81). On the branch $\eta \lambda^\perp_{\max} < 1$ where $r(\eta)$ is finite, the exact critical step size $\eta^\star_{\mathrm{batch,ms}}$ as in Eq. (6) is the smallest positive root of*

$$\frac{\eta}{2}\big(1 + r(\eta)\big) \sum_{j \neq k} \frac{\lambda_j}{1 - \eta \lambda_j} = 1.$$

Define $\bar{c} = 1 + r(2/\tau_\perp)$. If $\bar{c}$ is finite, then

$$\frac{2}{\bar{c}\,\tau_\perp + 2\lambda^\perp_{\max}} \leq \eta^\star_{\mathrm{batch,ms}} \leq \frac{2}{\tau_\perp}. \tag{16}$$

The exact alignment assumption fails at all 24 training checkpoints, so the theorem does not give their exact boundaries. Nevertheless, the scalar predictor values $2/(\mathrm{Tr}(\boldsymbol{H}) - \widehat{\boldsymbol{g}}^\top \boldsymbol{H} \widehat{\boldsymbol{g}})$ differ from the full-operator boundaries $\eta^\star_{\mathrm{batch,ms}}$ by only 0.042%–0.440%.

When $\bar{c} - 1 \ll 1$ and $\lambda^\perp_{\max}/\tau_\perp \ll 1$, the aligned predictor reduces to the clipping-independent leading form $\eta\tau_\perp \simeq 2$; its exact correction retains C-dependence through $r(\eta)$. The derivation and off-alignment analysis appear in Appendices C.12 and C.13.

4.1.2 Beyond the linear threshold: a nonlinear covariance balance

The preceding analysis linearizes the update at a nonstationary checkpoint to track an infinitesimal perturbation. However, such analysis fails to capture the dependence on clipping, so here we do not linearize the update: we retain the full nonlinear clipped update in a local quadratic model around a stationary point $\boldsymbol{x}^\star$. Let $\boldsymbol{\delta}_t = \boldsymbol{x}_t - \boldsymbol{x}^\star$ and $\boldsymbol{W}_t = \mathbb{E}[\boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top]$. The values $\mathrm{clip}_C(d)$, rather than only their derivative gates, then remain in the dynamics, so the amount of clipping changes with the displacement amplitude. At $\boldsymbol{W} = 0$ clipping is inactive, so stability of the zero state is governed by the unclipped mean-square threshold by Song et al. (2026), unlike $\eta^\star_{\mathrm{batch,ms}}$ obtained by linearizing at a generally nonstationary checkpoint where clipping may be active.

Scalar balance. Batch-clipped trajectories in the late training remain bounded beyond the linearized mean-square threshold because growing fluctuations make clipping suppress the squared random update more strongly than the mean contraction, producing a finite-covariance balance.

Consider the local quadratic model around a stationary point, $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{\delta}$. Within this model, the clipped recursion is exact:

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \eta\, \mathrm{clip}_C(\boldsymbol{u}_t^\top \boldsymbol{H} \boldsymbol{\delta}_t)\boldsymbol{u}_t, \qquad \boldsymbol{u}_t \sim \mathcal{N}(0, \boldsymbol{I}_d). \tag{17}$$

For a fixed displacement, Gaussian isotropy gives $d = \boldsymbol{u}^\top \boldsymbol{g} = \|\boldsymbol{g}\|\zeta$ with $\zeta \sim \mathcal{N}(0,1)$. Thus the calculation depends only on $\kappa = C/\|\boldsymbol{g}\|$. Define $p(\kappa) = \Pr(|\zeta| < \kappa) = \mathbb{E}[\mathrm{clip}'_C(d)]$, $\mu_2(\kappa) = \mathbb{E}[\min(\zeta^2, \kappa^2)] = \mathbb{E}[\mathrm{clip}_C(d)^2]/\|\boldsymbol{g}\|^2$, and $\mu_4(\kappa) = \mathbb{E}[\min(\zeta^2, \kappa^2)\zeta^2]$. Gaussian isotropy and Stein's identity give

$$\mathbb{E}[\mathrm{clip}_C(\boldsymbol{g}^\top \boldsymbol{u})\boldsymbol{u}] = p(\kappa)\boldsymbol{g}, \tag{18}$$

$$\mathbb{E}[\mathrm{clip}_C(\boldsymbol{g}^\top \boldsymbol{u})^2 \boldsymbol{u}\boldsymbol{u}^\top] = \|\boldsymbol{g}\|^2 \left[\mu_4 \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + \mu_2(\boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top)\right]. \tag{19}$$

These two moments determine the covariance update below.

Assumption 4.2 (Amplitude self-averaging)**.** *The clip moments depend on the random amplitude $\|\boldsymbol{g}\| = \|\boldsymbol{H}\boldsymbol{\delta}\|$. To close the dynamics in $\boldsymbol{W} = \mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top]$, we evaluate them at $\sqrt{T(\boldsymbol{W})}$, where $T(\boldsymbol{W}) = \mathbb{E}\|\boldsymbol{g}\|^2 = \mathrm{Tr}(\boldsymbol{H}^2\boldsymbol{W})$ and $\kappa(\boldsymbol{W}) = C/\sqrt{T(\boldsymbol{W})}$. Appendix B.3 gives the precise identities.*

Under this closure, the two conditional moments above give the covariance map

$$\mathcal{F}_{\mathrm{batch}}(\boldsymbol{W}) = \boldsymbol{W} - \eta p(\boldsymbol{H}\boldsymbol{W} + \boldsymbol{W}\boldsymbol{H}) + \eta^2[\mu_2 T(\boldsymbol{W})\boldsymbol{I} + (\mu_4 - \mu_2)\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}], \tag{20}$$

where all moments are evaluated at $\kappa(\boldsymbol{W})$. The negative $O(\eta)$ term contracts covariance, whereas the positive $O(\eta^2)$ terms inject diffusion; μ_2 is isotropic and $\mu_4 - \mu_2$ corrects the gradient direction. Define their drift-normalized coefficients $\gamma(\kappa) = \mu_2(\kappa)/p(\kappa)$ and $\beta(\kappa) = \eta(\mu_4(\kappa) - \mu_2(\kappa))/(2p(\kappa))$. As $T(\boldsymbol{W})$ grows, $\kappa = C/\sqrt{T(\boldsymbol{W})}$ and hence $\gamma(\kappa)$ decrease, providing negative feedback that can arrest linear growth.

Define a stationary covariance as a nonzero fixed point $\boldsymbol{W}^\star \succeq 0$ satisfying $\mathcal{F}_{\mathrm{batch}}(\boldsymbol{W}^\star) = \boldsymbol{W}^\star$. In an eigenbasis of $\boldsymbol{H}$, the off-diagonal fixed-point equation implies $W^\star_{ij} = 0$ under $\beta\lambda_1 < 1$. Thus, $\boldsymbol{W}^\star$ is diagonal in this basis. Writing $w_j = W^\star_{jj}$, the diagonal equation gives $2p\lambda_j w_j = \eta\mu_2 T(\boldsymbol{W}^\star) + \eta(\mu_4 - \mu_2)\lambda_j^2 w_j$. Solving for w_j and substituting into $T(\boldsymbol{W}^\star) = \sum_j \lambda_j^2 w_j$ yields the exact balance scalar and stationary condition

$$\tau(\kappa) = \gamma(\kappa)\sum_j \frac{\lambda_j}{1 - \beta(\kappa)\lambda_j}, \qquad \eta\tau(\kappa^\star) = 2. \tag{21}$$

At fixed η, it solves for $\kappa^\star = C/\sqrt{T(\boldsymbol{W}^\star)}$ and hence the nonzero covariance amplitude.

For comparison, let $\eta_0^\star$ denote the unclipped Gaussian-ZO mean-square critical step size at the zero state (Song et al., 2026), namely the unique root on $\eta_0^\star\lambda_1 < 1$ of

$$\frac{\eta_0^\star}{2}\sum_j \frac{\lambda_j}{1 - \eta_0^\star\lambda_j} = 1. \tag{22}$$

The nonlinear balance undergoes its bifurcation at this same step size.

Theorem 4.3 (Self-limiting batch-ZO balance)**.** *Let $\boldsymbol{H} \succ 0$ and let Assumption 4.2 hold. If $\eta \le \eta_0^\star$, the closed dynamics have only the zero stationary covariance. If $\eta > \eta_0^\star$, the origin is mean-square unstable and a unique nonzero stationary amplitude is selected by $\eta\tau(\kappa^\star) = 2$. At this balance, the relative error of the trace approximation $\gamma\,\mathrm{Tr}(\boldsymbol{H})$ satisfies*

$$0 \le \frac{\tau - \gamma\,\mathrm{Tr}(\boldsymbol{H})}{\tau} \le 2\frac{\lambda_1}{\mathrm{Tr}(\boldsymbol{H})}. \tag{23}$$

The exact scalar is τ, while the measurable proxy $\gamma\,\mathrm{Tr}(\boldsymbol{H})$ is accurate at high effective rank, where the balance becomes $\eta\gamma(\kappa^\star)\,\mathrm{Tr}(\boldsymbol{H}) \approx 2$. This is the experimental diagnostic.

For fixed η and $\boldsymbol{H}$, the balance $\eta\tau(\kappa^\star) = 2$ fixes the ratio $\kappa^\star = C/\sqrt{T(\boldsymbol{W}^\star)}$, rather than C or the covariance amplitude separately. Hence $T(\boldsymbol{W}^\star) = (C/\kappa^\star)^2$: increasing C weakens clipping and moves the balance to a larger covariance amplitude. As $C \to \infty$, this amplitude diverges and the finite balance disappears; the unclipped covariance contracts below $\eta_0^\star$ and diverges above it.

The balance can be compared with the linearized threshold in Eq. (16). At a nonzero balance, $\tau(\kappa^\star) = 2/\eta$, whereas the leading linearized boundary is $\tau_\perp = 2/\eta$. Thus $\tau(\kappa^\star) < \tau_\perp$ is equivalent

to $\eta\tau_\perp > 2$, placing the balance on the linearly unstable side of that boundary. In our runs, $\tau_\perp$ reaches $4.1 \times 2/\eta$, while the proxy $\gamma \operatorname{Tr}(\boldsymbol{H})$ remains within $[0.71, 1.00] \times 2/\eta$.

Existence does not ensure attraction. The elasticity $\mathfrak{e}$ is a computable scalar bound on how sharply the gate moments respond to amplitude; it depends only on $\kappa^\star$, $\eta\lambda_1$, and one-dimensional Gaussian moments (Appendix B.3). The condition $\mathfrak{e} \le 1$ is sufficient, not necessary or automatic: it is favored by a gradual clipping response and small $\eta\lambda_1$, and rules out feedback overshoot without forming the full Jacobian $D\mathcal{F}_{\text{batch}}$.

Theorem 4.4 (Local attraction of the batch balance)**.** *Under the assumptions of Theorem 4.3, suppose $\eta\lambda_1 \le 1/2$ and $\mathfrak{e}(\kappa^\star) \le 1$. Then the stationary covariance $\boldsymbol{W}^\star$ selected by $\kappa^\star$ is a locally exponentially attracting fixed point of $\mathcal{F}_{\text{batch}}$: for some $K < \infty$ and $q < 1$, locally,*

$$\rho(D\mathcal{F}_{\text{batch}}(\boldsymbol{W}^\star)) < 1, \qquad \|\boldsymbol{W}_t - \boldsymbol{W}^\star\| \le Kq^t \|\boldsymbol{W}_0 - \boldsymbol{W}^\star\|. \tag{24}$$

Thus, when its sufficient conditions hold, the zero covariance is unstable while the selected nonzero covariance is locally attracting. This is the theoretical boundary–balance distinction.

4.2 ZO-GD WITH PER-SAMPLE CLIPPING: WHY A MATRIX BALANCE IS NEEDED

Why there is no batch-style scalar threshold. Batch clipping has one shared gate and aggregate Hessian; per-sample clipping has distinct gates and sample Hessians. In the linearized update, $\boldsymbol{J}(\boldsymbol{u}) = n^{-1}\sum_i a_i(\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{H}_i$, the drift contains single-gate moments but the diffusion contains pairwise moments $a_i a_j$. Since distinct gates do not satisfy $a_i a_j = a$, their co-activation geometry remains in the mean-square operator (Lemma C.1), precluding a batch-style scalar threshold.

Nonlinear matrix balance. Per-sample clipping also lacks a scalar reduction in the nonlinear dynamics. Each sample's gate weights its own Hessian, so the effective drift changes in direction as well as magnitude when the gates change. The stationary state therefore requires a matrix balance.

Let $s_t = n^{-1}\sum_i \operatorname{clip}_C(d_{i,t})$, where $d_{i,t} = \boldsymbol{u}_t^\top \boldsymbol{g}_{i,t}$, so $\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \eta s_t \boldsymbol{u}_t$. Approximating the sample losses as quadratics with a common stationary point gives $\boldsymbol{g}_{i,t} = \boldsymbol{H}_i\boldsymbol{\delta}_t$. Stein's identity then gives $\mathbb{E}_{\boldsymbol{u}}[s_t\boldsymbol{u}_t] = \boldsymbol{H}_p(\boldsymbol{\delta}_t)\boldsymbol{\delta}_t$, where $\boldsymbol{H}_p(\boldsymbol{\delta}) = n^{-1}\sum_i p_i(\boldsymbol{\delta})\boldsymbol{H}_i$ and $p_i(\boldsymbol{\delta}) = \Pr_{\boldsymbol{u}}(|\boldsymbol{u}^\top \boldsymbol{H}_i\boldsymbol{\delta}| < C)$. Because p_i varies across samples, $\boldsymbol{H}_p$ is generally not a scalar multiple of $\boldsymbol{H}$; no single effective clipping multiplier remains.

Under Assumption B.2, the gate moments are approximated from $\boldsymbol{W}$, yielding $\mathcal{F}_{\text{ps}}(\boldsymbol{W}) = \boldsymbol{W} - \eta(\boldsymbol{H}_p(\boldsymbol{W})\boldsymbol{W} + \boldsymbol{W}\boldsymbol{H}_p(\boldsymbol{W})) + \eta^2\boldsymbol{\Sigma}_s(\boldsymbol{W})$. Hence, a stationary covariance satisfies

$$\boldsymbol{H}_p(\boldsymbol{W}^\star)\boldsymbol{W}^\star + \boldsymbol{W}^\star\boldsymbol{H}_p(\boldsymbol{W}^\star) = \eta\boldsymbol{\Sigma}_s(\boldsymbol{W}^\star). \tag{25}$$

Thus contraction and diffusion match direction by direction. For eigenpairs $\boldsymbol{H}_p\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$ with $\lambda_i > 0$, diagonal projection gives

$$\boldsymbol{v}_i^\top \boldsymbol{W}^\star \boldsymbol{v}_i = \frac{\eta \boldsymbol{v}_i^\top \boldsymbol{\Sigma}_s \boldsymbol{v}_i}{2\lambda_i}, \qquad \frac{\sum_i \boldsymbol{v}_i^\top \boldsymbol{\Sigma}_s \boldsymbol{v}_i/\lambda_i}{\sum_i \boldsymbol{v}_i^\top \boldsymbol{W}^\star \boldsymbol{v}_i} = \frac{2}{\eta}. \tag{26}$$

The first equality implies the ratio $2/\eta$ separately for every direction with nonzero variance; summation is not required theoretically. We use the ratio of sums as a computable experimental diagnostic: over leading eigenvectors of $\boldsymbol{H}_p$, it aggregates noisy windowed estimates of $\boldsymbol{W}^*$ and $\boldsymbol{\Sigma}_s$ while avoiding small-λ_i directions. This is a balance diagnostic, not a transition criterion.

Proposition 4.5 (Existence of a nonzero per-sample covariance balance)**.** *Under the per-sample self-averaging closure in Assumption B.2 and the isotropic reduction $\boldsymbol{\Sigma}_s(\boldsymbol{W}) = M_2(\boldsymbol{W})\boldsymbol{I}$, assume $\boldsymbol{H}_i \succeq 0$ and $\boldsymbol{H} = n^{-1}\sum_i \boldsymbol{H}_i \succ 0$. Suppose M_2 is continuous and nondecreasing along the covariance path used in the fixed-point construction: $\boldsymbol{W} \preceq \boldsymbol{W}'$ implies $M_2(\boldsymbol{W}) \le M_2(\boldsymbol{W}')$ there (Appendix C.7). If $\eta \operatorname{Tr}(\boldsymbol{H}) > 2$, Eq. (25) has a nonzero solution.*

Thus a nonzero stationary covariance exists under the isotropic reduction and pathwise monotonicity condition; the trace condition is sufficient, not a transition threshold. Attraction requires the stronger local condition below.

Theorem 4.6 (Local attraction of the per-sample balance)**.** *Under Assumption B.2 and $\boldsymbol{\Sigma}_s(\boldsymbol{W}) = M_2(\boldsymbol{W})\boldsymbol{I}$, let $\boldsymbol{W}^\star \succ 0$ solve (25) and let $\boldsymbol{H}_p$ and M_2 be continuously differentiable nearby. Suppose*

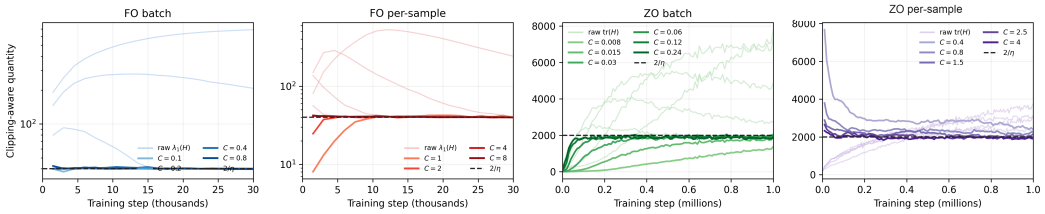

Figure 1: Clipping-aware proxies versus $2/\eta$ (dashed). From left: FO batch $\alpha\lambda_1(\boldsymbol{H})$, $\alpha = \min\{1, C/\|\boldsymbol{g}\|\}$; FO per-sample $\lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$; batch ZO $\gamma \operatorname{Tr}(\boldsymbol{H})$; and per-sample ZO the ratio in Eq. (26) over the leading 16 directions. Pale curves are raw-curvature baselines. FO tracks clipping-aware edges, whereas ZO approaches nonzero fluctuation balances.

$D\mathcal{F}_{\mathrm{ps}}(\boldsymbol{W}^\star)[\boldsymbol{\Delta}] \succeq 0$ for every $\boldsymbol{\Delta} \succeq 0$. If $D\mathcal{F}_{\mathrm{ps}}(\boldsymbol{W}^\star)[\boldsymbol{W}^\star] \prec \boldsymbol{W}^\star$, then $\rho(D\mathcal{F}_{\mathrm{ps}}(\boldsymbol{W}^\star)) < 1$, and $\boldsymbol{W}^\star$ is locally exponentially attracting: for some $K < \infty$ and $q \in (0,1)$, $\|\boldsymbol{W}_t - \boldsymbol{W}^\star\| \le Kq^t\|\boldsymbol{W}_0 - \boldsymbol{W}^\star\|$ for all $\boldsymbol{W}_0$ sufficiently close to $\boldsymbol{W}^\star$.

The theorem gives a sufficient, restrictive guarantee: isotropic diffusion and positivity of $D\mathcal{F}_{\mathrm{ps}}$ are not verified at our checkpoints, so it does not certify their attraction. We instead test the directional diagnostic in Eq. (26); the one-step loss balance gives a complementary check (Appendix B.5).

5 Experiments

We adapt the CIFAR-10 experiment with a CNN from Song et al. (2026) to a 1,000-example, four-class subset (Krizhevsky, 2009) and add our clipping schemes and diagnostics. Figure 1 tests the FO batch and per-sample edges, the batch-ZO nonlinear balance, and the per-sample directional balance. Appendices F and C.16 give the protocol and complete results.

Both FO proxies approach the predicted edge. Batch gradient-direction removal changes the exact scalar by less than 1.1%, making $\alpha\lambda_1(\boldsymbol{H})$ accurate. Under strong per-sample clipping, $\lambda_1(\boldsymbol{A})$ overestimates the edge by 24–35%, whereas $S = \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$ reduces the relative error $|2/\eta^\star_{\mathrm{ps}} - S|/S$ to 0.012%.

For batch ZO, raw $\operatorname{Tr}(\boldsymbol{H})$ reaches $4.16 \times 2/\eta$, while $\gamma \operatorname{Tr}(\boldsymbol{H})/(2/\eta)$ remains in $[0.71, 1.00]$. The strongest clip gives the evolving lower endpoint 0.71, suggesting it has not yet reached balance.

For per-sample ZO, we evaluate the ratio in Eq. (26) over the leading 16 eigenvectors of $\boldsymbol{H}_p$, replacing $\boldsymbol{W}^\star$ and $\boldsymbol{\Sigma}_s$ with estimates from single-trajectory time windows. We omit small-λ_i directions because they equilibrate slowly. Smaller-C curves remain farther above the predicted $2/\eta$: stronger clipping reduces update magnitudes and slows equilibration without changing the balance. The trajectories empirically approach this matrix balance, although the sufficient conditions of Proposition 4.5 and Theorem 4.6 are not verified at these checkpoints.

6 Conclusion

For FO, clipping shifts the linear-stability edge: batch clipping gives a rescaled, projected curvature and per-sample clipping a sample-weighted, generally nonsymmetric Jacobian; both proxies approach $2/\eta$. ZO instead separates linearized instability from a bounded nonlinear balance. Although batch-ZO linearization predicts $\eta\tau_\perp \simeq 2$, trajectories remain bounded up to $\tau_\perp = 4.1 \times 2/\eta$. Retaining amplitude-dependent clipping yields the batch balance $\eta\gamma \operatorname{Tr}(\boldsymbol{H}) \approx 2$ and the per-sample matrix balance in Eq. (25), which our experiments follow rather than the linearized boundary.

Future work should extend these balances beyond fixed quadratic geometry and self-averaging. Our guarantees are local rather than pathwise or global, and the per-sample results require order and cone conditions; broader empirical validation is also needed.

AI USE STATEMENT

We used ChatGPT to polish the writing of the manuscript; assist in identifying potentially relevant prior work and drafting the related-work section in the appendix; support mathematical derivations and the exploration of proof strategies; and assist with limited experimental code development, including the implementation of clipping, plot generation, and analysis of experimental results. We also used ChatGPT during research ideation to investigate whether a scalar balancing factor could be derived in the per-sample zeroth-order setting. All AI-assisted text, references, derivations, code, experimental results, and interpretations were independently reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318. URL https://arxiv.org/abs/1607.00133.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International Conference on Machine Learning*, 2022.

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

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, 2020.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, 2022.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *ACM Workshop on Artificial Intelligence and Security*, 2017.

Jeremy Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.00065.

Oswaldo L. V. Costa and Marcelo D. Fragoso. Stability results for discrete-time linear systems with markovian jumping parameters. *Journal of Mathematical Analysis and Applications*, 179(1): 154–178, 1993. doi: 10.1006/jmaa.1993.1341. URL https://doi.org/10.1006/jmaa.1993.1341.

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.

John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Wenzhi Fang, Ziyi Yu, Yuning Jiang, Yuanming Shi, Colin N. Jones, and Yong Zhou. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing*, 70:5058–5073, 2022.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/abd1c782880cc59759f4112fda0b8f98-Abstract.html.

Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *Journal of Machine Learning Research*, 23(173):1–65, 2022.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2013.

Kevin G. Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, 2012.

Tosio Kato. *Perturbation Theory for Linear Operators*. Springer, 2nd edition, 1995.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17343–17363. PMLR, 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

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

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bH-kCY6LdKg.

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

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.

Chao Ma and Lexing Ying. On linear stability of SGD and input-smoothness of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*, 2023. URL https://arxiv.org/abs/2305.17333.

Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. To clip or not to clip: the dynamics of sgd with gradient clipping in high-dimensions, 2024. URL https://arxiv.org/abs/2406.11733.

Michael Muehlebach and Michael Jordan. A dynamical systems perspective on Nesterov acceleration. In *International Conference on Machine Learning*, 2019.

Michael Muehlebach and Michael I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22(73):1–50, 2021.

Rotem Mulayoff and Tomer Michaeli. Exact mean square linear stability analysis for SGD. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3915–3969. PMLR, 2024. URL https://proceedings.mlr.press/v247/mulayoff24a.html.

Rotem Mulayoff and Sebastian U. Stich. On the stability of nonlinear dynamics in GD and SGD: Beyond quadratic potentials. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5210–5243. PMLR, 2026. URL https://proceedings.mlr.press/v336/mulayoff26a.html.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 1310–1318. PMLR, 2013. URL https://proceedings.mlr.press/v28/pascanu13.html.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1504–1512. PMLR, 2021. URL https://proceedings.mlr.press/v130/qian21a.html.

Zhen Qin, Daoyuan Chen, Bingchen Qian, Bolin Ding, Yaliang Li, and Shuiguang Deng. Federated full-parameter tuning of billion-sized language models with communication cost under 18 kilobytes. In *International Conference on Machine Learning*, 2024.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. In *Advances in Neural Information Processing Systems*, 2023.

Minhak Song, Liang Zhang, Bingcong Li, Niao He, Michael Muehlebach, and Sewoong Oh. Zeroth-order optimization at the edge of stability. *arXiv preprint arXiv:2604.14669*, 2026. URL https://arxiv.org/abs/2604.14669.

Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch SGD via generating functions: Conditions of convergence, phase transitions, benefit from negative momenta. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bzaPGEllsjE.

Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, 2018.

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*, 2022.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/6651526b6fb8f29a00507de6a49ce30f-Abstract.html.

Mengwei Xu, Dongqi Cai, Yaozong Wu, Xiang Li, and Shangguang Wang. FwdLLM: Efficient federated finetuning of large language models with perturbed inferences. In *USENIX Annual Technical Conference*, 2024.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgnXpVYwS.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. Dpzero: Private fine-tuning of language models without backpropagation. In *International Conference on Machine Learning*, 2024a.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, Michael Muehlebach, and Niao He. Zeroth-order optimization finds flat minima. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5375. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/ebc62a3af9342eb4ebc728e5c5bc4cca-Abstract-Conference.html.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *International Conference on Machine Learning*, 2024b.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *International Conference on Learning Representations*, 2023.

Liu Ziyin, Botao Li, Tomer Galanti, and Masahito Ueda. Type-ii saddles and probabilistic stability of stochastic gradient descent. *arXiv preprint arXiv:2303.13093*, 2023.

A Extended related work

This section places our results in four bodies of work: the theory of gradient clipping, dynamical stability analyses of optimization algorithms, the edge of stability, and zeroth-order optimization. Throughout, *stability* means dynamical stability of an optimization update near a reference trajectory or stationary point. This is distinct from algorithmic stability, which studies the sensitivity of a learned model to replacing training examples and is often used to derive generalization bounds.

A.1 Theory of gradient clipping

Gradient clipping was introduced in deep learning as a practical response to exploding gradients, particularly in recurrent networks (Pascanu et al., 2013). It has since become a standard optimization primitive and also plays a structural role in private learning: DP-SGD clips each example gradient to bound sensitivity before adding noise (Abadi et al., 2016). These two uses already highlight distinct effects of clipping. Batch clipping controls the magnitude of one aggregate update, whereas per-sample clipping changes how individual gradients are combined and can therefore change the update direction.

Existing theory has primarily studied the convergence, adaptivity, bias, or robustness produced by this modification. Under a relaxed smoothness model in which local smoothness can grow with the gradient norm, Zhang et al. (2020) interpret clipping as adaptive rescaling and derive convergence guarantees that can improve over those of fixed-step GD in that setting. For finite-sum objectives, Qian et al. (2021) show that clipping component gradients introduces a directional bias controlled by their norms and their angles with the full gradient. Koloskova et al. (2023) give tight convergence guarantees with explicit dependence on an arbitrary clipping threshold: in their setting, the threshold affects higher-order terms for deterministic GD, while stochastic clipping can introduce a bias that prevents convergence to the exact optimum under standard noise assumptions. A complementary line uses clipping to obtain high-probability guarantees under heavy-tailed stochastic gradients (Gorbunov et al., 2020). More recently, Marshall et al. (2024) derive deterministic high-dimensional equations for the loss dynamics of clipped streaming SGD in least-squares models and use them to quantify the interaction among clipping, noise, and problem geometry.

Unlike these analyses of convergence rates, stochastic bias, and robustness, we study clipping through the local stability of the resulting update dynamics. We distinguish whether clipping changes the stability edge itself or stabilizes a nonzero state that is not predicted by the infinitesimal boundary. We also separate batch from per-sample clipping because clipping before aggregation can change both the direction and the stochastic dependence structure of an update.

A.2 Dynamical stability of optimization algorithms

Local stability analysis asks whether a small perturbation decays when the geometry of the objective and the state of the optimizer are frozen at a checkpoint. For a deterministic update, this reduces to a spectral-radius condition on its Jacobian. Applied to GD, it gives the familiar threshold $\eta\lambda_1(\boldsymbol{H}) < 2$; momentum and adaptive methods require augmenting the optimizer state or using the corresponding preconditioned Jacobian (Cohen et al., 2021; Muehlebach & Jordan, 2019; 2021; Cohen et al., 2022). Our FO analysis uses the same principle, but applies it to the Jacobian of the clipped update rather than to the Hessian itself.

Random updates require a stronger notion of stability because the expected perturbation may decay even while its variance grows. Mean-square stability therefore tracks the covariance of a perturbation and tests the spectral radius of the resulting lifted operator, a classical construction for random linear systems (Costa & Fragoso, 1993). This framework has been used to characterize stable minima and critical step sizes for SGD, including the effects of batch size, gradient noise, and momentum (Wu et al., 2018; Granziol et al., 2022; Velikanov et al., 2023; Mulayoff & Michaeli, 2024). Related work also considers higher-moment and probabilistic notions of stability (Ma & Ying, 2021; Ziyin et al., 2023).

Song et al. (2026) apply mean-square stability to the randomness generated by ZO search directions, even in full-batch training. They derive exact frozen stability thresholds for ZO-GD, ZO with momentum, and frozen ZO-Adam and show that unclipped trajectories approach these spectrum-

dependent boundaries. We adopt this framework as the infinitesimal baseline for clipped ZO. The key difference is that clipping depends on the amplitude of the directional derivative: linearization yields a frozen gated operator, whereas explaining the observed finite-amplitude trajectory requires retaining the nonlinear clip and analyzing a nonzero covariance balance. A shared batch clip and distinct per-sample clips lead to different covariance operators, which is why the two cases admit different reductions in our analysis.

Recent work has also shown that linearized stability need not fully determine nonlinear behavior near a minimum (Mulayoff & Stich, 2026). Our setting isolates a distinct mechanism for this separation: clipping supplies an explicit amplitude-dependent feedback, and we characterize the resulting nonzero balance rather than only the nonlinear behavior of the zero state.

A.3 Edge-of-stability dynamics

The EoS phenomenon was documented systematically by Cohen et al. (2021): during full-batch GD, the largest Hessian eigenvalue progressively rises and then remains near $2/\eta$, even though the loss becomes nonmonotone. Large-step "catapult" dynamics, in which loss spikes are followed by movement toward lower-curvature regions, provided an earlier related picture (Lewkowycz et al., 2020). Subsequent work has studied progressive sharpening along the trajectory (Wang et al., 2022), unstable or oscillatory convergence beyond the textbook monotone-descent regime (Ahn et al., 2022), and EoS in tractable neural-network models (Zhu et al., 2023). Bifurcation and trajectory-alignment analyses explain how trajectories with different initializations can approach a common EoS diagram (Song & Yun, 2023), while self-stabilization analyses describe how oscillations in the sharp direction induce slower motion that controls curvature (Damian et al., 2023). Normalized-GD analyses give a related limiting flow along a manifold of minimizers (Arora et al., 2022), and central-flow models describe the smoothed trajectory of oscillatory optimizers more broadly (Cohen et al., 2025).

The EoS viewpoint has also been adapted to optimization rules whose relevant local operator is not the raw Hessian. Full-batch adaptive methods track an edge defined by a preconditioned Hessian (Cohen et al., 2022); SAM has a gradient-dependent stability edge that is observed empirically (Long & Bartlett, 2024); and mini-batch SGD has been described through an edge of stochastic stability based on batch-dependent curvature (Andreyev & Beneventano, 2024). These works reinforce a principle used in our FO analysis: the appropriate edge must be derived from the actual update rule.

Our work asks what remains of this principle after clipping. For FO, the EoS story survives once raw curvature is replaced by the Jacobian of the clipped update: trajectories track a clipping-aware frozen edge. For batch-clipped ZO, however, the frozen mean-square edge remains an infinitesimal boundary but does not select the observed finite-amplitude trajectory. Clipping can arrest growth beyond that boundary at a nonzero covariance balance. Per-sample ZO retains a matrix-valued balance but does not support the same scalar ordering relative to its frozen edge. We therefore distinguish an *edge*, which separates decay from growth near the zero-fluctuation state, from a *balance*, which is a nonzero state selected by nonlinear clipping. Our fixed-geometry analysis does not attempt to explain progressive sharpening itself; it asks which edge or balance describes a trajectory once it reaches a given local curvature regime.

A.4 Zeroth-order optimization

Classical ZO theory studies how to optimize using only function evaluations. Random one- and two-point estimators lead to dimension-dependent convergence and query-complexity guarantees for convex, stochastic, and bandit problems (Jamieson et al., 2012; Duchi et al., 2015; Shamir, 2017; Nesterov & Spokoiny, 2017; Wang et al., 2018). This literature primarily asks how accurately and how quickly function-value queries can emulate first-order optimization. ZO methods have also been developed for applications where backpropagation is unavailable or undesirable, including black-box adversarial attacks (Chen et al., 2017; Andriushchenko et al., 2020), evolution-strategy reinforcement learning (Salimans et al., 2017), communication-efficient federated learning (Fang et al., 2022), and memory-efficient language-model fine-tuning (Malladi et al., 2023; Zhang et al., 2024b; Xu et al., 2024; Qin et al., 2024). DPZero combines the forward-only memory profile with privacy guarantees (Zhang et al., 2024a), providing an additional motivation for understanding clipping and per-sample structure in ZO updates.

More recent theory asks which solutions ZO dynamics prefer, rather than only whether they converge. Zhang et al. (2025) show that the standard two-point estimator favors minima with small Hessian trace. Most directly related, Song et al. (2026) derive exact mean-square stability thresholds for ZO-GD, ZO with momentum, and frozen ZO-Adam. Unlike deterministic FO thresholds, these depend on the full spectrum of the (possibly preconditioned) Hessian and admit trace–top-eigenvalue bounds; their experiments connect these thresholds to EoS behavior in unclipped neural-network training.

We take this unclipped mean-square boundary as the frozen baseline and study what changes after scalar directional derivatives are clipped. The main new difficulty is not the finite-difference estimator itself, but the interaction between search-direction randomness, clipping, and sample aggregation. A shared batch gate can largely cancel between the drift and diffusion terms of the frozen boundary, while the full nonlinear clip changes their relative strength at finite amplitude and produces a covariance balance. Per-sample gates retain pairwise gate moments and lead instead to a matrix-valued balance. Thus our results complement classical ZO convergence theory and recent ZO implicit-bias and EoS theory by characterizing the local dynamics created by clipping.

B Details deferred from the main text

B.1 Rule-specific fixed-geometry scope

In every stability calculation, $\boldsymbol{H}$ in the batch rules and, where they enter, the collection $\{\boldsymbol{H}_i\}$ in the per-sample rules are constant, as are η and C. Checkpoint linearisation additionally holds the checkpoint gradients and clipping coefficients fixed; for ZO it also fixes the probe-dependent gate law and its moments. The nonlinear covariance maps instead allow the gradients, clip widths, coefficients, and moments to respond to the evolving fluctuation amplitude while retaining fixed quadratic geometry.

The local FO per-sample result accepts arbitrary checkpoint values $\{\boldsymbol{g}_i, \boldsymbol{H}_i\}$ and requires no shared minimizer. The general linearized ZO operator of Lemma C.1 likewise retains arbitrary $\boldsymbol{H}_i$, whereas the explicit closed pair-moment formula of Theorem C.14 assumes $\boldsymbol{H}_i \equiv \boldsymbol{H}$; neither linearized statement requires a shared minimizer. The batch nonlinear result requires the aggregate relation $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{\delta}$, where $\boldsymbol{\delta} = \boldsymbol{x} - \boldsymbol{x}^\star$. With $\boldsymbol{W} = \mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top]$, its amplitude closure uses $\kappa(\boldsymbol{W}) = C/\sqrt{\mathrm{Tr}(\boldsymbol{H}^2\boldsymbol{W})}$. The per-sample nonlinear results additionally use $\boldsymbol{g}_i = \boldsymbol{H}_i\boldsymbol{\delta}$ about a shared stationary point, which is a shared local minimizer when every $\boldsymbol{H}_i \succeq 0$.

B.2 Additional first-order bounds

For active batch clipping, Cauchy interlacing shows that if $\lambda_2(\boldsymbol{H}) \geq (1-\epsilon)\lambda_1(\boldsymbol{H})$, then the exact compressed scalar is at least $(1-\epsilon)(C/\|\boldsymbol{g}\|)\lambda_1(\boldsymbol{H})$. Thus a degenerate top cluster forces the compression gap to be small. EoS training has been reported to drive top eigenvalues together (Cohen et al., 2021), and our runs exhibit such a cluster, but we do not claim that EoS alone produces it. Without clipping the Jacobian is simply $\boldsymbol{H}$ and the lower endpoint of the interlacing interval is already exact; the nontrivial interval in the proof of Theorem 3.1 arises only because active clipping replaces $\boldsymbol{H}$ by its compression to $\widehat{\boldsymbol{g}}^\perp$.

The active phase is normalized GD, so the sharpness-reduction flow of Arora et al. (2022) applies directly and the self-stabilization mechanism of Damian et al. (2023) may plausibly transport to $\{(C/\|\boldsymbol{g}\|)\lambda_1(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}) \leq 2/\eta\}$ in the active regime; we do not prove that transport.

For the per-sample proxy, Proposition 3.2 states both the gap-free one-sided estimate and the two-sided refinement obtained under the cluster-separation condition of Proposition C.28. The exact nonsymmetric edge requires the relevant complex spectral branch; neither the largest real part nor the largest modulus alone suffices.

B.3 Deferred batch-ZO details

Updating the gate moments along a trajectory produces a nonautonomous tangent or covariance cocycle, but $a_t^2 = a_t$ still gives drift and diffusion the same instantaneous shared-gate factor. A Lyapunov exponent of that cocycle is a different stability object and need not determine the nonzero

stationary amplitude. The distinction is visible in

$$\eta\alpha\lambda_1 = 2, \quad \eta\tau_\perp = 2, \quad \eta\gamma(\kappa)\,\mathrm{Tr}(\boldsymbol{H}) = 2, \qquad \alpha = \kappa = \frac{C}{\|\boldsymbol{H}\boldsymbol{\delta}\|},$$

where the first and third relations depend on the fluctuation amplitude while the linearized ZO relation does not. The linearized boundary remains correct for an infinitesimal perturbation; it simply does not specify the finite amplitude at which clipping arrests growth.

The self-averaging closure used in the batch nonlinear map is the following set of three identities. With $r = \|\boldsymbol{H}\boldsymbol{\delta}_t\|$ and $r_\star = \sqrt{\mathrm{Tr}(\boldsymbol{H}^2\boldsymbol{W}_t)}$,

$$\begin{aligned} \mathbb{E}[p(r)\boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top] &= p(r_\star)\boldsymbol{W}_t, & \mathbb{E}[\mu_2(r)r^2] &= \mu_2(r_\star)\,\mathrm{Tr}(\boldsymbol{H}^2\boldsymbol{W}_t), \\ \mathbb{E}[(\mu_4-\mu_2)(r)\boldsymbol{H}\boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top\boldsymbol{H}] &= (\mu_4-\mu_2)(r_\star)\boldsymbol{H}\boldsymbol{W}_t\boldsymbol{H}. \end{aligned} \tag{27}$$

They hold exactly when r is almost surely constant; otherwise they define the approximation whose error is not propagated in the main theorems.

For the appendix calculations, write $\psi(\kappa) := (\mu_4(\kappa) - \mu_2(\kappa))/p(\kappa)$. The one-dimensional moments obey

$$\gamma(\kappa) = \sqrt{\pi/2}\,\kappa - \tfrac{2}{3}\kappa^2 + O(\kappa^3), \qquad \psi(\kappa) = \tfrac{2}{3}\kappa^2 + O(\kappa^4), \tag{28}$$

and are strictly increasing onto $(0,1)$ and $(0,2)$, respectively. Hence $\beta < \eta$ and the scalar stationary root is unique. The sufficient elasticity condition in Theorem 4.4 is

$$\mathfrak{e}(\kappa^\star) := \frac{\kappa^\star}{2}\frac{\gamma'(\kappa^\star)}{\gamma(\kappa^\star)} + \frac{\kappa^\star\psi'(\kappa^\star)}{4}\frac{\eta\lambda_1}{1-\psi(\kappa^\star)\eta\lambda_1/2} \le 1. \tag{29}$$

This sufficient condition is not necessary: the exact local test is $\rho(D\mathcal{F}_{\mathrm{batch}}(\boldsymbol{W}^\star)) < 1$, whose diagonal block is a diagonal-plus-rank-one matrix. The $O(\eta\lambda_1)$ trace reduction is sharp in order; for $\boldsymbol{H} = \boldsymbol{I}_d$ the exact ungated relation is $\eta(d+2) = 2$.

Equation (21) is a stationarity condition, not a claim that every smaller step is stable or that training reaches the local basin. At fixed $\boldsymbol{H}$ and η the root fixes $\kappa^\star$; translating a smaller threshold C into a larger trained trace requires a sharpening theory not supplied here.

B.4 Deferred per-sample ZO operators

The exact linearized operators below require no drift-collinearity assumption. Such an assumption would be a full matrix identity, not merely alignment along one checkpoint displacement: the latter identifies only $\boldsymbol{K}_{\mathrm{dr}}\boldsymbol{\delta}$, whereas covariance diagonalization requires the entire operator. Batch clipping has $\boldsymbol{K}_{\mathrm{dr}} = p\boldsymbol{H}$ exactly. Our per-sample diagnostics show that the corresponding scalar-multiple approximation fails under active clipping, which is why Section 4.2 uses a scalar loss balance instead.

Theorem B.1 (Saturated map; per-sample)**.** *Under the centered sample-quadratic model, set* $s_C = n^{-1}\sum_i \mathrm{clip}_C(\boldsymbol{a}_i^\top\boldsymbol{u})$, $\boldsymbol{a}_i = \boldsymbol{H}_i\boldsymbol{\delta}_t$, *and* $\boldsymbol{W} = \boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top$. *Conditional on* $\boldsymbol{\delta}_t$,

$$\mathbb{E}[\boldsymbol{\delta}_{t+1}\boldsymbol{\delta}_{t+1}^\top \mid \boldsymbol{\delta}_t] = \boldsymbol{W} - \eta(\boldsymbol{K}\boldsymbol{W} + \boldsymbol{W}\boldsymbol{K}) + \eta^2\mathbb{E}[s_C^2\boldsymbol{u}\boldsymbol{u}^\top], \qquad \boldsymbol{K} = \frac{1}{n}\sum_i p_i\boldsymbol{H}_i, \tag{30}$$

and

$$\mathbb{E}[s_C^2\boldsymbol{u}\boldsymbol{u}^\top] = \frac{1}{n^2}\sum_{i,j}\{\boldsymbol{E}_{ij}\boldsymbol{M}_{ij}\boldsymbol{E}_{ij}^\top + \Theta_{ij}\boldsymbol{\Pi}_{ij}\}, \tag{31}$$

where $\boldsymbol{E}_{ij}, \boldsymbol{\Pi}_{ij}$ *are as in Theorem C.14 and*

$$\Theta_{ij} := \mathbb{E}\big[\mathrm{clip}_C(d_i)\,\mathrm{clip}_C(d_j)\big] = \sigma_i\sigma_j\int_0^{\rho_{ij}} c_{ij}(r)\,dr$$

is the clipped-score *pair moment, not the gate co-activation* $c_{ij} = \mathbb{E}[a_ia_j]$ *of Lemma C.6 that carries the same role in Theorem C.14; the two differ because a clipped score is not an indicator. Finally* $\boldsymbol{M}_{ij}$ *is the* 2×2 *plane moment* $\mathbb{E}[\mathrm{clip}_i\,\mathrm{clip}_j\,\boldsymbol{x}\boldsymbol{x}^\top]$. *Its entries are the bivariate slab quadratures of Proposition C.18.*

The conditional identity does not close on $\boldsymbol{W}_t = \mathbb{E}[\boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top]$, because its probabilities and plane geometry depend on $\boldsymbol{\delta}_t$. Freezing only those scalar moments would moreover leave an affine expression. Instead freeze the soft multipliers $c_i(\boldsymbol{u}) = \mathrm{clip}_C(\boldsymbol{a}_i^\top \boldsymbol{u})/(\boldsymbol{a}_i^\top \boldsymbol{u})$ and define

$$\mathcal{Q}_{\mathrm{sat}}(\boldsymbol{W}) = \frac{1}{n^2} \sum_{i,j} \mathbb{E}[c_i(\boldsymbol{u}) c_j(\boldsymbol{u}) (\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{W} \boldsymbol{H}_j \boldsymbol{u}) \boldsymbol{u}\boldsymbol{u}^\top].$$

Writing $\boldsymbol{v} = n^{-1} \sum_i c_i(\boldsymbol{u}) \boldsymbol{H}_i \boldsymbol{u}$ shows $\mathcal{Q}_{\mathrm{sat}}(\boldsymbol{W}) = \mathbb{E}[(\boldsymbol{v}^\top \boldsymbol{W} \boldsymbol{v}) \boldsymbol{u}\boldsymbol{u}^\top]$, hence positivity on the PSD cone and recovery of (31) for $\boldsymbol{W} = \boldsymbol{\delta}\boldsymbol{\delta}^\top$. The matching drift operator is

$$\boldsymbol{K}_c = \frac{1}{n} \sum_i \boldsymbol{\Sigma}_{c_i} \boldsymbol{H}_i, \qquad \boldsymbol{\Sigma}_{c_i} = \bar{c}_i (\boldsymbol{I} - \widehat{\boldsymbol{a}}_i \widehat{\boldsymbol{a}}_i^\top) + p_i \widehat{\boldsymbol{a}}_i \widehat{\boldsymbol{a}}_i^\top, \qquad \bar{c}_i = \mathbb{E}[c_i(\boldsymbol{u})].$$

It agrees with $\boldsymbol{K}$ in (30) along the checkpoint direction, but need not be symmetric. Thus the fixed-coefficient linear operator is

$$\mathcal{T}_\eta(\boldsymbol{W}) = \boldsymbol{W} - \eta(\boldsymbol{K}_c \boldsymbol{W} + \boldsymbol{W} \boldsymbol{K}_c^\top) + \eta^2 \mathcal{Q}_{\mathrm{sat}}(\boldsymbol{W}),$$

where the transpose ensures that symmetric matrices remain symmetric. This gives the quasi-static boundary $\rho(\mathcal{T}_\eta) = 1$. Monte Carlo with 8×10^6 draws reproduces the conditional formulas within 0.1% for $C = 0.3, 1, 3, 10$.

These formulas validate the linearized per-sample operator itself. They do not validate a reduction to $\gamma_{\mathrm{ps}} \boldsymbol{H}$ or attraction of the training trajectory to a nonzero covariance.

B.5 Per-sample matrix closure: the derivation chain

Assumption B.2 (Per-sample self-averaging)**.** *This is the per-sample analogue of Assumption 4.2. The batch closure replaces one random gradient amplitude by its covariance-level root-mean-square value; here we replace the random collection of sample scores by a Gaussian vector whose covariance is determined by $\boldsymbol{W}$. For a covariance $\boldsymbol{W} \succeq 0$, define the sample-score covariance*

$$\boldsymbol{G}(\boldsymbol{W})_{ij} := \mathrm{Tr}(\boldsymbol{H}_i \boldsymbol{H}_j \boldsymbol{W}), \qquad \boldsymbol{z}(\boldsymbol{W}) \sim \mathcal{N}(0, \boldsymbol{G}(\boldsymbol{W})).$$

We replace the displacement-dependent score moments by those of $\boldsymbol{z}(\boldsymbol{W})$. In particular,

$$p_i(\boldsymbol{W}) := \Pr(|z_i(\boldsymbol{W})| < C), \qquad M_2(\boldsymbol{W}) := \mathbb{E}\left[\left(\frac{1}{n} \sum_i \mathrm{clip}_C(z_i(\boldsymbol{W})) \right)^2 \right],$$

and, with $p_i(\boldsymbol{\delta}) := \Pr_{\boldsymbol{u}}(|\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{\delta}| < C)$, we approximate

$$\mathbb{E}\big[p_i(\boldsymbol{\delta}) \boldsymbol{\delta}\boldsymbol{\delta}^\top\big] \approx p_i(\boldsymbol{W}) \boldsymbol{W}, \qquad \boldsymbol{W} = \mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top].$$

The isotropic reduction used in Proposition 4.5 and Theorem 4.6 further sets $\boldsymbol{\Sigma}_s(\boldsymbol{W}) = M_2(\boldsymbol{W}) \boldsymbol{I}$. These replacements define the per-sample closed covariance map; they are not asserted as identities for the original displacement dynamics.

Proof of Proposition 4.5.

Proof. For this proof, write $\mathcal{G}(\boldsymbol{W}) = \eta M_2(\boldsymbol{W}) \boldsymbol{H}_p(\boldsymbol{W})^{-1}/2$. $\boldsymbol{H} \succ 0$ and the strict positivity of every finite gate probability imply $\boldsymbol{H}_p(\boldsymbol{W}) \succ 0$ for every finite $\boldsymbol{W}$, so the inverse is well defined. $\mathcal{G}$ is order preserving. If $\boldsymbol{W} \preceq \boldsymbol{W}'$ then every σ_i rises, every κ_i falls and every p_i falls, so $\boldsymbol{H}_p(\boldsymbol{W}) \succeq \boldsymbol{H}_p(\boldsymbol{W}')$ because the $\boldsymbol{H}_i$ are positive semidefinite; inversion is order reversing, and M_2 is non-decreasing by hypothesis, so $\mathcal{G}(\boldsymbol{W}) \preceq \mathcal{G}(\boldsymbol{W}')$.

A sub-solution exists. As $\boldsymbol{W} \to 0$ every gate opens, $\boldsymbol{H}_p \to \boldsymbol{H}$ and $M_2(\boldsymbol{W}) = \mathrm{Tr}(\boldsymbol{H}^2 \boldsymbol{W}) + o(\|\boldsymbol{W}\|)$. Hence

$$\epsilon^{-1} \boldsymbol{H}^{1/2} \mathcal{G}(\epsilon \boldsymbol{H}^{-1}) \boldsymbol{H}^{1/2} \longrightarrow \frac{\eta \,\mathrm{Tr}(\boldsymbol{H})}{2} \boldsymbol{I}.$$

Because $\eta \,\mathrm{Tr}(\boldsymbol{H})/2 > 1$, it follows that $\mathcal{G}(\epsilon \boldsymbol{H}^{-1}) \succeq \epsilon \boldsymbol{H}^{-1}$ for all sufficiently small $\epsilon > 0$.

A super-solution exists. Take $\boldsymbol{W}_0 = \boldsymbol{H}^{-1} \succ 0$. Along $\boldsymbol{W} = \lambda \boldsymbol{W}_0$ the amplitudes grow as $\sqrt{\lambda}$, so $p_i = O(\lambda^{-1/2})$ and $\boldsymbol{H}_p^{-1} = O(\lambda^{1/2})$, while $|s| \le C$ bounds M_2; hence $\mathcal{G}(\lambda \boldsymbol{W}_0) = O(\lambda^{1/2})$ and falls below $\lambda \boldsymbol{W}_0$ for λ large.

Iterating $\mathcal{G}$ from the sub-solution gives an increasing sequence, which by order preservation and induction stays below the super-solution. The Loewner order is not a lattice, but an increasing sequence bounded above still converges: each $\boldsymbol{x}^\top \boldsymbol{W}_k \boldsymbol{x}$ increases and is bounded, hence converges, and polarisation recovers a limit $\boldsymbol{W}^\star$. Continuity of $\mathcal{G}$ makes $\boldsymbol{W}^\star$ a fixed point, and it is nonzero because it dominates the sub-solution. □

Pair-gate moments and the scalar loss reading. Although the matrix balance does not reduce to an effective Hessian, two scalar summaries remain useful as diagnostics. Define

$$\gamma_{\mathrm{ps}}(\mathcal{W}) := \frac{\mathbb{E}_{\mathcal{W}}[s_t^2]}{\mathbb{E}_{\mathcal{W}}[s_t d_t]},$$

so that, with $q_t = \boldsymbol{u}_t^\top \boldsymbol{H}_t \boldsymbol{u}_t$,

$$\tau_{\mathrm{ps}}^{\mathrm{LB}} = \gamma_{\mathrm{ps}} \frac{\mathbb{E}_{\mathcal{W}}[s_t^2 q_t]}{\mathbb{E}_{\mathcal{W}}[s_t^2]}.$$

This is a squared-update-per-descent ratio followed by an update-weighted curvature; it is not the matrix $\gamma_{\mathrm{ps}} \boldsymbol{H}$.

At a fixed centered sample-quadratic state, write $\boldsymbol{a}_i = \boldsymbol{H}_i \boldsymbol{\delta}$, $\boldsymbol{a} = n^{-1} \sum_i \boldsymbol{a}_i$, $\sigma_i = \|\boldsymbol{a}_i\|$, $\kappa_i = C/\sigma_i$, and $\rho_{ij} = \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle / (\sigma_i \sigma_j)$.

Corollary B.3 (Pair-moment formula for γ_{ps})**.** *With $c_{ij}(r) = \Pr(|x_1| < \kappa_i, |x_2| < \kappa_j)$ the co-activation probability at correlation r,*

$$\mathbb{E}[s_C d] = \frac{1}{n} \sum_i p_i \langle \boldsymbol{a}_i, \boldsymbol{a} \rangle, \qquad \mathbb{E}[s_C^2] = \frac{1}{n^2} \sum_{i,j} \sigma_i \sigma_j \int_0^{\rho_{ij}} c_{ij}(r)\, dr = \frac{1}{n^2} \sum_{i,j} \Theta_{ij}.$$

Their ratio is γ_{ps}. The off-diagonal terms retain sample-gradient alignment and pair-gate co-activation, which is precisely what prevents a single effective-curvature reduction.

Proposition B.4 (One-step loss balance; any scalar ZO rule)**.** *For $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta s_t \boldsymbol{u}_t$, Taylor expansion gives*

$$L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) = -\eta s_t d_t + \frac{\eta^2}{2} s_t^2 q_t + R_t,$$

where R_t is the Taylor and finite-difference remainder. For a window $\mathcal{W}$, define

$$D_{\mathcal{W}} := \mathbb{E}_{\mathcal{W}}[s_t d_t], \qquad N_{\mathcal{W}} := \mathbb{E}_{\mathcal{W}}[s_t^2 q_t], \qquad \tau_{\mathrm{ps}}^{\mathrm{LB}}(\mathcal{W}) := \frac{N_{\mathcal{W}}}{D_{\mathcal{W}}}.$$

Whenever $D_{\mathcal{W}} > 0$,

$$\frac{\eta \tau_{\mathrm{ps}}^{\mathrm{LB}}(\mathcal{W})}{2} = 1 + \frac{\mathbb{E}_{\mathcal{W}}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t)]}{\eta D_{\mathcal{W}}} - \frac{\mathbb{E}_{\mathcal{W}}[R_t]}{\eta D_{\mathcal{W}}}.$$

Consequently, a quadratic objective under a stationary trajectory law obeys

$$\eta \tau_{\mathrm{ps}}^{\mathrm{LB}} = 2.$$

This identity is exact with its two correction terms retained. It is a secondary scalar reading of the matrix balance: a descending finite trajectory is expected to approach the stationary ratio from below when the remainder is small.

At a fixed state, put $\boldsymbol{\Sigma}_s = \mathbb{E}[s^2 \boldsymbol{u}\boldsymbol{u}^\top]$. The associated positive semidefinite diagnostic is

$$\mathcal{A}_{\mathrm{ps}} := \frac{\boldsymbol{H}^{1/2} \boldsymbol{\Sigma}_s \boldsymbol{H}^{1/2}}{\mathbb{E}[sd]}, \qquad \mathrm{Tr}(\mathcal{A}_{\mathrm{ps}}) = \tau_{\mathrm{ps}}^{\mathrm{LB}}. \tag{32}$$

This is a loss-cost object, not an effective drift matrix.

Our synthetic sweeps further support the exploratory condition

$$\eta \lambda_{\max}(\boldsymbol{H}_p) < 1$$

as sufficient for local stability of the closed map. We leave it as a conjecture: the exact test is the gate-space criterion developed below, and the available sweeps do not exclude a failure outside their range.

The trace approximation to the per-sample loss balance.

Corollary B.5 (Trace approximation to the per-sample loss balance)**.** *Define the trajectory-weighted trace proxy and its directional-correlation factor by*

$$\widetilde{\tau}_{\mathrm{ps}}^{\mathrm{LB}}(\mathcal{W}) := \frac{\mathbb{E}_{\mathcal{W}}[s_t^2 \operatorname{Tr}(\boldsymbol{H}_t)]}{\mathbb{E}_{\mathcal{W}}[s_t d_t]}, \qquad r_{\mathrm{curv}}(\mathcal{W}) := \frac{\mathbb{E}_{\mathcal{W}}[s_t^2 q_t]}{\mathbb{E}_{\mathcal{W}}[s_t^2 \operatorname{Tr}(\boldsymbol{H}_t)]}. \tag{33}$$

Then, algebraically,

$$\tau_{\mathrm{ps}}^{\mathrm{LB}} = r_{\mathrm{curv}} \widetilde{\tau}_{\mathrm{ps}}^{\mathrm{LB}}.$$

Thus $r_{\mathrm{curv}} \approx 1$ justifies replacing the directional curvature q_t by $\operatorname{Tr}(\boldsymbol{H}_t)$ in this scalar balance. Only when the trace is also approximately constant over $\mathcal{W}$, or sufficiently decorrelated from s_t^2, may this be reduced further to $\widetilde{\tau}_{\mathrm{ps}}^{\mathrm{LB}} \approx \gamma_{\mathrm{ps}} \overline{\operatorname{Tr}(\boldsymbol{H})}$.

The distinction in Corollary B.5 is material. Our fixed-checkpoint probes give $r_{\mathrm{curv}} = 1.001$–$1.032$, so directional curvature correlation is small in the measured runs. Across a moving trajectory, however, separating the s_t^2 and trace averages changes the post-peak scalar by as much as 12%. The weighted proxy in Eq. (33), rather than the matrix $\gamma_{\mathrm{ps}}\boldsymbol{H}$, is therefore the quantity used in Section 5.

Theorem 4.6 proves local attraction under its cone and Perron-test conditions. Here we record the diagnostics and the route toward a stronger global statement. On the centered sample-quadratic model, where every coefficient is computable, the exact local test $\rho(D\mathcal{F}_{\mathrm{ps}}(\boldsymbol{W}^\star)) < 1$ of Theorem 4.4 holds at all 28 settings we examined, spanning $n = 1$ to 16, gate rates $\bar{p} = 0.11$ to 0.98, and curvature heterogeneity over a decade; the induced Lyapunov map is order-preserving in every trial, and iterating it from covariances spread over three orders of magnitude returns the same fixed point. Order preservation together with subhomogeneity would give uniqueness and global attraction by nonlinear Perron–Frobenius arguments; subhomogeneity, however, is violated by up to 1% once the gate rates are both high and unequal across samples. The mechanism is anisotropy of the gate response, not concavity of the diffusion. Writing the stationary map as $\mathcal{G}(\boldsymbol{W}) = \frac{\eta}{2} M_2(\boldsymbol{W}) \boldsymbol{H}_p(\boldsymbol{W})^{-1}$, each factor is separately subhomogeneous in every instance we tested; what fails is their product, because $\boldsymbol{H}_p$ changes shape and not merely scale with the amplitude. A sufficient condition is therefore

$$\frac{M_2(\lambda \boldsymbol{W})}{\lambda M_2(\boldsymbol{W})} \;\geq\; \lambda_{\max}\big(\boldsymbol{H}_p(\boldsymbol{W})^{-1} \boldsymbol{H}_p(\lambda \boldsymbol{W})\big), \qquad \lambda \in (0,1), \tag{34}$$

the scalar diffusion gain dominating the stiffest direction of the gate response. At the violating instances the two sides are 1.403 and 1.412.

Two reductions make (34) checkable. First, $\boldsymbol{H}_p(\boldsymbol{W})$ and $\boldsymbol{H}_p(\lambda \boldsymbol{W})$ are positive combinations of the same $\boldsymbol{H}_i$, and $\lambda \boldsymbol{W} \preceq \boldsymbol{W}$ opens every gate, so with $r = \max_i p_i(\lambda \boldsymbol{W})/p_i(\boldsymbol{W})$ one has $r\boldsymbol{H}_p(\boldsymbol{W}) - \boldsymbol{H}_p(\lambda \boldsymbol{W}) = n^{-1} \sum_i \big(r p_i - p_i(\lambda \boldsymbol{W})\big) \boldsymbol{H}_i \succeq 0$ whenever $\boldsymbol{H}_i \succeq 0$, whence the right-hand side of (34) is at most r. Second, $\boldsymbol{\Sigma}(\boldsymbol{W})_{ij} = \operatorname{Tr}(\boldsymbol{H}_i \boldsymbol{H}_j \boldsymbol{W})$ is *linear* in $\boldsymbol{W}$, so the scores scale as $\sqrt{\lambda}$ and $\operatorname{clip}_C(\sqrt{\lambda}\, d) = \sqrt{\lambda}\, \operatorname{clip}_{C/\sqrt{\lambda}}(d)$; writing $\Psi(c) := \mathbb{E}[(n^{-1} \sum_i \operatorname{clip}_c(d_i))^2]$ at fixed $\boldsymbol{W}$ and $t = 1/\sqrt{\lambda}$, the left-hand side is exactly $\Psi(tC)/\Psi(C)$. Both sides equal one at $t = 1$, and differentiating there gives the local criterion

$$\varepsilon_\Psi \;\geq\; \lambda_{\max}\Big(\boldsymbol{H}_p^{-1} \frac{1}{n} \sum_i p_i \varepsilon_i \boldsymbol{H}_i\Big), \qquad \varepsilon_\Psi := \frac{d \log M_2}{d \log C}, \qquad \varepsilon_i := \frac{2\kappa_i \phi(\kappa_i)}{p(\kappa_i)}. \tag{35}$$

Every quantity in (35) is measurable, and the right-hand side is where the condition is directional: each sample's gate elasticity ε_i is weighted by the direction its curvature occupies, not by its size alone. Across the eight closed-model settings we solved, (35) agrees with the sign of the realised violation in all eight, while the cruder bound $\max_i \varepsilon_i$ misclassifies two of them.

Because $\varepsilon_i < 1$ for every $\kappa_i > 0$, with supremum one as $\kappa_i \to 0$, (35) is implied by the single scalar condition

$$\varepsilon_\Psi \;\geq\; 1, \qquad \text{equivalently } C \mapsto M_2(C)/C \text{ non-decreasing}, \tag{36}$$

which involves no per-sample quantity at all. This is the precise form of "clipping is active": $\varepsilon_\Psi \to 2$ when every sample is strongly clipped and $\varepsilon_\Psi \to 0$ when none is, so the diffusion must still respond at least linearly to the threshold. For one shared gate the criterion reduces to $\gamma = \mu_2/p$ non-decreasing in κ, that is $\kappa p(1-p) \geq \phi(\kappa)\mu_2(\kappa)$, so batch clipping satisfies (34) unconditionally; this is the precise sense in which the per-sample problem is harder, and it replaces the batch elasticity scalar by a directional condition.

Uniqueness, after a change of coordinates. Subhomogeneity of $\mathcal{G}$ in $\boldsymbol{W}$ is false, which is what closes off the usual route to uniqueness. The obstruction disappears in gate coordinates. A fixed point has $\boldsymbol{W} = s\,\boldsymbol{A}(\boldsymbol{p})^{-1}$ with $\boldsymbol{A}(\boldsymbol{p}) = n^{-1}\sum_i p_i \boldsymbol{H}_i$ and $s = \eta M_2/2$, so $\sigma_i^2 = s\,t_i(\boldsymbol{p})$ with $t_i(\boldsymbol{p}) := \mathrm{Tr}(\boldsymbol{H}_i^2 \boldsymbol{A}(\boldsymbol{p})^{-1})$, and consistency closes on

$$\Phi_s(\boldsymbol{p})_i := p\Big(C/\sqrt{s\,t_i(\boldsymbol{p})}\Big), \qquad \boldsymbol{p} \in (0,1]^n.$$

Proposition B.6 (At most one gate vector)**.** *Φ_s is order preserving and satisfies $\Phi_s(\lambda \boldsymbol{p}) \succeq \sqrt{\lambda}\,\Phi_s(\boldsymbol{p})$ for $\lambda \in (0,1)$. Hence it has at most one fixed point in $(0,1]^n$, and the nonzero fixed points of $\mathcal{G}$ correspond to the roots of the scalar equation $\eta\tau_{\mathrm{ps}} = 2$.*

Proof. Raising $\boldsymbol{p}$ raises $\boldsymbol{A}$, lowers $\boldsymbol{A}^{-1}$ and hence each t_i, raises κ_i, and raises $p(\kappa_i)$, so Φ_s is order preserving. Because $\boldsymbol{A}$ is *linear* in $\boldsymbol{p}$, $\boldsymbol{A}(\lambda\boldsymbol{p})^{-1} = \lambda^{-1}\boldsymbol{A}(\boldsymbol{p})^{-1}$ and $t_i(\lambda\boldsymbol{p}) = t_i(\boldsymbol{p})/\lambda$, so κ_i scales by exactly $\sqrt{\lambda}$. Now $p(\kappa) = \mathrm{erf}(\kappa/\sqrt{2})$ is concave on $\kappa \geq 0$ with $p(0) = 0$, so $p(a\kappa) \geq a\,p(\kappa)$ for $a \in [0,1]$; taking $a = \sqrt{\lambda}$ gives the stated bound, and $\sqrt{\lambda} > \lambda$ makes it strict. An order-preserving, strictly subhomogeneous self-map of the interior of a cone has at most one fixed point there. □

Uniqueness of the nonzero stationary covariance therefore reduces from an n-dimensional question to a scalar one, exactly as Theorem 4.3 does for batch clipping, where the corresponding scalar $F_{\mathrm{batch}}(\kappa)$ is shown to be monotone. We verify every step: order preservation and the $\sqrt{\lambda}$ bound hold at all 27 settings we tested, with slack 10^{-8} to 4×10^{-3}; iterating Φ_s from a dozen random gate vectors returns the same fixed point to 10^{-14}; and sweeping s over six decades the equation $\eta\tau_{\mathrm{ps}} = 2$ crosses zero exactly once in every case. Note that the origin remains a fixed point throughout, so the claim is uniqueness of the *nonzero* state; started small enough, the iteration still collapses to zero.

Monotonicity of that scalar is the one step we do not prove, but its endpoints are forced and its mechanism is identifiable. As $s \to 0$ every gate opens, $\boldsymbol{A} \to \boldsymbol{H}$ and $\bar{\gamma}_{ij} \to 1$, so $\tau_{\mathrm{ps}} \to n^{-2}\sum_{ij}\mathrm{Tr}(\boldsymbol{H}_i\boldsymbol{H}^{-1}\boldsymbol{H}_j) = \mathrm{Tr}(\boldsymbol{H})$; as $s \to \infty$, $p_i \simeq \sqrt{2/\pi}\,C/\sqrt{s\,t_i}$ makes $\boldsymbol{A} = O(s^{-1/2})$ and the traces $O(s^{1/2})$ while $\bar{\gamma}_{ij} = O(s^{-1})$, so $\tau_{\mathrm{ps}} = O(s^{-1/2}) \to 0$. A monotone decrease between those endpoints therefore gives a unique root precisely when $\eta\,\mathrm{Tr}(\boldsymbol{H}) > 2$, which is the existence condition of Proposition 4.5: the two statements meet at the same threshold.

The mechanism separates cleanly. Writing $\tau_{\mathrm{ps}}(s) = M_2(s\boldsymbol{A}(s)^{-1})/s$, freezing the direction $\boldsymbol{A}$ makes the decrease exactly the statement $\varepsilon_\Psi \geq 0$, that $\Psi(c) = \mathbb{E}[(n^{-1}\sum_i \mathrm{clip}_c(d_i))^2]$ increases in c; what is left is the rotation of $\boldsymbol{A}$ as the gates close unevenly. Numerically the fixed-direction term is negative at every setting and always dominates, while the rotation term is positive but falls from a comparable size when the gates are nearly open to between a fiftieth and a thousandth of it once clipping bites. So uniqueness rests on $\varepsilon_\Psi \geq 0$ together with control of that rotation, just as existence rests on $\varepsilon_\Psi \leq 2$.

The pair is not a coincidence, and it collapses to a single quantity. Put $v_i := \mathrm{clip}_c(d_i)$; then $\partial v_i/\partial c = \mathbf{1}\{|d_i| > c\}\,\mathrm{sign}(d_i)$, which vanishes off the saturated event and equals v_i/c on it. Splitting the update accordingly,

$$u_i := c\,\mathrm{sign}(d_i)\mathbf{1}\{|d_i| > c\}, \quad w_i := d_i\mathbf{1}\{|d_i| < c\}, \quad v = w + u, \quad \partial v/\partial c = u/c, \tag{37}$$

and writing $\bar{u}, \bar{w}$ for the sample means of u, w, so that $s_c = \bar{u} + \bar{w}$, the two bounds read

$$\varepsilon_\Psi \geq 0 \iff \mathbb{E}[s_c\bar{u}] \geq 0, \qquad \varepsilon_\Psi \leq 2 \iff \mathbb{E}[s_c\bar{w}] \geq 0,$$

whose sum is $\mathbb{E}[s_c^2] \geq 0$ and hence automatic: the update must correlate non-negatively with each of its saturated and unsaturated parts separately. Expanding, $\mathbb{E}[s_c\bar{u}] = \mathbb{E}[\bar{u}^2] + \mathbb{E}[\bar{u}\bar{w}]$ and $\mathbb{E}[s_c\bar{w}] = \mathbb{E}[\bar{w}^2] + \mathbb{E}[\bar{u}\bar{w}]$, so $0 \leq \varepsilon_\Psi \leq 2$ is the single requirement $\mathbb{E}[\bar{u}\bar{w}] \geq -\min(\mathbb{E}[\bar{u}^2], \mathbb{E}[\bar{w}^2])$.

That cross-covariance is unusually well behaved. Since u_i and w_i have disjoint support, $\mathbb{E}[u_i w_i] = 0$ for every i, so $\mathbb{E}[\bar{u}\bar{w}]$ has no diagonal at all: it is a purely cross-sample, $O(\rho_{ij})$ effect, and it vanishes identically when a single gate is shared, which is exactly why both bounds are unconditional in the batch case. On realizable covariances we measure it not merely bounded but *positive* at every instance, from 1.4×10^{-3} to 0.47. That sign is not an accident of the measurement: it is forced, entry by entry, by the sign of the correlation.

Proposition B.7 (Cross-covariance of the saturated and unsaturated parts)**.** *Write $d_i = \sigma_i z_i$ with (z_i, z_j) standard bivariate normal of correlation ρ_{ij} and $\kappa_i := c/\sigma_i$. Then for every pair $i \neq j$ and every pair of thresholds,*

$$\operatorname{sign} \mathbb{E}[u_i w_j] = \operatorname{sign} \rho_{ij}, \tag{38}$$

with $\mathbb{E}[u_i w_j] = 0$ exactly when $\rho_{ij} = 0$. Consequently $\mathbb{E}[\bar{u}\bar{w}] \geq 0$ whenever $\rho_{ij} \geq 0$ for all $i \neq j$, and then $0 \leq \varepsilon_\Psi \leq 2$ holds unconditionally.

Proof. Since $\mathbb{E}[u_i w_j] = c\sigma_j \mathbb{E}[f(z_i)g(z_j)]$ with $f(x) = \operatorname{sign}(x)\mathbf{1}\{|x| > \kappa_i\}$ and $g(y) = y\mathbf{1}\{|y| < \kappa_j\}$, condition on z_i and set

$$\vartheta(x) := \mathbb{E}[g(z_j) \mid z_i = x] = \mathbb{E}[Y\mathbf{1}\{|Y| < \kappa_j\}], \qquad Y \sim \mathcal{N}(\bar{y},\, 1 - \rho_{ij}^2), \quad \bar{y} = \rho_{ij} x,$$

so that $\mathbb{E}[f(z_i)g(z_j)] = \mathbb{E}[f(z_i)\vartheta(z_i)]$. We claim that $\mathbb{E}[Y\mathbf{1}\{|Y| < \kappa\}] \geq 0$ for every $\kappa > 0$ whenever the mean $\bar{y}$ of Y is non-negative. Indeed, with f_τ the centred normal density of standard deviation τ, folding the interval about the origin gives

$$\mathbb{E}[Y\mathbf{1}\{|Y| < \kappa\}] = \int_0^\kappa y\,[f_\tau(y - \bar{y}) - f_\tau(y + \bar{y})]\,\mathrm{d}y,$$

and for $y, \bar{y} \geq 0$ one has $|y - \bar{y}| \leq y + \bar{y} = |y + \bar{y}|$, while f_τ is decreasing in $|\cdot|$; the bracket is therefore non-negative pointwise and so is the integral. This is the whole content of the claim: a symmetric unimodal density shifted to the right puts more of its mass on the positive half of any symmetric window than on the negative half.

Take $\rho_{ij} \geq 0$. For $x > 0$ the conditional mean $\rho_{ij}x$ is non-negative, so $\vartheta(x) \geq 0$ by the claim; ϑ is odd, hence $\vartheta(x) \leq 0$ for $x < 0$. Since $f(x)$ carries the sign of x, the product $f(z_i)\vartheta(z_i)$ is non-negative *pointwise*, and taking expectations gives $\mathbb{E}[u_i w_j] \geq 0$. Replacing ρ_{ij} by $-\rho_{ij}$ flips the sign of ϑ and hence of the product, which gives the reverse inequality and so (38); at $\rho_{ij} = 0$ the two factors are independent and both odd, so the expectation vanishes. The statement about $\mathbb{E}[\bar{u}\bar{w}]$ follows by summing over $i \neq j$, the diagonal being zero, and the elasticity bounds then follow from $\mathbb{E}[s_c\bar{u}] = \mathbb{E}[\bar{u}^2] + \mathbb{E}[\bar{u}\bar{w}]$ and $\mathbb{E}[s_c\bar{w}] = \mathbb{E}[\bar{w}^2] + \mathbb{E}[\bar{u}\bar{w}]$. □

Two remarks on why the proof has to be run this way. The natural first attempt, Price's theorem, does *not* close: differentiating in ρ and using Stein's identity, the μ-dependent terms cancel and one is left with

$$\frac{\partial}{\partial \rho_{ij}} \mathbb{E}[u_i w_j] \;\propto\; \int_a^b \phi(z)\,\mathrm{d}z - \frac{b-a}{2}\,[\phi(a) + \phi(b)], \qquad a, b = \frac{\mp\kappa_j - \rho_{ij}\kappa_i}{\sqrt{1 - \rho_{ij}^2}},$$

which asks whether the trapezoid rule underestimates the Gaussian integral on $[a, b]$. It does where ϕ is concave, but not on intervals lying in the convex tail — on $[2, 3]$ the integral is 0.0214 against a trapezoid of 0.0292 — and it is negative on 28% of a $(\kappa_i, \kappa_j, \rho)$ grid. So $\mathbb{E}[u_i w_j]$ is genuinely non-monotone in ρ_{ij}; only its sign is pinned, and the conditioning argument pins it directly. Second, the hypothesis $\rho_{ij} \geq 0$ is a restriction, but a mild one here: the covariance is the Gram matrix $\boldsymbol{\Sigma}(\boldsymbol{W})_{ij} = \operatorname{Tr}(\boldsymbol{H}_i \boldsymbol{H}_j \boldsymbol{W}) = \langle \boldsymbol{H}_i \boldsymbol{W}^{1/2}, \boldsymbol{H}_j \boldsymbol{W}^{1/2} \rangle_F$, whose entries are measured non-negative for 99% of pairs, and the sum tolerates far more than that — numerically $\mathbb{E}[\bar{u}\bar{w}]$ stays positive with a third of the off-diagonal entries negative. Since $\mathbb{E}[\bar{u}\bar{w}] \geq -\min(\mathbb{E}[\bar{u}^2], \mathbb{E}[\bar{w}^2])$ is all the elasticity bounds require, and $\mathbb{E}[\bar{u}^2], \mathbb{E}[\bar{w}^2]$ are bounded away from zero whenever the gate is neither fully open nor fully closed, the conclusion survives the violating minority with room to spare.

The local test reduces to n dimensions. The cone properties above belong to the global route. Local exponential stability needs only the linearisation, and there the per-sample problem is much smaller than it looks. Perturbing $\boldsymbol{W}^\star$ by $\varepsilon \boldsymbol{V}$, the coefficients move only through the n amplitudes $t_i(\boldsymbol{V}) = \langle \boldsymbol{H}_i^2, \boldsymbol{V} \rangle$ and through the scalar M_2, so with $\boldsymbol{B}_i := \boldsymbol{H}_i \boldsymbol{W}^\star + \boldsymbol{W}^\star \boldsymbol{H}_i$ and $\mathcal{L}_0(\boldsymbol{V}) := \boldsymbol{V} - \eta(\boldsymbol{H}_p \boldsymbol{V} + \boldsymbol{V} \boldsymbol{H}_p)$,

$$D\mathcal{F}_{\mathrm{ps}}(\boldsymbol{V}) = \mathcal{L}_0(\boldsymbol{V}) + \sum_{i=1}^n \varpi_i(\boldsymbol{V})\, \boldsymbol{B}_i, \qquad \varpi_i(\boldsymbol{X}) = \frac{\eta\, p_i \varepsilon_i}{2n\sigma_i^2} \langle \boldsymbol{H}_i^2, \boldsymbol{X} \rangle + \frac{\eta\, p_i}{M_2 n}\, dM_2(\boldsymbol{X}),$$

a correction of rank exactly n rather than $n+1$: the fixed-point equation $\boldsymbol{H}_p \boldsymbol{W}^\star + \boldsymbol{W}^\star \boldsymbol{H}_p = \eta M_2 \boldsymbol{I}$ puts $\boldsymbol{I}$ itself in $\operatorname{span}\{\boldsymbol{B}_i\}$, so the diffusion feedback contributes no new direction. We confirm the

rank numerically at $n = 2, 4, 8$ in a 36 dimensional symmetric space. Since $\mathcal{L}_0$ is diagonal in the eigenbasis of $\boldsymbol{H}_p$ with eigenvalues $1 - \eta(\nu_a + \nu_b)$, its resolvent $\mathcal{R}_z = (\mathcal{L}_0 - z)^{-1}$ is explicit and the eigenproblem collapses to the secular equation

$$\det\big(\boldsymbol{I}_n + \boldsymbol{M}(z)\big) = 0, \qquad \boldsymbol{M}(z)_{ij} = \varpi_i\big(\mathcal{R}_z(\boldsymbol{B}_j)\big), \tag{39}$$

an $n \times n$ determinant in place of a $d(d+1)/2$ dimensional eigenproblem. It reproduces $\rho(D\mathcal{F}_{\mathrm{ps}}(\boldsymbol{W}^\star))$ to 10^{-11} at every setting we solved. For one shared gate it is scalar, which is the diagonal-plus-rank-one structure Appendix B.3 already uses for the batch map.

Because $\det(\boldsymbol{I}_n + \boldsymbol{M}(z))$ is analytic outside the unit disc — the poles of $\boldsymbol{M}$ are the eigenvalues of $\mathcal{L}_0$, all inside it — and tends to one at infinity, the argument principle counts the secular roots that would violate stability,

$$\#\{z : \det(\boldsymbol{I}_n + \boldsymbol{M}(z)) = 0,\ |z| > 1\} = -\frac{1}{2\pi}\big[\arg\det(\boldsymbol{I}_n + \boldsymbol{M}(z))\big]_{|z|=1}, \tag{40}$$

so a winding number on the unit circle decides local stability exactly. That count is zero at every setting we examined, with $\min_{|z|=1} |\det(\boldsymbol{I}_n + \boldsymbol{M})|$ between 0.18 and 0.32, so the verdict is not numerically marginal. It is also zero at the checkpoints: assembling (39) from the measured per-sample Hessians gives no root outside the unit circle at any of the nine clipped settings for which a nonzero fixed point exists, spanning gate rates 0.32–0.92 and a tenfold range of C, with $\min_{|z|=1} |\det|$ between 0.098 and 0.163. The remaining settings are reported as subcritical rather than stable, the plane carrying too little of $\mathrm{Tr}\,\boldsymbol{H}$ for a nonzero fixed point; that is also why the checkpoint evaluation scales η, so it tests the map built from measured curvature and not those runs' trajectories. Note that $\rho(\mathcal{L}_0) = 1 - 2\eta\nu_{\min}$ is 0.9999 at these checkpoints, close enough to one that reading stability off a directly computed spectral radius would be inconclusive, whereas the winding number is unaffected by that proximity. We report (40) rather than a closed-form condition deliberately: replacing it by norm bounds is very lossy, a row-sum bound on $\boldsymbol{M}$ giving 5–18 where the truth is below one, and an entrywise contour bound giving 0.64–1.78 and certifying four of six settings. The per-sample analogue of the batch elasticity condition is therefore a computable n-dimensional test, not an inequality in closed form.

The criterion is a cone test, and it proves contraction. Condition (35) was obtained as a derivative comparison, but it has a second reading that supplies a proof. Order preservation of $\mathcal{G}$ makes $D\mathcal{G}(\boldsymbol{W}^\star)$ a positive operator on the semidefinite cone, and for such operators the Perron test applies: if some $\boldsymbol{V} \succ 0$ satisfies $D\mathcal{G}[\boldsymbol{V}] \preceq \theta \boldsymbol{V}$ then $\rho(D\mathcal{G}) \le \theta$. The fixed point supplies its own test vector. Since $\frac{d}{dt}\big[\mathcal{G}(t\boldsymbol{W}^\star)/t\big]_{t=1} = D\mathcal{G}[\boldsymbol{W}^\star] - \boldsymbol{W}^\star$, the choice $\boldsymbol{V} = \boldsymbol{W}^\star$, $\theta = 1$ asks for $\mathcal{G}$ to be subhomogeneous *infinitesimally at the fixed point* — not the global statement that (71) refutes.

Proposition B.8 (Contraction of the stationary map)**.** *Let $\boldsymbol{W}^\star \succ 0$ be a fixed point of $\mathcal{G}$ and suppose $\mathcal{G}$ is order preserving near it. Then*

$$D\mathcal{G}(\boldsymbol{W}^\star)[\boldsymbol{W}^\star] \preceq \boldsymbol{W}^\star \iff \lambda_{\max}\Big(\boldsymbol{H}_p^{-1}\,\frac{1}{n}\sum_i p_i \varepsilon_i \boldsymbol{H}_i\Big) \le \varepsilon_\Psi, \tag{41}$$

and either side implies $\rho(D\mathcal{G}(\boldsymbol{W}^\star)) \le 1$.

Proof. Write $\mathcal{G}(\boldsymbol{W}) = \frac{\eta}{2} M_2(\boldsymbol{W}) \boldsymbol{H}_p(\boldsymbol{W})^{-1}$ and differentiate along $t\boldsymbol{W}^\star$. On that ray σ_i^2 scales by t, so κ_i scales by $t^{-1/2}$ and $\dot{p}_i = -\frac{1}{2} p_i \varepsilon_i$, giving $\dot{\boldsymbol{H}}_p = -\frac{1}{2n}\sum_i p_i \varepsilon_i \boldsymbol{H}_i$; and $M_2(t\boldsymbol{W}^\star) = t\Psi(C/\sqrt{t})$ gives $\dot{M}_2/M_2 = 1 - \varepsilon_\Psi/2$. Hence $D\mathcal{G}[\boldsymbol{W}^\star] - \boldsymbol{W}^\star = \frac{\eta M_2}{2}\big[(\dot{M}_2/M_2 - 1)\boldsymbol{H}_p^{-1} - \boldsymbol{H}_p^{-1}\dot{\boldsymbol{H}}_p\boldsymbol{H}_p^{-1}\big]$. Conjugating by $\boldsymbol{H}_p^{1/2}$ turns $D\mathcal{G}[\boldsymbol{W}^\star] \preceq \boldsymbol{W}^\star$ into $-\frac{\varepsilon_\Psi}{2}\boldsymbol{I} \preceq \boldsymbol{H}_p^{-1/2}\dot{\boldsymbol{H}}_p\boldsymbol{H}_p^{-1/2}$, which is the stated eigenvalue inequality. Positivity of $D\mathcal{G}$ and the Perron test then give $\rho \le 1$. □

So (35) is not an ad hoc inequality: it is the cone test, and it delivers contraction with no norm bound anywhere — which is why it succeeds where the row-sum and contour estimates of (39) do not. We verify all of it: the Perron inequality holds at every setting we solved, with $\boldsymbol{W}^\star - D\mathcal{G}[\boldsymbol{W}^\star]$ bounded below by 0.032 to 0.186 times $\lambda_{\max}(\boldsymbol{W}^\star)$; the two sides of (41) agree in verdict throughout; and the realised $\rho(D\mathcal{G})$ is 0.53–0.74, so the bound is valid but not tight.

One step still separates this from the recursion. At the fixed point $\mathcal{F}$ and $\mathcal{G}$ are linked exactly by $\mathcal{F}(\boldsymbol{W}) = \boldsymbol{W} - \eta\,\mathcal{L}_{\boldsymbol{H}_p}(\boldsymbol{W} - \mathcal{G}(\boldsymbol{W}))$ with $\mathcal{L}_{\boldsymbol{A}}(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{A}$, so

$$D\mathcal{F} = \boldsymbol{I} - \eta\,\mathcal{L}_{\boldsymbol{H}_p}\big(\boldsymbol{I} - D\mathcal{G}\big),$$

which we confirm to 10^{-10}. Since $\mathcal{L}_{\boldsymbol{H}_p} \succ 0$ and $\rho(D\mathcal{G}) < 1$ places the spectrum of $\boldsymbol{I} - D\mathcal{G}$ in the right half plane, $\rho(D\mathcal{F}) < 1$ follows once that factor is accretive; we observe $\rho(D\mathcal{F}) = 0.76$–$0.87$ but do not prove accretivity.

Carrying the cone test to the recursion. Proposition B.8 contracts $\mathcal{G}$, but the dynamics is $\mathcal{F}$. The bridge is not accretivity of $\boldsymbol{I} - D\mathcal{G}$: that would suffice, since $\mathcal{L}_{\boldsymbol{H}_p} \succ 0$ and an accretive factor puts the spectrum of the product in the right half plane, but we measure $\lambda_{\max}$ of the symmetric part of $D\mathcal{G}$ at 1.08–1.60 once the curvature is heterogeneous, so it is simply false there, and $D\mathcal{G}$ is self-adjoint in none of the natural metrics we tried. The cone serves better a second time.

The test vector is supplied by the same fixed point, but it has to be earned. From $\mathcal{F}(\boldsymbol{W}) = \boldsymbol{W} - \eta\mathcal{L}_{\boldsymbol{H}_p}(\boldsymbol{W} - \mathcal{G}(\boldsymbol{W}))$ one gets $D\mathcal{F}[\boldsymbol{W}^\star] = \boldsymbol{W}^\star - \eta\mathcal{L}_{\boldsymbol{H}_p}(\boldsymbol{W}^\star - D\mathcal{G}[\boldsymbol{W}^\star])$, and although Proposition B.8 makes the bracket positive semidefinite, that alone is not enough: $\mathcal{L}_{\boldsymbol{A}}$ does *not* preserve the semidefinite cone, as $\boldsymbol{A} = \mathrm{diag}(1, 0)$ acting on the all-ones matrix already shows, returning eigenvalues -0.414 and 2.414.

Evaluating $\mathcal{L}_{\boldsymbol{H}_p}$ explicitly repairs this. With the gate-elasticity-weighted curvature $\boldsymbol{K}_\varepsilon := n^{-1}\sum_i p_i\varepsilon_i\boldsymbol{H}_i \succeq 0$, which is the matrix already appearing on the right of (35) and is not the drift $\boldsymbol{K}$ of (30), the proof of Proposition B.8 gives $\boldsymbol{W}^\star - D\mathcal{G}[\boldsymbol{W}^\star] = \frac{\eta M_2}{4}\boldsymbol{H}_p^{-1}(\varepsilon_\Psi\boldsymbol{I} - \boldsymbol{K}_\varepsilon\boldsymbol{H}_p^{-1})$, on which $\boldsymbol{H}_p\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{H}_p$ telescope to $\varepsilon_\Psi\boldsymbol{I} - \boldsymbol{K}_\varepsilon\boldsymbol{H}_p^{-1}$ and $\varepsilon_\Psi\boldsymbol{I} - \boldsymbol{H}_p^{-1}\boldsymbol{K}_\varepsilon$, so

$$\mathcal{L}_{\boldsymbol{H}_p}\big(\boldsymbol{W}^\star - D\mathcal{G}[\boldsymbol{W}^\star]\big) = \frac{\eta M_2}{4}\Big[2\varepsilon_\Psi\boldsymbol{I} - \big(\boldsymbol{K}_\varepsilon\boldsymbol{H}_p^{-1} + \boldsymbol{H}_p^{-1}\boldsymbol{K}_\varepsilon\big)\Big],$$

manifestly symmetric, and reproducing the directly computed quantity to 10^{-9}. Hence

$$D\mathcal{F}(\boldsymbol{W}^\star)[\boldsymbol{W}^\star] \preceq \boldsymbol{W}^\star \iff \lambda_{\max}\big(\boldsymbol{K}_\varepsilon\boldsymbol{H}_p^{-1} + \boldsymbol{H}_p^{-1}\boldsymbol{K}_\varepsilon\big) \le 2\varepsilon_\Psi. \tag{42}$$

Writing $\boldsymbol{K}_\varepsilon = \boldsymbol{H}_p^{1/2}\boldsymbol{S}_\varepsilon\boldsymbol{H}_p^{1/2}$, whose $\boldsymbol{S}_\varepsilon$ has the spectrum of $\boldsymbol{H}_p^{-1}\boldsymbol{K}_\varepsilon$, the quadratic form of the anticommutator is $2\langle\boldsymbol{H}_p^{1/2}\boldsymbol{x}, \boldsymbol{S}_\varepsilon\boldsymbol{H}_p^{-1/2}\boldsymbol{x}\rangle$, and the Kantorovich bound on $\|\boldsymbol{H}_p^{1/2}\boldsymbol{x}\|\|\boldsymbol{H}_p^{-1/2}\boldsymbol{x}\|$ gives

$$\lambda_{\max}\big(\boldsymbol{K}_\varepsilon\boldsymbol{H}_p^{-1} + \boldsymbol{H}_p^{-1}\boldsymbol{K}_\varepsilon\big) \le \lambda_{\max}(\boldsymbol{H}_p^{-1}\boldsymbol{K}_\varepsilon)\big(\sqrt{\varkappa} + 1/\sqrt{\varkappa}\big), \tag{43}$$

$\varkappa$ the condition number of $\boldsymbol{H}_p$. So (42) follows from the Perron criterion (35) strengthened by the factor $(\sqrt{\varkappa} + 1/\sqrt{\varkappa})/2$, which is one for a well conditioned $\boldsymbol{H}_p$. We find $\varkappa$ between 2.4 and 8.0 and (42) satisfied at every setting, its two sides being 1.44–1.98 against 1.92–3.83.

Hence if $D\mathcal{F}$ also maps the semidefinite cone into itself and (42) is strict, then $\boldsymbol{W}^\star - D\mathcal{F}[\boldsymbol{W}^\star] \succ 0$. Defining θ by the contraction bound

$$\theta = 1 - \lambda_{\min}\Big[(\boldsymbol{W}^\star)^{-1/2}\big(\boldsymbol{W}^\star - D\mathcal{F}(\boldsymbol{W}^\star)[\boldsymbol{W}^\star]\big)(\boldsymbol{W}^\star)^{-1/2}\Big]$$

gives $D\mathcal{F}[\boldsymbol{W}^\star] \preceq \theta\boldsymbol{W}^\star$ with $\theta < 1$. The cone Perron test therefore yields $\rho(D\mathcal{F}(\boldsymbol{W}^\star)) \le \theta < 1$, and the discrete-time linearisation theorem proves the local exponential bound in Theorem 4.6. With a non-strict test one obtains only $\rho(D\mathcal{F}) \le 1$, not local exponential attraction.

Cone positivity is where the step size enters, and we can locate it but not derive it. Completing the square isolates what is missing:

$$\boldsymbol{V} - \eta(\boldsymbol{H}_p\boldsymbol{V} + \boldsymbol{V}\boldsymbol{H}_p) = (\boldsymbol{I} - \eta\boldsymbol{H}_p)\boldsymbol{V}(\boldsymbol{I} - \eta\boldsymbol{H}_p) - \eta^2\boldsymbol{H}_p\boldsymbol{V}\boldsymbol{H}_p, \tag{44}$$

an identity we confirm to machine precision. The first term is a congruence and preserves the cone; the deficit is exactly $\eta^2\boldsymbol{H}_p\boldsymbol{V}\boldsymbol{H}_p$, and its growth as η^2 is why a step-size threshold appears at all.

It is tempting to read the deficit as an artifact of the isotropic truncation $\boldsymbol{\Sigma}_s = M_2\boldsymbol{I}$, since the batch map keeps a finite-rank diffusion term and the same completion turns its linear part into $(\boldsymbol{I} - \eta p\boldsymbol{H})\boldsymbol{V}(\boldsymbol{I} - \eta p\boldsymbol{H}) + \eta^2[(\mu_4 - \mu_2) - p^2]\boldsymbol{H}\boldsymbol{V}\boldsymbol{H}$. That reading is wrong: $(\mu_4 - \mu_2) - p^2$ reaches $2 - 1$ only as the clip disappears, and is negative at every finite κ: both terms vanish as $\kappa \to 0$, where the difference behaves as $-\frac{2}{\pi}\kappa^2$, so the minimum is interior and equals -0.109 at $\kappa = 0.732$, in the moderate-gating regime $p = 0.536$. So the batch map is not unconditionally cone positive either, and restoring the dropped per-sample term would not remove the threshold.

Nor does the feedback simply cover the deficit: sampling rank-one directions, $\sum_i \varpi_i(\boldsymbol{V})\boldsymbol{B}_i \succeq \eta^2\boldsymbol{H}_p\boldsymbol{V}\boldsymbol{H}_p$ holds in only 0–4% of them, while $D\mathcal{F}[\boldsymbol{V}] \succeq 0$ holds throughout the admissible range.

The congruence does not suffice alone either, since for $\boldsymbol{V} = \boldsymbol{v}\boldsymbol{v}^\top$ it leaves a difference of two rank-one terms, positive semidefinite only when they are parallel. All three pieces are needed together, which is why no inequality among them separately reproduces the condition. What adversarial search over rank-one directions first suggests is the clean threshold

$$2\eta\,\lambda_{\max}(\boldsymbol{H}_p) \le 1, \tag{45}$$

which separates our original sweep without exception: the six settings satisfying (45) keep $D\mathcal{F}[\boldsymbol{V}] \succeq 0$ with margin 0.013 to 0.027, and the five that violate it all fail. It is the per-sample counterpart of $\eta\lambda_1 \le \frac{1}{2}$ in Theorem 4.4. We will show below that (45) is genuinely one of the two requirements the problem imposes, but only one: over a wider search it is *not* sufficient, and the complete condition is (48).

The constant is not arbitrary. Pushing (44) one step further, $\boldsymbol{U} := \boldsymbol{I} - \eta\boldsymbol{H}_p$ and $\boldsymbol{Z} := \eta\boldsymbol{H}_p\boldsymbol{U}^{-1}$ are both functions of $\boldsymbol{H}_p$, hence commute, and $\boldsymbol{U}\boldsymbol{Z} = \eta\boldsymbol{H}_p$, so the linear part factors as

$$\mathcal{L}_0(\boldsymbol{V}) = \boldsymbol{U}\big[\boldsymbol{V} - \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}\big]\boldsymbol{U}, \qquad \boldsymbol{Z} = \eta\boldsymbol{H}_p(\boldsymbol{I} - \eta\boldsymbol{H}_p)^{-1},$$

an identity we confirm to 2×10^{-15}. Congruence by the symmetric $\boldsymbol{U}$ preserves the cone, so $\mathcal{L}_0$ is cone positive exactly when $\boldsymbol{V} \mapsto \boldsymbol{V} - \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}$ is, and the eigenvalues of $\boldsymbol{Z}$ are $\eta\nu_a/(1-\eta\nu_a)$, so (45) is precisely $\|\boldsymbol{Z}\| \le 1$. The threshold therefore says that the deficit operator $\boldsymbol{V} \mapsto \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}$ is a contraction, which is what the adversarial search is detecting.

It is still not a derivation. In the eigenbasis of $\boldsymbol{Z}$ the map $\boldsymbol{V} - \boldsymbol{Z}\boldsymbol{V}\boldsymbol{Z}$ is a Schur product with $[1 - r_a r_b]$, of the form $\mathbf{1}\mathbf{1}^\top - \boldsymbol{r}\boldsymbol{r}^\top$, which is positive semidefinite only when every r_a coincides; and indeed at $\|\boldsymbol{Z}\| \le 1$ we still measure smallest eigenvalues of -0.047 to -0.208 on rank-one probes. Even at the threshold the linear part alone fails, and the feedback remains necessary.

A third reading says the same thing and places the question in a known one. Because $\ell_{ab} = x_a + x_b$ with $x_a = \frac{1}{2} - \eta\nu_a$, the entrywise action is an anticommutator,

$$\mathcal{L}_0(\boldsymbol{V}) = \boldsymbol{X}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{X}, \qquad \boldsymbol{X} := \tfrac{1}{2}\boldsymbol{I} - \eta\boldsymbol{H}_p, \tag{46}$$

and (45) is exactly $\boldsymbol{X} \succeq 0$. So cone positivity of the linear part is the classical question of when the anticommutator of two positive semidefinite matrices is positive semidefinite, and the answer is that it generally is not: we find violations even at $\varkappa(\boldsymbol{X}) = 1.5$, and $\varkappa(\boldsymbol{X})$ in any case diverges as the threshold is approached, reaching 90 and 239 at our closest settings. The feedback is therefore required on both sides of the threshold, and what the threshold decides is only the sign of $\boldsymbol{X}$.

Cone positivity is a two-by-two condition. The question is finite, and smaller than it looks. $D\mathcal{F}$ is linear and the semidefinite cone is generated by its rank-one elements, so cone positivity is $D\mathcal{F}[\boldsymbol{v}\boldsymbol{v}^\top] \succeq 0$ for every unit $\boldsymbol{v}$, and on such a probe the linear part (46) is an outer-product pair,

$$\mathcal{L}_0(\boldsymbol{v}\boldsymbol{v}^\top) = (\boldsymbol{X}\boldsymbol{v})\boldsymbol{v}^\top + \boldsymbol{v}(\boldsymbol{X}\boldsymbol{v})^\top,$$

supported on the plane spanned by $\boldsymbol{v}$ and the part of $\boldsymbol{X}\boldsymbol{v}$ orthogonal to it. In an orthonormal basis $(\boldsymbol{v}, \boldsymbol{w})$ of that plane it is $\left[\begin{smallmatrix} 2a & b \\ b & 0 \end{smallmatrix}\right]$, and both entries are moments of the spectrum of $\boldsymbol{H}_p$ in the state $\boldsymbol{v}$:

$$a = \tfrac{1}{2} - \eta\,\mathbb{E}_{\boldsymbol{v}}[\nu], \qquad b = \eta\sqrt{\mathrm{Var}_{\boldsymbol{v}}(\nu)}, \tag{47}$$

where $\mathbb{E}_{\boldsymbol{v}}$ averages the eigenvalues ν_a of $\boldsymbol{H}_p$ against the weights $\langle \boldsymbol{v}, \boldsymbol{e}_a\rangle^2$; the second follows from $\|P_{\boldsymbol{v}^\perp}\boldsymbol{H}_p\boldsymbol{v}\|^2 = \|\boldsymbol{H}_p\boldsymbol{v}\|^2 - (\boldsymbol{v}^\top\boldsymbol{H}_p\boldsymbol{v})^2$, and we confirm it to 10^{-16}. Its eigenvalues are $a \pm \sqrt{a^2+b^2}$, so for $b \neq 0$ exactly one is negative, of size $\|\boldsymbol{X}\boldsymbol{v}\| - \boldsymbol{v}^\top\boldsymbol{X}\boldsymbol{v} = \sqrt{a^2+b^2} - a$; for $b = 0$ they are $2a$ and 0, and neither is negative once $a \ge 0$.

Splitting $\mathbb{R}^d$ into that plane and its complement and writing $\Phi := D\mathcal{F} - \mathcal{L}_0$ blockwise, the Schur complement $\tilde{\boldsymbol{P}}$ of $\Phi(\boldsymbol{v}\boldsymbol{v}^\top)$ onto the plane makes the test exact:

$$D\mathcal{F}[\boldsymbol{v}\boldsymbol{v}^\top] \succeq 0 \iff (\tilde{p}_{11} + 2a)\,\tilde{p}_{22} \ge (\tilde{p}_{12} + b)^2, \qquad \tilde{p}_{22} \ge 0, \quad \tilde{p}_{11} + 2a \ge 0. \tag{48}$$

The two sides agree in verdict on every probe at every setting we solved.

Two things follow at once. First, b vanishes exactly when $\boldsymbol{v}$ is an eigenvector of $\boldsymbol{H}_p$, and (48) is then automatic: the diagonal is always covered and every obstruction is off-diagonal — which is what the

adversarial search reports — and $b = \eta\sqrt{\mathrm{Var}_{\boldsymbol{v}}(\nu)}$ says exactly how much obstruction a direction carries. Second, the step-size threshold is nothing but $a \geq 0$ for every $\boldsymbol{v}$, since (45) $\iff \boldsymbol{X} \succeq 0$. Below it the $(1,1)$ entry of $-\mathcal{L}_0|_2$ is non-positive and only the off-diagonal must be covered; dropping $\tilde{p}_{11}$ from (48), a strictly weaker sufficient condition, already separates our original sweep without exception, holding at every setting below the threshold and failing at every setting above it.

What has to do the covering is the feedback, and its closed form explains the n dependence. The gates move only through the amplitudes $\sigma_i^2 = \langle \boldsymbol{H}_i^2, \boldsymbol{W}\rangle$, and $\mathrm{d}\log p_i / \mathrm{d}\log \sigma_i^2 = -\varepsilon_i/2$, so

$$\Phi(\boldsymbol{V}) = \frac{\eta}{2}\big(\mathcal{H}(\boldsymbol{V})\boldsymbol{W}^\star + \boldsymbol{W}^\star\mathcal{H}(\boldsymbol{V})\big) + \eta^2\dot{M}_2[\boldsymbol{V}]\boldsymbol{I}, \qquad \mathcal{H}(\boldsymbol{V}) := \frac{1}{n}\sum_i p_i q_i \boldsymbol{H}_i,$$

with $q_i := \varepsilon_i\langle \boldsymbol{H}_i^2, \boldsymbol{V}\rangle/\sigma_i^2 \geq 0$ for $\boldsymbol{V} \succeq 0$; this reproduces a finite-difference Φ to 10^{-9}. Stationarity pins its isotropic part exactly, because $\boldsymbol{W}^\star$ solves $\boldsymbol{H}_p\boldsymbol{W}^\star + \boldsymbol{W}^\star\boldsymbol{H}_p = \eta M_2\boldsymbol{I}$ and hence $n^{-1}\sum_i p_i\boldsymbol{B}_i = \eta M_2\boldsymbol{I}$, an identity we confirm to 10^{-13}. So if the q_i coincide then $\mathcal{H} = q\boldsymbol{H}_p$ and

$$\Phi(\boldsymbol{V}) = \Big(\tfrac{1}{2}q\,\eta^2 M_2 + \eta^2\dot{M}_2[\boldsymbol{V}]\Big)\boldsymbol{I},$$

purely isotropic, $\tilde{p}_{12} = 0$ and $\tilde{p}_{11} = \tilde{p}_{22}$. That is the batch case, and at $n = 1$ we measure the gate spread and the anisotropy of Φ to be zero identically, while for $n > 1$ the two rise together, from 0.04 to 0.57 as heterogeneity and n grow. The anisotropy that (48) has to absorb is exactly the spread of the gate elasticities across samples — the same mechanism that separates per-sample from batch in (32) and in Proposition B.7.

This also shows that the threshold cannot be the whole story. Under uniform gates (48) collapses to the scalar

$$\big(\gamma + 1 - 2\eta\,\mathbb{E}_{\boldsymbol{v}}[\nu]\big)\gamma \;\geq\; \eta^2\,\mathrm{Var}_{\boldsymbol{v}}(\nu), \qquad \gamma := \tfrac{1}{2}q\,\eta^2 M_2 + \eta^2\dot{M}_2, \tag{49}$$

and as (45) is approached the slack $1 - 2\eta\mathbb{E}_{\boldsymbol{v}}[\nu]$ vanishes on the top eigenvector, leaving $\gamma \geq \eta\sqrt{\mathrm{Var}_{\boldsymbol{v}}(\nu)}$ — a condition on the feedback against the curvature spread, in which the step-size threshold has been spent and only the feedback can close the gap. The step size itself has not disappeared: it enters γ at order η^2 and the right-hand side at order η. At the closest homogeneous setting we solve, $2\eta\lambda_{\max}(\boldsymbol{H}_p) = 0.996$, the two sides are $\gamma = 0.0772$ against $b = 0.0695$: an 11% margin, and nothing forces it to stay positive.

It does not. Searching over dimension, sample count, heterogeneity, clip level and step size, and keeping only settings that satisfy (45), produces explicit failures. At $d = 4$, $n = 2$, $C = 1$ and η sixteen times $2/\mathrm{Tr}(\boldsymbol{H})$, the fixed point is solved to a residual of 9×10^{-15} and $2\eta\lambda_{\max}(\boldsymbol{H}_p) = 0.9973$, yet the worst rank-one probe gives $\lambda_{\min} D\mathcal{F}[\boldsymbol{v}\boldsymbol{v}^\top] = -0.0030$, stable across finite-difference steps from 3×10^{-6} to 3×10^{-5}, and (48) reports the same verdict with $(\tilde{p}_{11} + 2a)\tilde{p}_{22} = 0.0190$ against $(\tilde{p}_{12} + b)^2 = 0.0205$. There $a = 0.18$ is comfortably positive while $b = 0.164$ is not: the threshold has done its job on the mean and the variance has defeated it anyway. So (45) is necessary for the argument as structured and sufficient on the settings of our original sweep, but it is not sufficient in general, and the correct statement is (48).

The Schur entries, bounded by the gate spread. (48) is exact but still refers to the feedback only through its Schur block. Both can be removed. Whenever $\Phi(\boldsymbol{v}\boldsymbol{v}^\top) \succ 0$, conjugating the whole test by $\Phi^{-1/2}$ leaves the linear part an outer-product pair again, $\Phi^{-1/2}\mathcal{L}_0\Phi^{-1/2} = \tilde{\boldsymbol{g}}\tilde{\boldsymbol{v}}^\top + \tilde{\boldsymbol{v}}\tilde{\boldsymbol{g}}^\top$ with $\tilde{\boldsymbol{v}} = \Phi^{-1/2}\boldsymbol{v}$ and $\tilde{\boldsymbol{g}} = \Phi^{-1/2}\boldsymbol{X}\boldsymbol{v}$, so the criterion collapses to a single scalar:

$$D\mathcal{F}[\boldsymbol{v}\boldsymbol{v}^\top] \succeq 0 \iff \delta_\Phi(\boldsymbol{v}) := \sqrt{(\boldsymbol{v}^\top\Phi^{-1}\boldsymbol{v})(\boldsymbol{v}^\top\boldsymbol{X}\Phi^{-1}\boldsymbol{X}\boldsymbol{v})} - \boldsymbol{v}^\top\boldsymbol{X}\Phi^{-1}\boldsymbol{v} \;\leq\; 1. \tag{50}$$

This is the deficit $\|\boldsymbol{X}\boldsymbol{v}\| - \boldsymbol{v}^\top\boldsymbol{X}\boldsymbol{v}$ of the unweighted problem, re-measured in the metric the feedback itself induces; no Schur complement and no plane appear. It agrees with the direct spectrum on every probe we test.

What controls Φ is one scalar per direction. Conjugate the feedback by $\boldsymbol{H}_p^{-1/2}$: with $\boldsymbol{G}_i := \boldsymbol{H}_p^{-1/2}\boldsymbol{H}_i\boldsymbol{H}_p^{-1/2}$ and $w_i := p_i/n$, the identity $\sum_i w_i\boldsymbol{G}_i = \boldsymbol{I}$ of the global route makes

$$\boldsymbol{S}_q := \boldsymbol{H}_p^{-1/2}\mathcal{H}\boldsymbol{H}_p^{-1/2} = \sum_i w_i q_i \boldsymbol{G}_i$$

a $\boldsymbol{G}$-weighted *average* of the gate elasticities. If the $\boldsymbol{H}_i$ are positive semidefinite the weights $w_i \boldsymbol{x}^\top \boldsymbol{G}_i \boldsymbol{x}$ are convex and $\mathrm{spec}(\boldsymbol{S}_q) \subseteq [q_{\min}, q_{\max}]$ outright; in general they need not be, and the negative-mass count $N^\star := \sum_i w_i \max\big(0, -\lambda_{\min}(\boldsymbol{G}_i)\big)$ — the same one that later rescues (34) without assuming $\boldsymbol{H}_i \succeq 0$ — applies verbatim and gives $\mathrm{spec}(\boldsymbol{S}_q) \subseteq [q_{\min} - N^\star \Delta q,\ q_{\max} + N^\star \Delta q]$, hence

$$\lambda_{\max}(\boldsymbol{S}_q) - \lambda_{\min}(\boldsymbol{S}_q) \ \le\ \Delta q\,(1 + 2N^\star), \qquad \Delta q(\boldsymbol{v}) := \max_i q_i(\boldsymbol{v}) - \min_i q_i(\boldsymbol{v}).$$

Subtracting the midpoint $\bar{q}$ of $\mathrm{spec}(\boldsymbol{S}_q)$ from the gates leaves the feedback isotropic plus a remainder, $\Phi - \bar{\gamma}\boldsymbol{I} = \frac{\eta^2 M_2}{4}(\boldsymbol{Y}\boldsymbol{H}_p^{-1} + \boldsymbol{H}_p^{-1}\boldsymbol{Y})$ with $\boldsymbol{Y} = \boldsymbol{H}_p^{1/2}(\boldsymbol{S}_q - \bar{q}\boldsymbol{I})\boldsymbol{H}_p^{1/2}$, whose quadratic form is $2(\boldsymbol{H}_p^{1/2}\boldsymbol{x})^\top(\boldsymbol{S}_q - \bar{q}\boldsymbol{I})(\boldsymbol{H}_p^{-1/2}\boldsymbol{x})$. Kantorovich bounds the two factors exactly as in (43), but now from below as well as above, and therefore

$$\big\|\Phi(\boldsymbol{v}\boldsymbol{v}^\top) - \bar{\gamma}\boldsymbol{I}\big\| \ \le\ \epsilon(\boldsymbol{v}) := \frac{\eta^2 M_2}{4}\, K\,(1 + 2N^\star)\,\Delta q(\boldsymbol{v}), \qquad K := \frac{\sqrt{\varkappa} + 1/\sqrt{\varkappa}}{2}, \tag{51}$$

with $\bar{\gamma} := \frac{1}{2}\eta^2 M_2 \bar{q} + \eta^2 \dot{M}_2[\boldsymbol{v}\boldsymbol{v}^\top]$ and $\varkappa$ the condition number of $\boldsymbol{H}_p$. In particular the Schur entries of (48) are bounded below by the gate spread alone,

$$\tilde{p}_{11},\ \tilde{p}_{22} \ \ge\ \lambda_{\min}\Phi(\boldsymbol{v}\boldsymbol{v}^\top) \ \ge\ \bar{\gamma} - \epsilon(\boldsymbol{v}), \tag{52}$$

the first inequality because $\tilde{\boldsymbol{P}} = [(\Phi^{-1})|_{\mathrm{plane}}]^{-1}$ once $\bar{\gamma} > \epsilon$ forces $\Phi \succ 0$; and the off-diagonal is bounded by it too, since the $\bar{q}$ part of Φ is a multiple of the identity and $\boldsymbol{v} \perp \boldsymbol{w}$, leaving $\Phi_{12} = \frac{\eta^2 M_2}{2}\boldsymbol{v}^\top \mathrm{sym}(\boldsymbol{Y}\boldsymbol{H}_p^{-1})\boldsymbol{w}$ and $|\Phi_{12}| \le \frac{\eta^2 M_2}{4}(1 + 2N^\star)\sqrt{\varkappa}\,\Delta q(\boldsymbol{v})$.

Feeding the two-sided bound (51) into (50) closes the chain. Using $(\bar{\gamma} + \epsilon)^{-1}\boldsymbol{I} \preceq \Phi^{-1} \preceq (\bar{\gamma} - \epsilon)^{-1}\boldsymbol{I}$ termwise gives the sufficient condition

$$\bar{\gamma}\,\delta(\boldsymbol{v}) + \epsilon(\boldsymbol{v})\big(a + \sqrt{a^2 + b^2}\big) \ \le\ \bar{\gamma}\big(\bar{\gamma} - \epsilon(\boldsymbol{v})\big), \qquad \delta(\boldsymbol{v}) = \sqrt{a^2 + b^2} - a, \tag{53}$$

in which every quantity is a moment: a and b the mean and standard deviation of the curvature spectrum in the state $\boldsymbol{v}$ by (47), and ϵ the spread of the gate elasticities across samples by (51). At $\Delta q = 0$ the spread term vanishes and (53) reduces to $b^2 \le \bar{\gamma}(\bar{\gamma} + 2a)$, which is exactly the scalar condition (49): the per-sample criterion is the batch criterion degraded by one term, and that term is proportional to the gate spread.

All of this is checked. (50) agrees with the direct spectrum at every setting and every probe; (51) holds without exception; and at $n = 1$ we find $\Delta q \equiv 0$, $\epsilon \equiv 0$ and (52) an equality. For $n > 1$ the bound on $\lambda_{\min}\Phi$ stays within a factor 1.01 to 1.58 of the truth, and the loss is attributable: the Kantorovich step in (51) costs only 1.03 to 1.50, while the envelope $\sigma_{\boldsymbol{S}_q} \le \Delta q(1 + 2N^\star)/2$ costs 1.3 to 7.7, so the sharper form of ϵ is the one written with $\sigma_{\boldsymbol{S}_q}$ and the Δq version is what makes it legible. The assembled condition (53) then holds throughout the low-spread regime — every setting with $\Delta q \le 0.2$ — and goes vacuous once the gate spread is large, reaching $\Delta q = 3.9$ at our most heterogeneous settings, where ϵ swamps $\bar{\gamma}$. So the reduction is complete and the constants are not: what (53) still needs is an envelope for $\sigma_{\boldsymbol{S}_q}$ that does not pass through Δq.

Weakening (48) any further costs the conclusion. Dominating the negative eigenvalue isotropically, $\Phi(\boldsymbol{v}\boldsymbol{v}^\top) \succeq (\sqrt{a^2 + b^2} - a)\boldsymbol{I}$, is sufficient but already fails below the threshold, reaching -0.037; bounding $\tilde{p}_{12}$ by $\sqrt{\tilde{p}_{11}\tilde{p}_{22}}$ to remove it from (48) fails at -0.032, because the realised $\tilde{p}_{12}$ carries a helpful sign rather than the worst one; and replacing the Schur entries by $\lambda_{\min}\Phi$ and $\lambda_{\max}\Phi$ loses an order of magnitude. Maximising instead over the free constant subtracted from the q_i recovers $\lambda_{\min}\Phi$ exactly, but that is the tautology $\max_t\{t - \|\boldsymbol{A} - t\boldsymbol{I}\|\} = \lambda_{\min}(\boldsymbol{A})$ and carries no information. We therefore have local stability conditional on (48), now an explicit two-by-two inequality in closed spectral quantities rather than an unquantified cone statement, and verified throughout our range; what remains unproven is a lower bound on the Schur entries of the feedback in terms of the gate spread.

C Extended results and deferred statements

How to read this appendix. Its material is of two kinds and they have different standing. The first kind supports the main results directly: proofs, the exact finite-μ Jacobians, the experimental

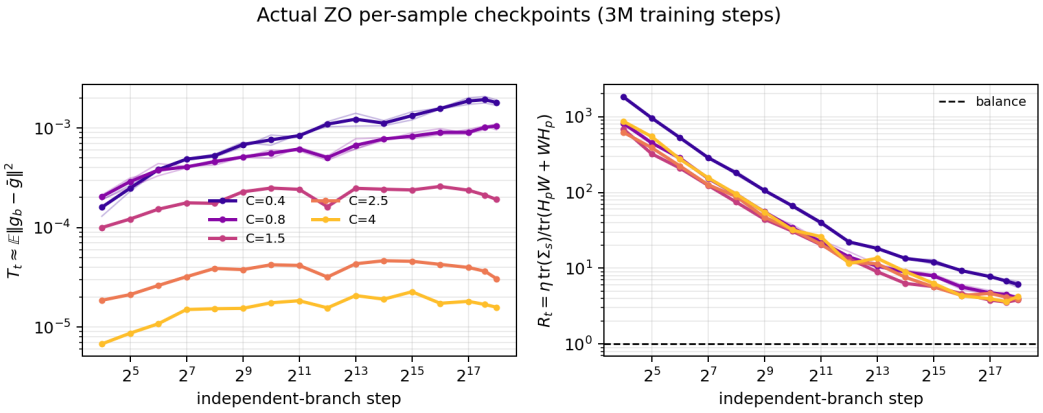

Figure 2: Branch diagnostics at real per-sample ZO checkpoints, eight independent-noise branches per clip, 262,144 steps. Faint curves are replicate seeds. **(a)** $T_t = \mathbb{E}\|\boldsymbol{g}_b - \bar{\boldsymbol{g}}\|^2$, which equals $\mathrm{Tr}(\boldsymbol{H}^2\boldsymbol{W}_t)$ in the local common-$\boldsymbol{H}$ quadratic model; it is flat over the last two decades. **(b)** The balance ratio R_t, which the matrix closure sets to one (dashed). It falls by three orders of magnitude and then flattens at 3.7–6.7. **(c)** Growth of $\mathrm{Tr}(\boldsymbol{W}_t)$ (solid) and of $\mathrm{Tr}(\boldsymbol{H}_p\boldsymbol{W}_t + \boldsymbol{W}_t\boldsymbol{H}_p)$ (dashed) relative to the first point, against a slope-one reference (dotted). The two agree until roughly 2^{11} steps; after that the curvature-weighted trace bends over and saturates while the raw trace continues to grow linearly. Panel (c) is why (b) does not reach one: the covariance keeps accumulating in directions that $\boldsymbol{H}_p$ barely sees, and the full-space trace ratio cannot equal one when $\boldsymbol{H}_p$ is near-singular.

protocol, and the pointwise numbers behind the sweeps. The second kind — the alignment analysis of Appendix C.13, the Krylov-plane corrected root of Appendix C.14, and the batch-ZO checkpoint comparisons in Appendix C.16 — concerns the accuracy of $\tau_\perp$ as the linearized batch-ZO threshold. Section 4.1.2 shows that this is not the edge a clipped run sits at, so those results should be read as statements about how well a scalar reduction approximates $\rho(\mathcal{T}_\eta) = 1$, and not as predictions about trajectories. We keep them because the failure of that linearized threshold to describe the runs is what motivates the saturated treatment, and a reader is entitled to see that the threshold was computed carefully before it was set aside. Where the older text says that $\tau_\perp$ is *the* batch criticality scalar, it should be read as the criticality scalar of the corresponding linear perturbation dynamics.

Beyond that, the appendix collects the justification for restricting to the positive-real branch, the deferred per-sample cluster refinement, the deferred ZO remarks, and the full empirical section with all figures and tables.

C.1 Gate order in the mean-square recursion

Lemma C.1 (Gate order in the mean-square recursion)**.** *For $\boldsymbol{J}(\boldsymbol{u}) = n^{-1}\sum_i a_i(\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{H}_i$ with symmetric $\boldsymbol{H}_i$,*

$$\boldsymbol{K} = \frac{1}{n}\sum_i \mathbb{E}[a_i\boldsymbol{u}\boldsymbol{u}^\top]\boldsymbol{H}_i, \qquad \mathcal{Q}(\boldsymbol{V}) = \frac{1}{n^2}\sum_{i,j}\mathbb{E}\big[a_ia_j(\boldsymbol{u}^\top\boldsymbol{H}_i\boldsymbol{V}\boldsymbol{H}_j\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top\big]. \tag{54}$$

Thus drift contains one gate, whereas diffusion contains a pair of gates. For a shared batch gate, $a_i \equiv a$ and $a_ia_j = a^2 = a$; distinct per-sample gates retain pairwise co-activation.

Proof. Since $\boldsymbol{J}(\boldsymbol{u}) = n^{-1}\sum_i a_i(\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{H}_i$, linearity of expectation gives $\boldsymbol{K} = \mathbb{E}[\boldsymbol{J}]$ and hence the first identity in (54). Expanding $\boldsymbol{J}(\boldsymbol{u})\boldsymbol{W}\boldsymbol{J}(\boldsymbol{u})^\top$ produces one factor a_i from the left copy of $\boldsymbol{J}$ and one factor a_j from the right, which gives the second identity. □

Mean-square stability balances an $O(\eta)$ drift term against an $O(\eta^2)$ diffusion term. An idempotent shared gate enters both at single-gate order, whereas a soft multiplier enters linearly and quadratically. For a Bernoulli gate independent of $\boldsymbol{u}$, Lemma C.9 makes cancellation exact for every gate rate p.

The slab gate used by clipped ZO depends on $\boldsymbol{u}$; Section C.9 therefore computes its moments exactly, and obtains gate-rate cancellation in the dominant gradient-orthogonal terms of the aligned model rather than assuming the independent-gate conclusion. Distinct gates retain pair moments, whose departure from the shared case is measured below.

Lemma C.6 in Appendix C makes this quantitative. With $p_i = \mathbb{E}[a_i]$, $\bar{p} = n^{-1}\sum_i p_i$ and the joint activation probabilities $c_{ij} = \mathbb{E}[a_i a_j]$, the *co-activation ratio* $\varsigma = n^{-2}\sum_{i,j} c_{ij}/\bar{p}^2$ lies in $[1, 1/\bar{p}]$; batch clipping attains the upper endpoint and decoupled gates the lower, and in the isotropic model, with $\tau = \mathrm{Tr}(\boldsymbol{H})$, $\eta^\star \simeq 2/(\varsigma\bar{p}\tau)$ — gate-free at the batch endpoint, proportional to $1/\bar{p}$ at the decoupled one. The c_{ij} are bivariate slab probabilities, available in closed form up to one quadrature (Proposition C.18), so ς is computed rather than sampled throughout.

Proofs are in Appendix E.

C.2 Why the positive-real branch

Remark C.2 (Why the positive-real branch). Restricting to $\Re\lambda > 0$ is not a convenience. The spectrum of $\mathbf{J}_m$ contains two other kinds of directions and neither carries information about the step size. (i) Gradient-direction removal contributes exact zeros: the deleted directions are neutral for the linearization and correspond to reparameterizations of the update scale, which the nonlinear clipping map fixes. (ii) Negative-curvature directions give $|1 - s\lambda| > 1$ for every $s > 0$; they describe escape from a saddle and are unstable with or without clipping, so including them would make $\eta_m^\star = 0$ for every rule and every step size. The EoS boundary is, here as in the unclipped literature, a statement about the positively curved block, and all comparisons with $2/\eta$ should be read that way. Our measured Hessians do have negative tail eigenvalues, so this is a genuine restriction and not a vacuous one.

C.3 Deferred statements

C.3.1 Blockwise FO clipping

Partition the parameter coordinates into L disjoint blocks, and let $\boldsymbol{E}_\ell$ embed block ℓ into the full parameter vector, so that $\boldsymbol{E}_\ell^\top \boldsymbol{E}_\ell = \boldsymbol{I}$, $\boldsymbol{E}_\ell^\top \boldsymbol{E}_j = 0$ for $\ell \neq j$, and $\sum_{\ell=1}^L \boldsymbol{E}_\ell \boldsymbol{E}_\ell^\top = \boldsymbol{I}$. The blockwise update is

$$\boldsymbol{x}_{t+1}^{\mathrm{FO,blk}} = \boldsymbol{x}_t - \eta \sum_{\ell=1}^{L} \boldsymbol{E}_\ell \,\mathrm{vclip}_C(\boldsymbol{E}_\ell^\top \boldsymbol{g}_t).$$

Write $\boldsymbol{g}^{(\ell)} = \boldsymbol{E}_\ell^\top \boldsymbol{g}$, $\alpha_\ell = \min\{1, C/\|\boldsymbol{g}^{(\ell)}\|\}$, with $\alpha_\ell = 1$ for a zero block, and $\chi_\ell = \mathbf{1}\{\|\boldsymbol{g}^{(\ell)}\| > C\}$. If $\mathcal{B}$ is the set of active blocks and $k = |\mathcal{B}|$, define

$$\boldsymbol{\Pi}_k = \sum_{\ell \in \mathcal{B}} \boldsymbol{E}_\ell \widehat{\boldsymbol{g}}^{(\ell)} \widehat{\boldsymbol{g}}^{(\ell)\top} \boldsymbol{E}_\ell^\top, \qquad \boldsymbol{Q}_k = \boldsymbol{I} - \boldsymbol{\Pi}_k,$$

and the derivative of the blockwise clipping map is

$$\boldsymbol{D} = \sum_{\ell=1}^{L} \alpha_\ell \boldsymbol{E}_\ell \big(\boldsymbol{I} - \chi_\ell \widehat{\boldsymbol{g}}^{(\ell)} \widehat{\boldsymbol{g}}^{(\ell)\top}\big) \boldsymbol{E}_\ell^\top. \tag{55}$$

Thus $\mathbf{J}_{\mathrm{FO,blk}} = \boldsymbol{D}\boldsymbol{H}$. When all L blocks are active and their coefficients are equal, $\boldsymbol{D} = \alpha \boldsymbol{Q}_L$. Theorem C.8 below quantifies the resulting removal of multiple directions.

Corollary C.3 (Clipping cannot escape a degenerate top cluster)**.** *Let $\boldsymbol{H} \succeq 0$. If $\boldsymbol{H}$ has k' eigenvalues within a relative window ϵ of λ_1, then any scale-invariant rule with fewer than k' active clipping blocks satisfies $q^{\mathrm{ex}} \geq (1-\epsilon)\alpha_{\min}\lambda_1(\boldsymbol{H})$. Gradient-direction removal alone therefore cannot lower the effective sharpness below the cluster; only the coefficient α can.*

Corollary C.4 (The corrected trace scalar)**.** *Under the hypotheses of Theorem 4.1, if in addition $\lambda_{\max}^{\perp}/\tau_\perp \leq \varepsilon_{\mathrm{rank}}$, then at $\eta = \eta_{\mathrm{ms}}^\star$*

$$\frac{1}{\bar{c} + 2\varepsilon_{\mathrm{rank}}} \frac{2}{\eta} \;\leq\; \tau_\perp \;\leq\; \frac{2}{\eta}. \tag{56}$$

Proposition C.5 (Exact per-sample drift). *For the per-sample ZO update, with $\kappa_i = C/\|\boldsymbol{g}_i\|$ and $(p_i, m_{2,i})$ the moments (75) evaluated at κ_i,*

$$\boldsymbol{K} = \frac{1}{n}\sum_i \boldsymbol{\Sigma}_{a_i}\boldsymbol{H}_i, \qquad \boldsymbol{\Sigma}_{a_i} = p_i\big(\boldsymbol{I} - \widehat{\boldsymbol{g}}_i\widehat{\boldsymbol{g}}_i^\top\big) + m_{2,i}\,\widehat{\boldsymbol{g}}_i\widehat{\boldsymbol{g}}_i^\top, \qquad \mathrm{Tr}(\boldsymbol{K}) = \frac{1}{n}\sum_i \mathrm{Tr}\big(\boldsymbol{\Sigma}_{a_i}\boldsymbol{H}_i\big). \tag{57}$$

No pair moments are required, and each $\boldsymbol{\Sigma}_{a_i}$ is a soft radial projection in the sense of Corollary C.12, now along the individual sample-gradient directions.

Lemma C.6 (Co-activation and the failure of cancellation). *Let $p_i = \mathbb{E}[a_i]$, $\bar{p} = n^{-1}\sum_i p_i$, let $c_{ij} = \mathbb{E}[a_i a_j]$ be the co-activation matrix, and define the* co-activation ratio

$$\varsigma = \frac{n^{-2}\sum_{i,j} c_{ij}}{\bar{p}^{\,2}} \;\in\; \Big[1,\; \tfrac{1}{\bar{p}}\Big]. \tag{58}$$

Batch clipping attains the upper endpoint $\varsigma = 1/\bar{p}$, and gates that decouple, $c_{ij} \to p_i p_j$ for $i \neq j$, give $\varsigma \to 1$. In the isotropic model $\boldsymbol{H}_i \equiv \boldsymbol{H} \succeq 0$ with the a_i independent of $\boldsymbol{u}$, write $\tau = \mathrm{Tr}(\boldsymbol{H})$. The mean-square boundary is $\frac{\eta\varsigma\bar{p}^{\,2}}{2}\sum_j \frac{\lambda_j}{\bar{p}-\eta\varsigma\bar{p}^{\,2}\lambda_j} = 1$, so at high effective rank

$$\eta^\star \;\simeq\; \frac{2}{\varsigma\,\bar{p}\,\tau}. \tag{59}$$

The critical step is therefore gate-free at the batch endpoint and proportional to $1/\bar{p}$ at the decoupled endpoint.

Lemma C.7 (Projected-Hessian spectrum). *The nonzero eigenvalues of $\boldsymbol{Q}_k\boldsymbol{H}$ are exactly the eigenvalues of the compression $\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k$ acting on $\mathrm{range}(\boldsymbol{Q}_k)$. If $\boldsymbol{H} \succeq 0$, then $\boldsymbol{Q}_k\boldsymbol{H}$ is diagonalizable with nonnegative real eigenvalues; in particular, its zero eigenvalue is semisimple.*

Theorem C.8 (Codimension-k compression). *Let $\lambda_1(\boldsymbol{H}) \geq \lambda_2(\boldsymbol{H}) \geq \cdots$ and let $q^{\mathrm{ex}}_{\mathrm{blk}} = \lambda_1(\mathbf{J}_{\mathrm{FO,blk}})$ be the top positive-real eigenvalue.* (i) *If all $k = L$ blocks are active and share a coefficient α, then $\mathbf{J}_{\mathrm{FO,blk}} = \alpha\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k$ on $\mathrm{range}(\boldsymbol{Q}_k)$ and*

$$\alpha\lambda_{k+1}(\boldsymbol{H}) \;\leq\; q^{\mathrm{ex}}_{\mathrm{blk}} \;\leq\; \alpha\lambda_1(\boldsymbol{H}). \tag{60}$$

Writing $\gamma_k = (\lambda_1 - \lambda_{k+1})/\lambda_1$ and $\eta^\star_{\mathrm{blk}} = 2/q^{\mathrm{ex}}_{\mathrm{blk}}$ therefore gives

$$\frac{2}{\alpha\lambda_1(\boldsymbol{H})} \;\leq\; \eta^\star_{\mathrm{blk}} \;\leq\; \frac{2}{(1-\gamma_k)\,\alpha\lambda_1(\boldsymbol{H})}. \tag{61}$$

(ii) *If instead $\boldsymbol{H} \succeq 0$ and the block coefficients differ, then*

$$\alpha_{\min}\lambda_{k+1}(\boldsymbol{H}) \;\leq\; q^{\mathrm{ex}}_{\mathrm{blk}} \;\leq\; \alpha_{\max}\lambda_1(\boldsymbol{H}), \tag{62}$$

where $\alpha_{\min} = \min_{1\leq\ell\leq L}\alpha_\ell$ and $\alpha_{\max} = \max_{1\leq\ell\leq L}\alpha_\ell$ range over all *L blocks, with $\alpha_\ell = 1$ on the inactive ones, so that $\alpha_{\max} = 1$ as soon as some block is inactive.*

Lemma C.9 (An idempotent shared gate does not move the mean-square edge). *Let $\boldsymbol{H} \succeq 0$ and consider $\boldsymbol{e}_{t+1} = (\boldsymbol{I} - \eta\boldsymbol{J}_t)\boldsymbol{e}_t$ with $\boldsymbol{J}_t = b_t\boldsymbol{u}_t\boldsymbol{u}_t^\top\boldsymbol{H}$, where $\boldsymbol{u}_t \sim \mathcal{N}(0, \boldsymbol{I}_d)$ and $b_t \in \{0,1\}$ is independent of $\boldsymbol{u}_t$ with $\Pr(b_t = 1) = p \in (0,1]$. Then the mean-square critical step $\sup\{s : \rho(\mathcal{T}_s) < 1\}$ is the same for every p, and in particular equals the ungated value $p = 1$.*

Proof. Conditioning on b_t gives $\mathcal{T}_s^{(p)} = (1-p)\mathcal{I} + p\mathcal{T}_s^{(1)}$ on the symmetric covariance space. The identity commutes with $\mathcal{T}_s^{(1)}$, so every eigenvalue λ is mapped to $1 - p + p\lambda$. Because $\mathcal{T}_s^{(1)}$ preserves the PSD cone, its spectral radius is a nonnegative Perron eigenvalue. Thus the first crossing of one occurs at the same s for every $p > 0$. □

Proposition C.10 (Exact covariance recursion). *For $\boldsymbol{e}_{t+1} = (\boldsymbol{I} - \eta\boldsymbol{J}(\boldsymbol{u}_t))\boldsymbol{e}_t$ with fresh independent directions, $\boldsymbol{V}_t = \mathbb{E}[\boldsymbol{e}_t\boldsymbol{e}_t^\top]$ satisfies*

$$\boldsymbol{V}_{t+1} = \mathcal{T}_\eta(\boldsymbol{V}_t), \quad \mathcal{T}_\eta(\boldsymbol{V}) = \boldsymbol{V} - \eta(\boldsymbol{K}\boldsymbol{V} + \boldsymbol{V}\boldsymbol{K}^\top) + \eta^2\mathcal{Q}(\boldsymbol{V}), \quad \mathcal{Q}(\boldsymbol{V}) = \mathbb{E}[\boldsymbol{J}(\boldsymbol{u})\boldsymbol{V}\boldsymbol{J}(\boldsymbol{u})^\top]. \tag{63}$$

Mean-square stability is governed by $\rho(\mathcal{T}_\eta)$, not by the conditional mean matrix $\boldsymbol{I} - \eta\boldsymbol{K}$.

For the Gaussian-reference diagnostics used below, whenever the symmetric drift $\boldsymbol{K}_s := (\boldsymbol{K} + \boldsymbol{K}^\top)/2$ is positive semidefinite we write

$$\boldsymbol{G} := \boldsymbol{K}_s, \qquad \tau := \mathrm{Tr}(\boldsymbol{G}) = \mathrm{Tr}(\boldsymbol{K}), \qquad \mathcal{Q}_{\boldsymbol{G}}(\boldsymbol{W}) := 2\boldsymbol{G}\boldsymbol{W}\boldsymbol{G} + \mathrm{Tr}(\boldsymbol{G}\boldsymbol{W}\boldsymbol{G})\boldsymbol{I}. \tag{64}$$

The associated finite-rank reference band is $\tau \le 2/\eta \le \tau + 2\lambda_1(\boldsymbol{G})$. It is exact for the Gaussian reference recursion with fourth moment $\mathcal{Q}_{\boldsymbol{G}}$; for per-sample clipping it is only a diagnostic band unless the pair-gate fourth moments are shown to obey that reference.

Corollary C.11 (Exact nonnormality of the drift)**.** *$\boldsymbol{K} - \boldsymbol{K}^\top = (m_2 - p)\big(\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top\boldsymbol{H} - \boldsymbol{H}\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top\big)$ and hence*

$$\|\boldsymbol{K} - \boldsymbol{K}^\top\|_2 = 2\kappa\phi(\kappa)\,\|\boldsymbol{Q}\boldsymbol{H}\widehat{\boldsymbol{g}}\|.$$

In particular $\boldsymbol{K}$ is exactly symmetric iff $\widehat{\boldsymbol{g}}$ is an eigenvector of $\boldsymbol{H}$, and the nonnormality vanishes at the Gaussian rate $2\kappa\phi(\kappa)$ as gating weakens.

Corollary C.12 (Gating is a soft radial projection)**.** *$\boldsymbol{\Sigma}_a = p\big(\boldsymbol{Q} + \theta\,\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top\big)$ with $\theta = m_2/p \in (0,1)$, $\theta \to 1$ as $\kappa \to \infty$ and $\theta = \kappa^2/3 + O(\kappa^4)$ as $\kappa \to 0$. Hence ZO scalar clipping attenuates the radial direction strictly more than the transverse ones and interpolates continuously between no gradient-direction suppression ($\theta = 1$) and complete gradient-direction removal ($\theta = 0$), as performed by active FO vector clipping.*

$$\boldsymbol{\Gamma} = \frac{1}{n}\sum_i \alpha_i\chi_i\widehat{\boldsymbol{g}}_i\widehat{\boldsymbol{g}}_i^\top, \qquad \overline{\alpha\chi} = \mathrm{Tr}\,\boldsymbol{\Gamma} = \frac{1}{n}\sum_i \alpha_i\chi_i, \qquad k_{\mathrm{eff}} = \frac{\mathrm{Tr}\,\boldsymbol{\Gamma}}{\lambda_1(\boldsymbol{\Gamma})}.$$

We use $\boldsymbol{\Gamma}$ and k_{eff} only as empirical summaries of the clipped sample-gradient directions; the stability result below is stated directly in terms of S and does not pass through $\lambda_1(\boldsymbol{A})$.

C.4 Why the scalar nonlinear balance is batch-only

The scalar reduction in Theorem 4.3 uses the one-gate batch identity: the anisotropic correction is positive semidefinite and bounded by a computable multiple of $\boldsymbol{H}\boldsymbol{W}\boldsymbol{H}$. Distinct per-sample gates instead produce pair-plane terms of either sign, so neither endpoint transfers. More fundamentally, the measured per-sample drift is not a scalar multiple of $\boldsymbol{H}$. Section 4.2 therefore replaces the attempted covariance bracket by the closure-free loss balance $\tau_{\mathrm{ps}}^{\mathrm{LB}} = \mathbb{E}[s^2\boldsymbol{u}^\top\boldsymbol{H}\boldsymbol{u}]/\mathbb{E}[sd]$. The closed form $\gamma(\kappa) = \mu_2/p$ remains batch-only: per sample, $\mathbb{E}[sd] = n^{-1}\sum_i p_i\langle\boldsymbol{H}_i\boldsymbol{\delta}, \boldsymbol{H}\boldsymbol{\delta}\rangle$ and $\kappa = C/\|\boldsymbol{g}\|$ does not determine γ_{ps}.

C.5 The stability edge and the saturation balance

Scope: a boundary and a balance. Theorems 4.1 and 4.3 describe different objects, and the difference is what clipping creates. The first bounds the step at which an *infinitesimal* perturbation in the frozen local model at a checkpoint starts to grow; the second selects the amplitude at which the nonlinear clipped covariance is stationary. Only the first is a stability boundary. The second is a fixed-point balance and, when Theorem 4.4 applies, is locally approached from both amplitude directions. Without clipping their scalar equalities coincide, which is why the edge-of-stability literature does not need this distinction. With clipping they separate, and by how much depends on the update rule.

In the first-order row the separation is small, and that can be derived rather than observed.

Proposition C.13 (FO batch as a saturated map)**.** *On $L = \frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{H}\boldsymbol{\delta}$ the active FO batch update is $\delta_j \mapsto \delta_j(1 - \eta\alpha\lambda_j)$ with $\alpha = C/\|\boldsymbol{H}\boldsymbol{\delta}\|$. A period-two orbit requires $(1 - a\lambda_j)(1 - b\lambda_j) = 1$ for every live j, with $a = \eta\alpha_t$ and $b = \eta\alpha_{t+1}$; dividing by $ab\lambda_j$ gives $1/a + 1/b = \lambda_j$. The left-hand side does not depend on j, so all live directions share a single eigenvalue $\lambda_\bullet$: they span one eigenspace of $\boldsymbol{H}$, which is one direction exactly when $\lambda_\bullet$ is simple. Taking $\lambda_\bullet = \lambda_1$, the mode that destabilises first and the one the simulations below select, and writing $A = a\lambda_1$, $B = b\lambda_1$,*

$$\frac{1}{A} + \frac{1}{B} = 1. \tag{65}$$

The linearized boundary $\eta\alpha\lambda_1 = 2$ is the symmetric solution $A = B = 2$ of (65), so it is not exact but sits inside the family the orbits occupy. Simulated at η from two to two hundred times $2/\lambda_1$ the

orbit is period-two to machine precision with one live mode and $1/A + 1/B = 1.00000$ throughout, with $\sqrt{AB} = 2.01$–2.12: on this range the linearised scalar is right to about 6%. The harmonic relation (65) constrains the pair (A, B) to a curve but leaves $\sqrt{AB}$ unbounded along it, so 6% is what the simulation found, not what the proposition guarantees. The same check run on the FO per-sample map puts $\eta\lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)/2$ at 0.985 and 0.990 at eight and thirty times threshold. So Theorem 3.1 and Proposition 3.2 may be read as statements about trajectories with an error of a few percent, and the first-order row needs no separate saturated treatment.

The zeroth-order row is where the boundary and balance separate. Theorem 4.3 already says how: $\eta \operatorname{Tr}(\boldsymbol{H}) = 2$ is the *bifurcation point* of the saturated dynamics up to its $O(\eta\lambda_1)$ correction, and past it the stationary amplitude sits at $\eta\tau = 2$, hence at $\eta\gamma \operatorname{Tr}(\boldsymbol{H}) \approx 2$ under the same correction, a distance γ^{-1} away that is unbounded as $\kappa^\star \to 0$. In the sweep summarized in Section 5 the measured separation $\eta \operatorname{Tr}(\boldsymbol{H})/2$ reaches 4.16, which is below γ^{-1} there because the trajectory sits at $\eta\gamma \operatorname{Tr}(\boldsymbol{H})/2 = 0.71$ rather than at one; the two numbers should not be equated. Propagating a tangent along the clipped trajectory confirms both halves: on a synthetic quadratic the mean-square tangent rate crosses zero at $\eta \operatorname{Tr}(\boldsymbol{H}) = 2$ and keeps rising past it ($+0.001$ at $1.1\times$ to $+0.024$ at $8\times$), the gate rate falling from 1.000 to 0.085 does not shift the crossing — the cancellation verified along a trajectory rather than at a chosen checkpoint — and the iterate stays bounded at every step size, so past the boundary the trajectory separates without the iterate growing. Why the separation is negligible in one row and not the other, and whether a smooth clip would remove it, we leave to Section 6.

Nothing in the derivation used the fact that the clipped scalar came from one gate. Every rule in the ZO row multiplies the shared probe by an aggregate scalar — $s_C = \operatorname{clip}_C(d)$ for batch, $s_C = n^{-1}\sum_i \operatorname{clip}_C(d_i)$ for per-sample, $s_C = G^{-1}\sum_a \operatorname{clip}_C(\bar{d}_a)$ for groups — against the same unclipped aggregate $d = \boldsymbol{u}^\top \boldsymbol{g}$, and the mean-square balance only ever sees two scalars built from that pair.

C.6 The linearized per-sample diffusion operator

Theorem C.14 (Exact per-sample diffusion operator)**.** *Assume $\boldsymbol{H}_i \equiv \boldsymbol{H}$, and let $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}$. For each pair let $\boldsymbol{E}_{ij}$ be the $d \times 2$ orthonormal basis of $\mathcal{P}_{ij}$, $\boldsymbol{\Pi}_{ij} = \boldsymbol{I} - \boldsymbol{E}_{ij}\boldsymbol{E}_{ij}^\top$, $\boldsymbol{M}_{\mathcal{P}} = \boldsymbol{E}_{ij}^\top \boldsymbol{M} \boldsymbol{E}_{ij}$, and let $G_{ij} = \mathbb{E}[a_i a_j]$, $\boldsymbol{S}_{ij} = \mathbb{E}[a_i a_j\, \boldsymbol{x}\boldsymbol{x}^\top]$ and $(T^4_{ij})_{abcd} = \mathbb{E}[a_i a_j\, x_a x_b x_c x_d]$ be the moments of Proposition C.18. Then*

$$\mathcal{Q}_{\mathrm{ps}}(\boldsymbol{W}) = \frac{1}{n^2}\sum_{i,j}\Big\{\boldsymbol{E}_{ij}\big[T^4_{ij}\cdot \boldsymbol{M}_{\mathcal{P}} + (\operatorname{Tr}\boldsymbol{M} - \operatorname{Tr}\boldsymbol{M}_{\mathcal{P}})\boldsymbol{S}_{ij}\big]\boldsymbol{E}_{ij}^\top + 2\big(\boldsymbol{E}_{ij}\boldsymbol{S}_{ij}\boldsymbol{E}_{ij}^\top \boldsymbol{M}\boldsymbol{\Pi}_{ij} + \boldsymbol{\Pi}_{ij}\boldsymbol{M}\boldsymbol{E}_{ij}\boldsymbol{S}_{ij}\boldsymbol{E}_{ij}^\top\big) + \langle \boldsymbol{M}_{\mathcal{P}}, \boldsymbol{S}_{ij}\rangle\, \boldsymbol{\Pi}_{ij} + G_{ij}\big(2\boldsymbol{\Pi}_{ij}\boldsymbol{M}\boldsymbol{\Pi}_{ij} + (\operatorname{Tr}\boldsymbol{M} - \operatorname{Tr}\boldsymbol{M}_{\mathcal{P}})\boldsymbol{\Pi}_{ij}\big)\Big\}, \tag{66}$$

and together with the exact drift (57) this determines $\mathcal{T}_\eta$ and hence the mean-square boundary $\rho(\mathcal{T}_\eta) = 1$. For $i = j$ the plane degenerates and (66) reduces term by term to (77).

This closes the analysis. The remaining modelling step is $\boldsymbol{H}_i \equiv \boldsymbol{H}$, which Lemma C.6 already assumes and which the exact drift does not need; we report both so its size is visible. What is no longer assumed is anything about the gate geometry.

C.7 The per-sample proof chain in full

This appendix examines the order-preservation hypothesis of Proposition 4.5 and the global route that Section 4.2 sets aside. The existence proposition assumes, explicitly, that M_2 is continuous and $\boldsymbol{W} \preceq \boldsymbol{W}' \Rightarrow M_2(\boldsymbol{W}) \le M_2(\boldsymbol{W}')$ on the covariance cone on which the closed map is iterated. This condition is not claimed on the entire PSD cone below; the discussion records where it can be checked and where the obstruction sits.

Two caveats. Positivity of the $\boldsymbol{H}_i$ fails at our checkpoints, where order preservation is nevertheless observed without exception; and the monotonicity of M_2 is genuinely directional. It is *not* true in the full Loewner order: writing M_2 as a function of the score covariance and differentiating by Price's theorem gives $\nabla_{\boldsymbol{\Sigma}} M_2 = n^{-2}\mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top] + n^{-1}\mathbb{E}[s\,\nabla^2 s]$ with $\boldsymbol{a}_i = \mathbf{1}\{|d_i| < C\}$, a second moment plus a boundary term at $\pm C$, and we find its smallest eigenvalue to be negative once $n \ge 2$. The

directions the proof uses are only the Gram cone $\{\mathbf{\Sigma}(\mathbf{\Delta}) : \mathbf{\Delta} \succeq 0\}$, of dimension $d(d+1)/2$ rather than $n(n+1)/2$, and there we observe no violation.

Along a ray the requirement has a clean form. Since $M_2(\lambda \boldsymbol{W}) = \lambda \Psi(C/\sqrt{\lambda})$, monotonicity in λ is $\varepsilon_\Psi \le 2$. We use throughout the saturation split of (37): $\bar{u}$ is the mean of the saturated parts $c\,\mathrm{sign}(d_i)\mathbf{1}\{|d_i| > c\}$, $\bar{w}$ the mean of the unsaturated parts $d_i \mathbf{1}\{|d_i| < c\}$, and $s_c = \bar{u} + \bar{w}$, so that $S_c = \bar{u}/c$. Since $\bar{u} - s_c = -\bar{w}$ identically,

$$\varepsilon_\Psi \le 2 \iff \mathbb{E}[s_c \bar{w}] \ge 0, \qquad \mathbb{E}[s_c \bar{w}] = \mathbb{E}[\bar{w}^2] + \mathbb{E}[\bar{u}\bar{w}].$$

The diagonal of $\mathbb{E}[\bar{u}\bar{w}]$ vanishes because $\mathbf{1}\{|d_i| > c\}$ and $\mathbf{1}\{|d_i| < c\}$ exclude each other, so only $O(\rho_{ij})$ cross terms remain. For one shared gate there are none and $\mathbb{E}[s_c \bar{w}] = \mathbb{E}[\bar{w}^2] \ge 0$ unconditionally, so batch clipping satisfies the hypothesis outright.

For $n > 1$ the inequality is not a general fact about Gaussian vectors. Price's theorem gives the cross term's derivative, $2[\int_{-c}^{c} \phi_2(c, y; \rho) dy - c(\phi_2(c, c; \rho) + \phi_2(c, -c; \rho))]$, of no evident sign, and the Hermite expansion is no better: its leading term is $\boldsymbol{a}^\top \boldsymbol{\rho}\, \boldsymbol{b} - \sum_i a_i b_i$ with $a_i = 2\phi(\kappa_i)$ and $b_i = \sigma_i m_2(\kappa_i)$, and $\boldsymbol{a}^\top \boldsymbol{\rho}\, \boldsymbol{b}$ can be negative for positive semidefinite $\boldsymbol{\rho}$ once $\boldsymbol{a} \ne \boldsymbol{b}$. Sampling arbitrary positive semidefinite score covariances with spread diagonals indeed produces violations, the worst reaching $\mathbb{E}[s_c \bar{w}] = -0.46\, \mathbb{E}[\bar{w}^2]$.

What makes the hypothesis plausible in the measured model is that arbitrary covariances do not occur. Only $\mathbf{\Sigma}(\boldsymbol{W})_{ij} = \mathrm{Tr}(\boldsymbol{H}_i \boldsymbol{H}_j \boldsymbol{W})$ arises, a Gram matrix of $\{\boldsymbol{H}_i \boldsymbol{W}^{1/2}\}$ lying in a set of dimension $d(d+1)/2$ rather than $n(n+1)/2$. Positivity of $\boldsymbol{H}_i$ and $\boldsymbol{W}$ does not itself make the Gram entries non-negative: $\mathrm{Tr}(\boldsymbol{H}_i \boldsymbol{H}_j \boldsymbol{W})$ can have either sign. In our draws the correlations are nevertheless mostly positive and large, with $\mathrm{mean}\,|\rho_{ij}| = 0.75$–$0.81$, which is the regime in which the leading Hermite term has the right sign. On these measured realizable covariances we find $\mathbb{E}[s_c \bar{w}] \ge 1.15\, \mathbb{E}[\bar{w}^2]$ throughout, so the cross term is not merely harmless but positive.

That structure can be used. Expanding the pair moments in Hermite polynomials, and keeping only odd orders because both factors are odd,

$$\mathbb{E}[s_c \bar{w}] = \sum_{k \text{ odd}} \frac{1}{n^2 k!}\, \boldsymbol{f}_k^\top \boldsymbol{\rho}^{\circ k} \boldsymbol{g}_k, \qquad (\boldsymbol{f}_k)_i = \sigma_i \mathbb{E}[\mathrm{clip}_{\kappa_i}(z) He_k(z)], \quad (\boldsymbol{g}_k)_j = \sigma_j \mathbb{E}[z \mathbf{1}\{|z| < \kappa_j\} He_k(z)], \tag{67}$$

where each $\boldsymbol{\rho}^{\circ k}$ is positive semidefinite by the Schur product theorem precisely because $\mathbf{\Sigma}$ is a Gram matrix. The leading term is unambiguous: $(\boldsymbol{f}_1)_i = \sigma_i p(\kappa_i)$ and $(\boldsymbol{g}_1)_j = \sigma_j m_2(\kappa_j)$ are strictly positive, so $\boldsymbol{f}_1^\top \boldsymbol{\rho}\, \boldsymbol{g}_1 \ge 0$ whenever $\boldsymbol{\rho}$ is entrywise non-negative. The higher orders are suppressed twice over, by $1/k!$ and by $\boldsymbol{\rho}^{\circ k}$, and across our realizable draws the tail never exceeds 0.71 of the leading term and is in fact itself positive at every instance, so the sum is positive with room to spare.

The coefficients in (67) are available in closed form. Gaussian integration by parts, $\mathbb{E}[F(z) He_k(z)] = \mathbb{E}[F^{(k)}(z)]$, with $\mathrm{clip}_\kappa^{(k)} = \delta^{(k-2)}(z + \kappa) - \delta^{(k-2)}(z - \kappa)$ and $\phi^{(m)} = (-1)^m He_m \phi$, gives for every odd $k \ge 3$

$$\mathbb{E}[\mathrm{clip}_\kappa(z) He_k(z)] = -2\phi(\kappa) He_{k-2}(\kappa), \qquad \mathbb{E}[z \mathbf{1}\{|z| < \kappa\} He_k(z)] = -2\phi(\kappa)\big[He_{k-2}(\kappa) + \kappa He_{k-1}(\kappa)\big], \tag{68}$$

matching quadrature to six digits, while $k = 1$ returns $p(\kappa)$ and $m_2(\kappa)$. This also explains the alternation: both coefficients carry the sign of $-He_{k-2}(\kappa)$, so $\boldsymbol{f}_k$ and $\boldsymbol{g}_k$ agree in sign except where He_{k-2} and $He_{k-2} + \kappa He_{k-1}$ straddle a zero, which is exactly the narrow window we observe near $\kappa \approx 2$ at $k = 5$.

Assembling these into a norm bound does not work. Cramér's inequality $|He_m(x)| \le K\sqrt{m!}\, e^{x^2/4}$ together with $\|\boldsymbol{\rho}^{\circ k}\| \le 1 + (n-1) r^k$, $r = \max_{i \ne j} |\rho_{ij}|$, produces a tail estimate exceeding the leading term by factors of 60 to 10^6, because the realizable r is 0.58–0.94 and $(n-1)r^k$ is already large at $k = 3$. The large correlation is counted purely as a hazard, when it is the very thing that makes the leading term positive; every norm bound we have tried in this section fails for the same reason. Nor does a termwise argument survive (68): using $He_m' = m He_{m-1}$, the two coefficients agree in sign exactly when $(1 + \kappa^2) He_m(\kappa) - \kappa He_m'(\kappa)$ shares the sign of $He_m(\kappa)$, which fails on one side of every zero of He_m, and He_m has $\lfloor m/2 \rfloor$ positive zeros.

A third route is exact but empty. Realizability lets one write $d_i = \langle \boldsymbol{v}_i, \boldsymbol{\xi} \rangle$ for a standard Gaussian $\boldsymbol{\xi}$ with $\boldsymbol{v}_i = \boldsymbol{H}_i \boldsymbol{W}^{1/2}$, so that $\langle \nabla_{\boldsymbol{\xi}} s_c, \boldsymbol{\xi} \rangle = \bar{w}$: the unsaturated part is the radial derivative of the

update. Stein's identity on s_c^2 then gives $\mathbb{E}[s_c\bar{w}] = \mathbb{E}[|\nabla s_c|^2] + \mathbb{E}[s_c \Delta s_c]$, whose first term is

$$\mathbb{E}[|\nabla s_c|^2] = \frac{1}{n^2}\operatorname{Tr}(\boldsymbol{C}\boldsymbol{\Sigma}), \qquad \boldsymbol{C} := \mathbb{E}[\boldsymbol{a}\boldsymbol{a}^\top], \quad a_i = \mathbf{1}\{|d_i| < c\}, \tag{69}$$

non-negative in one line because $\boldsymbol{C}$ is a second moment and $\boldsymbol{\Sigma}$ a Gram matrix. But the decomposition is the Ornstein–Uhlenbeck identity rearranged, so it carries no new information, and the two terms cancel to between 97% and 99.6%: $\mathbb{E}[s_c\bar{w}]$ is a small positive residual of large opposing quantities. No bound on the pieces can therefore establish it.

Three routes, three different obstructions, all of one kind: the quantity is a residual, and every estimate that discards structure discards more than the residual is worth.

What closes the question is not a sharper estimate but a change of object. Splitting the update at the clipping threshold rather than expanding it, $s_c = \bar{u} + \bar{w}$ as in (37), turns both elasticity bounds into the single cross-covariance $\mathbb{E}[\bar{u}\bar{w}]$, and that quantity has no diagonal: u_i and w_i have disjoint support, so only the off-diagonal pairs survive and each of them is settled in closed form by Proposition B.7, whose sign is exactly the sign of ρ_{ij}. On realizable Gram correlations that additionally satisfy $\rho_{ij} \geq 0$ entrywise, this gives $0 \leq \varepsilon_\Psi \leq 2$ for all n. Without that entrywise sign condition the elasticity bound remains a separate assumption to be checked. This is why the three routes above were failing: each tried to bound a residual of two large opposing quantities, where the split removes the cancellation instead of estimating around it. The Hermite and Stein material is kept because it explains the obstruction, and because (68) and (69) are used elsewhere.

The global route, and why it is not the one to take. Everything above is local. We now return to condition (34), whose role was to deliver uniqueness and global attraction by nonlinear Perron–Frobenius arguments, and record why that route closes off. Only the first of its two reductions used $\boldsymbol{H}_i \succeq 0$, and that hypothesis fails at our checkpoints: 46 of 48 projected per-sample Hessians are indefinite, although their gated mean stays positive definite along the whole iteration. It can be dropped. Put $\boldsymbol{G}_i := \boldsymbol{H}_p^{-1/2}\boldsymbol{H}_i\boldsymbol{H}_p^{-1/2}$, so that $\sum_i w_i \boldsymbol{G}_i = \boldsymbol{I}$ with $w_i = p_i/n$: the $\boldsymbol{G}_i$ resolve the identity but need not be positive. For a unit $\boldsymbol{v}$ the numbers $\gamma_i = w_i \boldsymbol{v}^\top \boldsymbol{G}_i \boldsymbol{v}$ sum to one while individual terms may be negative, so with $N(\boldsymbol{v}) := \sum_i \max(0, -\gamma_i)$ the positive mass is $1 + N$ and $\sum_i q_i\gamma_i \leq q_{\max}(1+N) - q_{\min}N$. Bounding $\max(0, -\boldsymbol{v}^\top \boldsymbol{G}_i \boldsymbol{v}) \leq \max(0, -m_i)$ with $m_i := \lambda_{\min}(\boldsymbol{G}_i)$ gives

$$\lambda_{\max}\big(\boldsymbol{H}_p(\boldsymbol{W})^{-1}\boldsymbol{H}_p(\lambda\boldsymbol{W})\big) \;\leq\; q_{\max} + N^\star\big(q_{\max} - q_{\min}\big), \qquad N^\star := \sum_i w_i \max(0, -m_i), \tag{70}$$

which needs only $\boldsymbol{H}_p \succ 0$. Since $N^\star = 0$ exactly when every $\boldsymbol{H}_i$ is positive semidefinite, (70) contains the positivity-based bound and prices the indefiniteness explicitly; the sufficient condition (36) becomes $\varepsilon_\Psi \geq 1 + N^\star$. The sharp criterion (35) never used positivity at all, being an exact eigenvalue derivative at $\boldsymbol{H}_p^{-1/2}\boldsymbol{H}_p\boldsymbol{H}_p^{-1/2} = \boldsymbol{I}$.

The indefiniteness is also quantitatively mild. On the synthetic control $N^\star = 0$ to machine precision and (70) reproduces $q_{\max}$ exactly; at the $C = 0.4$ checkpoint $N^\star = 0.042$, the bound is within 10^{-4} of the exact ratio at $\lambda = 0.99$ and within 2.8% at $\lambda = 0.2$, and the positivity-based $q_{\max}$, although no longer justified, is not in fact violated at any λ we tested.

From λ near one to all λ. Both sides of (34) equal one at $t = 1/\sqrt{\lambda} = 1$ and are exponentials of integrated logarithmic derivatives, so the statement for all $\lambda \in (0, 1)$ is the accumulated comparison

$$\int_1^t \Big[\varepsilon_\Psi(uC) - \varepsilon_R(u)\Big]\frac{du}{u} \;\geq\; 0, \qquad \varepsilon_R(u) := \frac{d\log}{d\log u}\lambda_{\max}\big(\boldsymbol{H}_p(\boldsymbol{W})^{-1}\boldsymbol{H}_p(u^{-2}\boldsymbol{W})\big), \tag{71}$$

of which (35) is the case $t \to 1$. Two cruder routes are available and both are too lossy to be worth using. Bounding $\varepsilon_R \leq 1 + N^\star$, which uses only $\varepsilon_i < 1$, asks for $\varepsilon_\Psi \geq 1 + N^\star$ along the whole path; and comparing the two elasticities pointwise asks for $\varepsilon_\Psi(uC) \geq \varepsilon_R(u)$ at every u. Neither certifies settings in which (34) holds by a wide margin: at $\eta \operatorname{Tr} \boldsymbol{H}/2 = 6$ the realised margin at $\lambda = 0.01$ is $+11.5$, while ε_Ψ has fallen to 0.86 against $\varepsilon_R = 0.97$ and both routes decline to certify. The reason is visible in (71): ε_Ψ exceeds ε_R by a wide margin early on the path and the surplus accumulated there is not undone by a small late deficit.

The surplus does not stay non-negative, so no proof of the global statement is possible. At $\eta \operatorname{Tr} \boldsymbol{H}/2 = 1.5$ the local criterion passes with margin $\delta(1) = +0.380$, yet the integrand turns negative immediately and the accumulated surplus crosses zero at $\lambda \approx 0.134$; the realised margins $+0.027, +0.009, -0.056, -0.098$ at $\lambda = 0.5, 0.2, 0.05, 0.01$ confirm the crossing without any quadrature. Since ε_Ψ is non-increasing at every setting we examined, the integrand changes sign at most once and $I(t)$ is unimodal, so what a per-sample argument can deliver is a basin in λ rather than all of $(0, 1)$.

That basin is explicit. If $\delta(1) = m > 0$ and $|u\,\delta'(u)| \le L$ along the path, then $\delta(u) \ge m - L \log u$ and $I(t) \ge m \log t - \frac{1}{2} L (\log t)^2$, so

$$\text{(34) holds for every } \lambda \;\ge\; \lambda_\star := \exp(-4m/L), \tag{72}$$

with m and L both measurable. Across our settings (72) gives $\lambda_\star = 0.83, 0.24, 0.028$ at $\eta \operatorname{Tr} \boldsymbol{H}/2 = 1.5, 2, 6$ against true crossings of 0.134 and, for the latter two, below 0.01: conservative in each case. This suffices for local exponential stability, which is what Theorem 4.4 claims in the batch case as well; only the global Perron–Frobenius route needs every λ, and (71) shows that route is unavailable in general. For one shared gate the question does not arise: the comparison holds pointwise at *every* κ, being $\gamma = \mu_2/p$ non-decreasing, so batch clipping satisfies (34) for all λ at once.

Bounding L. The constant in (72) can be bounded a priori. Differentiating $\varepsilon_\Psi = c\Psi'(c)/\Psi(c)$ gives the identity $d\varepsilon_\Psi / d\log c = \varepsilon_\Psi - \varepsilon_\Psi^2 + c^2\Psi''(c)/\Psi(c)$ with $\Psi' = 2\mathbb{E}[s_c S_c]$ and $\Psi'' = 2\mathbb{E}[S_c^2] + 2\mathbb{E}[s_c\, dS_c/dc]$, where $S_c = n^{-1}\sum_i \operatorname{sign}(d_i)\mathbf{1}\{|d_i| > c\}$. Cauchy–Schwarz bounds $\mathbb{E}[S_c^2] \le (n^{-1}\sum_i \sqrt{1-p_i})^2$, and $|s_c| \le c$ with the density $2\phi(\kappa_i^c)/\sigma_i$ of $|d_i|$ at c, $\kappa_i^c := c/\sigma_i$ being κ_i at the running threshold c, bounds $|\mathbb{E}[s_c\, dS_c/dc]| \le 2n^{-1}\sum_i \kappa_i^c \phi(\kappa_i^c) \le 2\phi(1)$. The piece that looks missing, a lower bound on Ψ, is free: $\varepsilon_\Psi \ge 0$ makes Ψ non-decreasing, so $\Psi(uC) \ge \Psi(C)$ along the path. Together these give an explicit bound on $|d\varepsilon_\Psi / d\log u|$, and it holds at every u in every setting we checked.

It is, however, loose by a factor of 18–27, which returns $\lambda_\star \approx 0.96$ and so a basin of no practical width. The looseness is localised: both $|s_c| \le c$ and the Cauchy–Schwarz step discard the cancellation among samples, and since $\mathbb{E}[\operatorname{sign}(d_i)\mathbf{1}\{|d_i| > c\}] = 0$ by symmetry the discarded cross terms are $O(\rho_{ij})$.

Refining them recovers part, but not all, of the lost factor. The diagonal of $\mathbb{E}[S_c^2]$ is $n^{-1}\operatorname{mean}_i(1-p_i)$, already a factor n below Cauchy–Schwarz, and Price's theorem with $(\kappa_i^c)^2 + (\kappa_j^c)^2 - 2r\kappa_i^c\kappa_j^c \ge (1-r)((\kappa_i^c)^2 + (\kappa_j^c)^2)$, $r = \max_{i\ne j}|\rho_{ij}|$, bounds the off-diagonal by $|\rho_{ij}|\,\frac{2}{\pi\sqrt{1-r^2}}\exp\!\big(-\frac{(\kappa_i^c)^2+(\kappa_j^c)^2}{2(1+r)}\big)$. Conditioning the boundary term on $|d_i| = c$ makes $\operatorname{clip}_C(d_i)\operatorname{sign}(d_i) = c$, so its self term is c/n rather than c, and its cross terms are $O(\rho_{ij})$ because clip_C is odd and 1-Lipschitz while $d_j \mid d_i = c$ has mean $\rho_{ij}(\sigma_j/\sigma_i)c$. Both refined bounds hold against Monte Carlo at every threshold we checked.

The gain is nevertheless conditional on weak correlation, since the n^2 cross terms improve on the diagonal only when $\sum_{i\ne j}|\rho_{ij}| \lesssim n$. On the synthetic model the refinement gives $9.7\times$, $8.1\times$ and $2.9\times$ at $\eta \operatorname{Tr} \boldsymbol{H}/2 = 1.5, 2, 6$ with $n = 16$; on the checkpoint curvature it gives $2.3\times$ and $1.2\times$. The reason is that the scores are not weakly correlated under the closed map's own fixed point: there $\operatorname{mean}|\rho_{ij}| = 0.37$ at $n = 48$, so $\sum_{i\ne j}|\rho_{ij}| \approx 835 \gg n$. In the measured positive semidefinite model the obstruction persists because the realised $\operatorname{Tr}(\boldsymbol{H}_i\boldsymbol{H}_j\boldsymbol{W})$ are predominantly non-negative; positive semidefiniteness alone would not force this sign. A usable basin therefore needs something other than a termwise Lipschitz bound — the signed sum $\sum_{i\ne j}\mathbb{E}[s_i s_j]$ is far smaller than $\sum_{i\ne j}|\mathbb{E}[s_i s_j]|$, and exploiting that cancellation is the step we leave open.

Monotonicity of ε_Ψ also gives a basin that avoids L altogether. Since $\varepsilon_R \le 1 + N^\star$, the surplus obeys $I(t) \ge \log\big(\Psi(tC)/\Psi(C)\big) - (1+N^\star)\log t$, and when ε_Ψ is non-increasing its log-average over $[1, t]$ is at least its endpoint value, so (34) holds whenever $\varepsilon_\Psi(C/\sqrt{\lambda}) \ge 1 + N^\star$. Neither basin dominates: at $\eta \operatorname{Tr} \boldsymbol{H}/2 = 2$ the L-based bound gives $\lambda_\star = 0.24$ against 0.56 for this one, and at $\eta \operatorname{Tr} \boldsymbol{H}/2 = 6$ it gives 0.028 against 0.019.

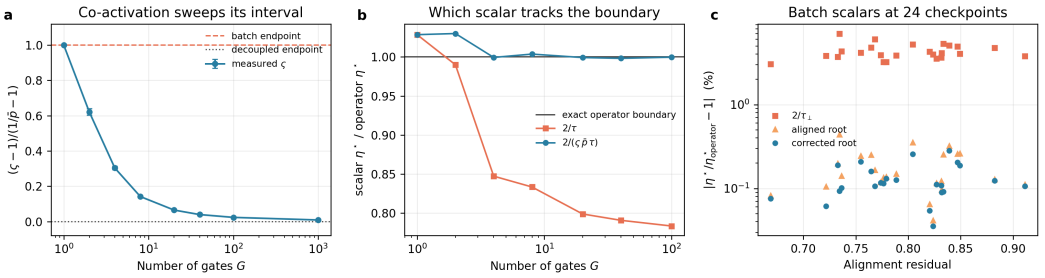

Figure 3: The ZO column as a continuum. **(a)** Position of the exact co-activation ratio inside its own interval, $(\varsigma - 1)/(1/\bar{p} - 1)$, which is one at the batch endpoint and zero at the decoupled one. It is exactly one at $G = 1$ and 0.008 at $G = 1000$. Normalising matters because the realised gate rate drifts by a few percent between branches, so $1/\bar{p}$ is not common to them; error bars are the spread over three late checkpoints. **(b)** The two candidate scalars divided by the exact operator boundary of Theorem C.14. They cross between $G = 2$ and $G = 4$: the trace-only threshold is right near the batch endpoint and the co-activation surrogate is right everywhere above it, which is the content of Corollary C.15. **(c)** Batch checkpoints: absolute relative error of the three scalars against the alignment-free operator boundary, plotted against the alignment residual. The vertical axis is logarithmic, and the aligned and corrected roots sit an order of magnitude below the trace threshold.

C.8 Group clipping

Group sweep: the co-activation ratio decides, and it is measurable. Branching eight runs from one checkpoint with per-G thresholds calibrated to a common gate rate makes the number of gates the only difference between runs. The co-activation ratio then leaves the batch endpoint exactly at $G = 1$ — measured $\varsigma = 1/\bar{p}$ to five digits, which is Lemma C.6's upper endpoint attained, not approached — and falls to within 0.2% of the decoupled endpoint at $G = 1000$. Against the exact boundary of Theorem C.14 the two candidate scalars *cross* between $G = 2$ and $G = 4$: the trace-only threshold $2/\tau$ is accurate at $G = 1, 2$ ($+2.9\%, -1.0\%$) and then falls away monotonically to -21.6% by $G = 100$, while the co-activation surrogate $2/(\varsigma\bar{p}\tau)$ is off by $+3.0\%$ at $G = 2$ but sits within 0.4% of the exact boundary for every $G \geq 4$, reaching -0.04% at $G = 100$. Over the 21 checkpoints with an exact operator the median absolute errors are 18.1% and 0.3% respectively. This is Corollary C.15 made quantitative: which scalar is correct is not a property of the rule's name but of where its gates sit in $[1, 1/\bar{p}]$, and the crossing is where the shared-gate cancellation stops paying for itself.

Group clipping interpolates the two columns. Deciding between $\tau_\perp$ and the gate-weighted trace is a question about *how many gates there are*, and nothing forces that number to be 1 or n. Partition the batch into G groups, average the directional derivative inside each group and gate that group scalar:

$$\boldsymbol{x}_{t+1}^{\mathrm{ZO,grp}} = \boldsymbol{x}_t - \eta\Big[\frac{1}{G}\sum_{a=1}^{G} \mathrm{clip}_C\big(\bar{d}_a^{\mu}(\boldsymbol{x}_t; \boldsymbol{u}_t)\big)\Big]\boldsymbol{u}_t, \qquad \bar{d}_a^{\mu} = \frac{G}{n}\sum_{i \in \mathcal{G}_a} d_i^{\mu}, \tag{73}$$

which is batch clipping at $G = 1$ and per-sample clipping at $G = n$. The gates are now $a_a(\boldsymbol{u}) = \mathbf{1}\{|\boldsymbol{u}^\top \bar{\boldsymbol{g}}_a| < C\}$ on the group gradients, so Lemma C.6 applies verbatim with G in place of n, and ς becomes a knob rather than a pair of endpoints. Because $\|\bar{\boldsymbol{g}}_a\|$ shrinks as groups grow, comparing group counts requires one threshold C_G per G, chosen so the gate rate $\bar{p}$ is matched; Proposition C.18 makes both the calibration and ς exact rather than sampled. Section 5 reports the resulting sweep.

Corollary C.15 (Which scalar summarises which rule)**.** *In the isotropic model of Lemma C.6, with $\tau = \mathrm{Tr}(\boldsymbol{H})$, at criticality*

$$\mathrm{Tr}(\boldsymbol{K}) \;\approx\; \bar{p}\,\tau \;\approx\; \bar{p}\cdot\frac{2}{\eta\,\varsigma\,\bar{p}} \;=\; \frac{1}{\varsigma}\cdot\frac{2}{\eta}.$$

Batch clipping has $\varsigma = 1/\bar{p}$, so the gate-weighted trace undershoots $2/\eta$ by the factor $\bar{p}$ and the correct scalar is $\tau_\perp$, as in Corollary C.25. Fully decoupled gates have $\varsigma = 1$, so the gate-weighted trace is *the criticality scalar. Per-sample clipping lies between the two endpoints, at a value of ς computable from the co-activation matrix.*

The two ZO rules are therefore the two endpoints of the same interval $[1, 1/\bar{p}]$, the group family fills it, and which scalar is correct at any point of it is decided by a quantity one can compute exactly from Proposition C.18. The shared-slab batch formulas, including Theorem C.20, do not apply to per-sample runs; (66) replaces them. The cost is $O(G^2)$ bivariate quadratures, one per gate pair, which is the reason the group family is the practical route to the per-sample endpoint: at $G = n$ the pair count is 10^6 for our runs, and the exact operator is assembled for $G \le 128$ and the co-activation ratio estimated without bias beyond that.

C.9 The batch linear analysis in full

Write $\kappa = C/\|\boldsymbol{g}\|$, $\boldsymbol{Q} = \boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$, and let ϕ, Φ be the standard normal density and cdf. Split the probe along the gradient as $\boldsymbol{u} = z\widehat{\boldsymbol{g}} + \boldsymbol{u}_\perp$ with $z = \boldsymbol{u}^\top \widehat{\boldsymbol{g}} \sim \mathcal{N}(0,1)$ and $\boldsymbol{u}_\perp = \boldsymbol{Q}\boldsymbol{u}$ independent of z. Since $\boldsymbol{g} = \|\boldsymbol{g}\|\widehat{\boldsymbol{g}}$, the batch gate reads

$$a(\boldsymbol{u}) = \mathbf{1}\{|\boldsymbol{u}^\top \boldsymbol{g}| < C\} = \mathbf{1}\{|z| < \kappa\}, \tag{74}$$

so the gate is a slab indicator in the single coordinate z: it is closed exactly when the probe aligns with the gradient, and it does not see $\boldsymbol{u}_\perp$ at all. Every gate moment we need is therefore a moment of z against that indicator,

$$\begin{aligned} p &= \mathbb{E}[a] = \Pr(|z| < \kappa) = 2\Phi(\kappa) - 1, \\ m_2 &= \mathbb{E}[a\, z^2] = p - 2\kappa\phi(\kappa), \\ m_4 &= \mathbb{E}[a\, z^4] = 3p - 2\kappa(\kappa^2 + 3)\phi(\kappa), \end{aligned} \tag{75}$$

available in closed form, so the two-sided envelope assumed in earlier analyses is unnecessary. The three are the same indicator weighted by successive powers of the radial coordinate: p is the *gate rate*, the probability that a step is taken at all, and m_2, m_4 are the second and fourth radial moments it survives with. As $\kappa \to \infty$ the gate is always open and they collapse to the ungated Gaussian moments $(p, m_2, m_4) \to (\mathbb{E}[1], \mathbb{E}[z^2], \mathbb{E}[z^4]) = (1, 1, 3)$, which is the limit in which every formula below must return the ungated theory.

The same slab supplies the exact gated second- and fourth-moment operators $\boldsymbol{\Sigma}_a = \mathbb{E}[a\, \boldsymbol{u}\boldsymbol{u}^\top]$ and $\mathcal{Q}$, which the linear analysis needs and we defer to Appendix C.10.

Two consequences replace previously empirical statements. The drift nonnormality is exactly $\|\boldsymbol{K} - \boldsymbol{K}^\top\|_2 = 2\kappa\phi(\kappa)\|\boldsymbol{Q}\boldsymbol{H}\widehat{\boldsymbol{g}}\|$, so $\boldsymbol{K}$ is symmetric iff $\widehat{\boldsymbol{g}}$ is an eigenvector of $\boldsymbol{H}$ (Corollary C.11); and $\boldsymbol{\Sigma}_a = p(\boldsymbol{Q} + \theta\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top)$ with $\theta = m_2/p \in (0,1)$ (Corollary C.12), so the gate rescales the transverse block by p and the radial block by $p\theta$. That factor is the one place the clip threshold survives in the drift, and it reappears below in $\mathrm{Tr}(\boldsymbol{K})$.

Why the gate rate cancels: on $\widehat{\boldsymbol{g}}^\perp$, gating is a time change. By (74) the gate is a function of z alone, and $\boldsymbol{u}_\perp$ is independent of z. Conditioned on the gate being open, the transverse probe is therefore still standard Gaussian on $\widehat{\boldsymbol{g}}^\perp$: gating changes nothing about the transverse dynamics except *whether a step is taken*. This is visible term by term. The transverse block of $\boldsymbol{\Sigma}_a$ in (76) is exactly $p\boldsymbol{Q}$, and the transverse diffusion in (77) is exactly $p\big(2\boldsymbol{M}_\perp + \mathrm{Tr}(\boldsymbol{M}_\perp)\boldsymbol{Q}\big)$ — the ungated transverse operator times the same p — plus the radial feed $m_2 m\boldsymbol{Q}$. Drift and diffusion therefore carry a common factor, and a common factor is a time change, which cannot move the boundary between mean-square growth and decay. Were the gate independent of $\boldsymbol{u}$ the statement would be exact and immediate: $\mathcal{T}_\eta = (1-p)\,\mathrm{Id} + p\,\mathcal{T}_\eta^{\mathrm{ung}}$, whose spectral radius is one precisely when $\rho(\mathcal{T}_\eta^{\mathrm{ung}}) = 1$, for *every* $p \in (0,1]$. Idempotence is what supplies the common factor: if the diffusion were a genuine pair-gate object it would carry p^2 against the drift's p, and the cancellation would fail — which is exactly what happens per sample (Section 4.2). The gradient direction is the one place where the gate does correlate with the probe, since it closes precisely when $\boldsymbol{u}$ aligns with $\widehat{\boldsymbol{g}}$; that correlation, together with the radial feed above, is the entire content of the residual $r(\eta)$ in Theorem C.20. This is why the criticality scalar is $\tau_\perp$ rather than the trace or the drift trace: $\tau_\perp$ is the trace of exactly the part of the spectrum on which clipping is only a time change.

The exact operator closes on a d-dimensional system when the gradient direction is a Hessian eigenvector, which by Corollary C.11 is exactly when $\boldsymbol{K}$ is symmetric. What that premise buys is not correctness but *reduction*: $\mathcal{T}_\eta$ and the boundary $\rho(\mathcal{T}_\eta) = 1$ are known exactly for every $\widehat{\boldsymbol{g}}$, and alignment merely collapses a spectral condition on an operator to a root of a scalar equation.

Under radial alignment ($\widehat{\boldsymbol{g}}$ an eigenvector of $\boldsymbol{H} \succeq 0$) the operator boundary $\rho(\mathcal{T}_\eta) = 1$ collapses to a scalar root, and that root brackets the critical step between reciprocals of the transverse trace $\tau_\perp = \operatorname{Tr} \boldsymbol{H} - \widehat{\boldsymbol{g}}^\top \boldsymbol{H} \widehat{\boldsymbol{g}}$, with the bracket closing at high effective rank and expressible in $\operatorname{Tr}(\boldsymbol{H})$ and λ_1 alone (Appendix C.12). Two features of that linearized threshold matter here. *The clip threshold does not appear in it*: all residual gate dependence sits in a single scalar attached to the radial mode, of relative size $O(\lambda_r/\tau_\perp)$, measured at 3%–7%. And the drift trace is $\operatorname{Tr}(\boldsymbol{K}) = p(\theta\lambda_r + \tau_\perp) \approx p \cdot 2/\eta$ at criticality, so a drift trace comfortably below $2/\eta$ does not indicate a margin created by clipping. Section 4.1.2 shows that this C-free edge is not the one a clipped run sits at, and what replaces it.

Off alignment. Assumption C.19 fails at every checkpoint (relative residual 0.669–0.911), so the scalar roots of Theorem C.20 are approximations. They survive anyway: the alignment residual enters the boundary only at second order, one coordinate carries the whole effect, and retaining it gives a corrected root that is exact under a Krylov-plane premise. Since Section 4.1.2 shows that this linearized boundary is not the one the trajectories obey, we defer the analysis to Appendix C.13 and use the alignment-free operator boundary $\rho(\mathcal{T}_\eta) = 1$ wherever a linearized threshold is needed.

C.10 Exact gated moments

Proposition C.16 (Exact gated second and fourth moments; batch)**.** *With $a(\boldsymbol{u}) = \mathbf{1}\{|\boldsymbol{u}^\top \boldsymbol{g}| < C\}$ we have, exactly,*

$$\boldsymbol{\Sigma}_a := \mathbb{E}[a(\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top] = m_2 \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + p\boldsymbol{Q}, \qquad \boldsymbol{K} = \boldsymbol{\Sigma}_a \boldsymbol{H}, \qquad \operatorname{Tr}(\boldsymbol{K}) = m_2\, \widehat{\boldsymbol{g}}^\top \boldsymbol{H} \widehat{\boldsymbol{g}} + p\, \operatorname{Tr}(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}), \tag{76}$$

and, for any $\boldsymbol{W}$, writing $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}$, $m = \widehat{\boldsymbol{g}}^\top \boldsymbol{M} \widehat{\boldsymbol{g}}$, $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{M}\widehat{\boldsymbol{g}}$, $\boldsymbol{M}_\perp = \boldsymbol{Q}\boldsymbol{M}\boldsymbol{Q}$,

$$\mathcal{Q}(\boldsymbol{W}) = \big(m_4 m + m_2 \operatorname{Tr} \boldsymbol{M}_\perp\big)\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + 2m_2\big(\widehat{\boldsymbol{g}}\boldsymbol{b}^\top + \boldsymbol{b}\widehat{\boldsymbol{g}}^\top\big) + m_2\, m\, \boldsymbol{Q} + p\big(2\boldsymbol{M}_\perp + \operatorname{Tr}(\boldsymbol{M}_\perp)\boldsymbol{Q}\big). \tag{77}$$

As $\kappa \to \infty$, $(p, m_2, m_4) \to (1, 1, 3)$ and (77) reduces to $2\boldsymbol{M} + \operatorname{Tr}(\boldsymbol{M})\boldsymbol{I}$.

Corollary C.17 (The clipping-aware trace never exceeds the threshold; batch)**.** *The exact stationary condition (99) of Theorem 4.3 is*

$$\frac{\eta\, \mu_2}{2\, p} \sum_j \frac{\lambda_j}{1 - \beta\lambda_j} = 1, \qquad \beta = \frac{\eta(\mu_4 - \mu_2)}{2p} \in (0, \eta),$$

on the admissible range $\beta\lambda_1 < 1$; the root selected in Theorem 4.3 lies in this range. Since $\min(\zeta^2, \kappa^2)$ and ζ^2 are both nondecreasing functions of ζ^2, Chebyshev's association inequality gives $\mu_4 \geq \mu_2$, so $\beta \geq 0$ and $\sum_j \lambda_j/(1 - \beta\lambda_j) \geq \operatorname{Tr}(\boldsymbol{H})$. Hence at any stationary state

$$\eta\gamma \operatorname{Tr}(\boldsymbol{H}) \;\leq\; \eta\tau \;=\; 2.$$

Separately, and for a different reason, $\gamma \operatorname{Tr}(\boldsymbol{H}) \leq \operatorname{Tr}(\boldsymbol{H})$, because $\gamma < 1$ for every finite C. The two inequalities do not compose: nothing here implies $2/\eta \leq \operatorname{Tr}(\boldsymbol{H})$. Nor does the condition $\eta > \eta_0^\star$ of Theorem 4.3: below the pole, $\sum_j \lambda_j/(1 - \eta\lambda_j)$ strictly exceeds $\operatorname{Tr}(\boldsymbol{H})$, so a nonzero stationary amplitude can coexist with $\eta \operatorname{Tr}(\boldsymbol{H}) < 2$. The raw trace exceeds $2/\eta$ only up to the same inverse-effective-rank correction of (23).

C.11 Exact pair-gate slab moments

Proposition C.18 (Exact pair-gate slab moments; per-sample)**.** *Let $\rho = \widehat{\boldsymbol{g}}_i^\top \widehat{\boldsymbol{g}}_j$ with $|\rho| < 1$, $s = \sqrt{1 - \rho^2}$, and take the orthonormal plane basis $\boldsymbol{e}_1 = \widehat{\boldsymbol{g}}_i$, $\boldsymbol{e}_2 = (\widehat{\boldsymbol{g}}_j - \rho\widehat{\boldsymbol{g}}_i)/s$ of $\mathcal{P}_{ij} = \operatorname{span}\{\widehat{\boldsymbol{g}}_i, \widehat{\boldsymbol{g}}_j\}$. Writing $\boldsymbol{x} = (x_1, x_2)$ for the standard normal coordinates in that basis, the joint gate region is $R = \{|x_1| < \kappa_i\} \cap \{|\rho x_1 + s x_2| < \kappa_j\}$ and, for $a + b \leq 4$,*

$$\mathbb{E}\big[a_i a_j\, x_1^a x_2^b\big] = \int_{-\kappa_i}^{\kappa_i} x_1^a \phi(x_1)\, I_b\big(\ell(x_1), \upsilon(x_1)\big)\, dx_1, \qquad \ell, \upsilon = \frac{\mp\kappa_j - \rho x_1}{s}, \tag{78}$$

where $I_b(\ell, \upsilon) = \int_\ell^\upsilon x^b \phi(x)\, dx$ obeys $I_0 = \Phi(\upsilon) - \Phi(\ell)$, $I_1 = \phi(\ell) - \phi(\upsilon)$ and $I_b = (b-1)I_{b-2} + \ell^{b-1}\phi(\ell) - \upsilon^{b-1}\phi(\upsilon)$. Every pair-gate moment is therefore one smooth one-dimensional quadrature. The collinear case $|\rho| = 1$ reduces to the single slab $\min(\kappa_i, \kappa_j)$ with x_2 free.

C.12 THE LINEARIZED BATCH-ZO THRESHOLD

Assumption C.19 (Radial alignment)**.** *$\boldsymbol{H} \succeq 0$ with eigenpairs $(\lambda_j, \boldsymbol{v}_j)$, and $\widehat{\boldsymbol{g}} = \boldsymbol{v}_r$ for some index r. Write $\tau_\perp = \sum_{j\neq r} \lambda_j = \operatorname{Tr} \boldsymbol{H} - \widehat{\boldsymbol{g}}^\top \boldsymbol{H}\widehat{\boldsymbol{g}}$ and $\lambda_{\max}^\perp = \max_{j\neq r} \lambda_j$.*

Theorem C.20 (Exact nonstationary-point mean-square threshold for clipped ZO)**.** *Under Assumption C.19, suppose in addition that $\lambda_r > 0$ and $\tau_\perp > 0$. On $\operatorname{range}(\boldsymbol{H})$, the covariance recursion restricted to matrices diagonal in $\{\boldsymbol{v}_j\}$ is exact and reads*

$$w_r^{t+1} = w_r^t\big(1 - 2\eta m_2 \lambda_r\big) + \eta^2\big[m_4 \lambda_r^2 w_r^t + m_2 T_\perp^t\big], \tag{79}$$

$$w_j^{t+1} = w_j^t\big(1 - 2\eta p \lambda_j\big) + \eta^2\big[2p\lambda_j^2 w_j^t + m_2 \lambda_r^2 w_r^t + p T_\perp^t\big], \quad j \neq r, \tag{80}$$

with $T_\perp^t = \sum_{j\neq r} \lambda_j^2 w_j^t$. The spectral radius of the full covariance operator on $\operatorname{range}(\boldsymbol{H})$ is attained on this diagonal subspace. Consequently, the mean-square critical step $\eta_{\mathrm{ms}}^\star$ on $\operatorname{range}(\boldsymbol{H})$ is the smallest positive root of

$$\boxed{\Xi(\eta) := \frac{\eta}{2}\Big(1 + r(\eta)\Big) \sum_{j\neq r} \frac{\lambda_j}{1 - \eta\lambda_j} = 1, \qquad r(\eta) = \frac{\eta\, m_2^2\, \lambda_r}{p\,\big(2m_2 - \eta m_4 \lambda_r\big)},} \tag{81}$$

on the admissible range $\eta\lambda_j < 1$ for $j \neq r$ and $\eta m_4 \lambda_r < 2m_2$, on which Ξ is strictly increasing.

Equation (81) is exact — no closure constant, no operator envelope — and its structure is the key point: *the gate rate p does not appear in the transverse sum*. All residual gate dependence sits in the scalar $r(\eta)$ attached to the radial mode. Two limits identify it with known objects: as $\kappa \to \infty$ it is algebraically equivalent to the ungated boundary of Song et al. (2026), and as $\kappa \to 0$ it becomes the ungated boundary of the spectrum with the gradient direction *deleted*, the ZO counterpart of the FO gradient-direction removal in Theorem 3.1 (Corollaries C.26 and C.27).

The boundary (81) brackets the critical step between two reciprocals of the gate-free transverse trace, and because $\boldsymbol{H} \succeq 0$ the gradient direction enters that bracket only through two inequalities, the band can be stated in the raw spectrum alone.

Theorem 4.1 states the resulting trace band in the main text. Write $\bar{c} = 1 + r(2/\tau_\perp)$. Under Assumption C.19 one also has $\tau_\perp = \operatorname{Tr} \boldsymbol{H} - \widehat{\boldsymbol{g}}^\top \boldsymbol{H}\widehat{\boldsymbol{g}} \in [\operatorname{Tr} \boldsymbol{H} - \lambda_1, \operatorname{Tr} \boldsymbol{H}]$ and $\lambda_{\max}^\perp \leq \lambda_1$, so (16) implies

$$\frac{2}{\bar{c}\,\operatorname{Tr}(\boldsymbol{H}) + 2\lambda_1} \;\leq\; \eta_{\mathrm{ms}}^\star \;\leq\; \frac{2}{\operatorname{Tr}(\boldsymbol{H}) - \lambda_1}, \tag{82}$$

equivalently, a checkpoint sitting at its own mean-square edge satisfies

$$\frac{2/\eta - 2\lambda_1}{\bar{c}} \;\leq\; \operatorname{Tr}(\boldsymbol{H}) \;\leq\; \frac{2}{\eta} + \lambda_1. \tag{83}$$

This is the ZO counterpart of Theorem 3.1. Both bracket the critical step between reciprocals of a computable scalar, and in both the width of the bracket is controlled by a spectral property that the runs supply: the top gap $\lambda_1 - \lambda_2$ there, the inverse effective rank $\lambda_1/\operatorname{Tr}(\boldsymbol{H})$ here. The relative width of (82) is $3\lambda_1/\operatorname{Tr}(\boldsymbol{H}) + O((\lambda_1/\operatorname{Tr}\boldsymbol{H})^2)$ when $\bar{c} \approx 1$; at the late checkpoints $\lambda_1/\operatorname{Tr}(\boldsymbol{H}) = 0.054$–$0.062$ and $\bar{c} = 1.001$–1.008, so the band is 17%–21% wide, which is wider than the $+3.05\%$–$+7.04\%$ by which $2/\tau_\perp$ misses the alignment-free operator boundary of Appendix C.13. The band is therefore not tighter than the approximation it rests on.

Equation (83) needs no gradient direction and no eigenvector, only the Hutchinson trace and one power iteration, both of which are logged online during training; and *neither side of it contains the clip threshold C*. It is therefore directly falsifiable, and we report in Section 5 that it *is* falsified as a description of training trajectories: at $\eta = 0.001$ and clip thresholds spanning an order of magnitude in gate rate, $\operatorname{Tr}(\boldsymbol{H})$ peaks at 1.00–4.16 times $2/\eta$, monotonically in the strength of the clip. The band is not wrong; it bounds the object it is about, which is the growth rate of an *infinitesimal* perturbation in the frozen local model at a checkpoint. Section 4.1.2 shows that this is not the object edge of stability is about, and derives the nonlinear balance tracked by the trajectories. At high effective rank the bracket closes on $2/\tau_\perp$, so $\tau_\perp$ — and not the drift trace — is the batch criticality scalar, up to the computable factor $\bar{c}$; Corollary C.4 states that consequence for $\tau_\perp$ itself, and $\operatorname{Tr}(\boldsymbol{H}) - \tau_\perp = \widehat{\boldsymbol{g}}^\top \boldsymbol{H}\widehat{\boldsymbol{g}}$ is

0.8%–1.5% of $\mathrm{Tr}(\boldsymbol{H})$ at the measured checkpoints, five to seven times below the stochastic-Lanczos standard error on $\mathrm{Tr}(\boldsymbol{H})$ itself, so the two are not separated by our measurements.

Under alignment, at criticality the drift trace is $\mathrm{Tr}(\boldsymbol{K}) = p(\theta\lambda_r + \tau_\perp) \approx p \cdot 2/\eta$ whenever $\lambda_r \ll \tau_\perp$ (Corollary C.25): it undershoots $2/\eta$ by approximately the gate rate. A drift trace comfortably below $2/\eta$ therefore need not indicate a margin created by clipping. The trace identity itself is alignment-free; interpreting $\tau_\perp$ as the boundary is conditional on the scalar reduction or must be checked through $\rho(\mathcal{T}_\eta)$.

C.13 Off alignment: why the scalar survives, and its correction

Assumption C.19 fails at every checkpoint (relative residual 0.669–0.911), so Theorem C.20 does not apply to the measured runs. Two facts, one negative and one constructive, settle what happens instead. Write

$$\boldsymbol{h}_\perp = \boldsymbol{Q}\boldsymbol{H}\widehat{\boldsymbol{g}}, \quad \xi = \|\boldsymbol{h}_\perp\|, \quad \widehat{\boldsymbol{h}} = \boldsymbol{h}_\perp/\xi, \quad \boldsymbol{H} = \boldsymbol{H}_{\mathrm{a}} + \boldsymbol{E}, \quad \boldsymbol{H}_{\mathrm{a}} = \lambda_r \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + \boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}, \quad \boldsymbol{E} = \widehat{\boldsymbol{g}}\boldsymbol{h}_\perp^\top + \boldsymbol{h}_\perp\widehat{\boldsymbol{g}}^\top,$$

so ξ is exactly the nonnormality scale of Corollary C.11 and the alignment residual is $\xi/\|\boldsymbol{H}\widehat{\boldsymbol{g}}\|$.

Proposition C.21 (First-order alignment insensitivity)**.** *Let $\mathcal{D} = \{\boldsymbol{W} = \boldsymbol{W}^\top : \boldsymbol{Q}\boldsymbol{W}\widehat{\boldsymbol{g}} = 0\}$ be the block-diagonal subspace and $\mathcal{D}^\perp$ its orthogonal complement in the symmetric matrices. Then $\mathcal{T}_\eta[\boldsymbol{H}_{\mathrm{a}}]$ preserves both $\mathcal{D}$ and $\mathcal{D}^\perp$, while the part of $\mathcal{T}_\eta$ linear in $\boldsymbol{E}$ maps $\mathcal{D} \to \mathcal{D}^\perp$ and $\mathcal{D}^\perp \to \mathcal{D}$. Assume further that the eigenvalue of $\mathcal{T}_\eta[\boldsymbol{H}_{\mathrm{a}}]$ attaining $\rho(\mathcal{T}_\eta[\boldsymbol{H}_{\mathrm{a}}])$ is algebraically simple and isolated, with separation Γ from the rest of the spectrum. Then the first-order perturbation of that eigenvalue vanishes and*

$$\rho(\mathcal{T}_\eta[\boldsymbol{H}]) = \rho(\mathcal{T}_\eta[\boldsymbol{H}_{\mathrm{a}}]) + O(\xi^2/\Gamma) \qquad \text{for } \xi \text{ small compared with } \Gamma. \tag{84}$$

Simplicity is not free notation. Analytic perturbation of a spectral radius is a statement about one eigenvalue, and a repeated Perron eigenvalue can split at $O(\xi)$ under a perturbation that annihilates the diagonal of the degenerate block, which is what the $\mathcal{D} \leftrightarrow \mathcal{D}^\perp$ structure gives. In our setting $\mathcal{T}_\eta[\boldsymbol{H}_{\mathrm{a}}]$ restricted to $\mathcal{D}$ is a positive operator whose Perron root is simple whenever the transverse spectrum of $\boldsymbol{H}$ has no repeated top eigenvalue; we check this numerically at every checkpoint where (84) is used, and the residual$^{1.94}$ scaling reported below is the corresponding empirical confirmation.

This is the missing explanation for an otherwise puzzling observation: the aligned root is accurate at residuals as large as 0.9 not by accident but because the residual enters the boundary only at second order. It also says where the whole $O(\xi^2)$ effect lives: in the round trip $\mathcal{D} \to \mathcal{D}^\perp \to \mathcal{D}$ through the cross block $\boldsymbol{c} = \boldsymbol{Q}\boldsymbol{W}\widehat{\boldsymbol{g}}$. At leading order that block is generated along $\widehat{\boldsymbol{h}}$ alone — from the proof, the $O(\boldsymbol{E})$ image of a block-diagonal $\boldsymbol{W}$ is spanned by $\widehat{\boldsymbol{g}}\boldsymbol{b}^\top + \boldsymbol{b}\widehat{\boldsymbol{g}}^\top$ with $\boldsymbol{b} \propto \boldsymbol{h}_\perp$ — so retaining the single coordinate $y = \widehat{\boldsymbol{h}}^\top\boldsymbol{W}\widehat{\boldsymbol{g}}$ captures the correction, and it does so exactly when $\mathcal{P}$ is $\boldsymbol{H}$-invariant. That replaces radial alignment by a strictly weaker premise.

The correction this localization implies is explicit. Retaining y alongside the radial coordinate closes the recursion on the Krylov plane $\mathcal{P} = \mathrm{span}\{\widehat{\boldsymbol{g}}, \boldsymbol{H}\widehat{\boldsymbol{g}}\}$ and yields a corrected root $\Xi_{\mathrm{c}}(\eta) = 1$ of the same shape as (81): the transverse resolvent sum still carries no gate rate, and all gate dependence sits in one scalar, now attached to the plane rather than to the radial mode alone. The cross block $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{M}\widehat{\boldsymbol{g}}$, which Section C.9 identified as the boundary information no diagonal functional retains, is thereby kept exactly along its leading direction $\widehat{\boldsymbol{h}}$. The premise is one Krylov step weaker than radial alignment — it asks that $\boldsymbol{H}^2\widehat{\boldsymbol{g}}$ lie in the plane spanned by $\widehat{\boldsymbol{g}}$ and $\boldsymbol{H}\widehat{\boldsymbol{g}}$, rather than that $\boldsymbol{H}\widehat{\boldsymbol{g}}$ be parallel to $\widehat{\boldsymbol{g}}$ — and the root returns to (81) at $\xi = 0$. Appendix C.14 states the assumption, the closed recursion and the root, and reports the synthetic validation.

Two numerical facts fix the status of the three scalars in play. On synthetic spectra where Assumption C.22 holds by construction, $\eta_{\mathrm{c}}^\star$ matches the exact operator boundary to 7×10^{-13} relative while the aligned root errs by up to 4.5% and the trace-only threshold by up to 40.8%; on the same family the aligned error grows as residual$^{1.94}$, confirming (84) numerically. At the 24 trained checkpoints, where the second Krylov residual is 0.223–0.524 and so the corrected root is itself only an approximation, the relative errors against the alignment-free operator boundary are $+3.05\%$ to $+7.04\%$ for $2/\tau_\perp$, $+0.042\%$ to $+0.440\%$ for the aligned root, and $+0.036\%$ to $+0.284\%$ for the corrected root, which is closer at all 24 points (Section 5). This is a comparison of boundary values rather than of binary labels.

C.14 THE KRYLOV-PLANE CORRECTED BATCH-ZO EDGE

Proposition C.21 localizes the whole $O(\xi^2)$ alignment effect in the single coordinate $y = \widehat{\boldsymbol{h}}^\top \boldsymbol{W}\widehat{\boldsymbol{g}}$. Retaining it gives the corrected boundary quoted in Appendix C.13.

Assumption C.22 (Krylov-plane alignment)**.** *$\boldsymbol{H} \succeq 0$ and the plane $\mathcal{P} = \mathrm{span}\{\widehat{\boldsymbol{g}}, \boldsymbol{H}\widehat{\boldsymbol{g}}\}$ is $\boldsymbol{H}$-invariant. Write $\lambda_h = \widehat{\boldsymbol{h}}^\top \boldsymbol{H}\widehat{\boldsymbol{h}}$, the counterpart of λ_r on $\widehat{\boldsymbol{h}}$, and let $\{(\lambda_j, \boldsymbol{v}_j)\}$ be the eigenpairs of the compression of $\boldsymbol{H}$ to $\mathcal{P}^\perp$.*

Assumption C.19 is the special case $\xi = 0$. The premise is one Krylov step weaker: instead of requiring $\boldsymbol{H}\widehat{\boldsymbol{g}}$ to be parallel to $\widehat{\boldsymbol{g}}$, it requires $\boldsymbol{H}^2\widehat{\boldsymbol{g}}$ to lie in the plane those two vectors span. We call the relative size of the part of $\boldsymbol{H}\widehat{\boldsymbol{h}}$ leaving $\mathcal{P}$ the *second Krylov residual*; it plays for Theorem C.23 the role the alignment residual plays for Theorem C.20.

Theorem C.23 (Krylov-plane corrected ZO edge)**.** *Under Assumption C.22 the covariance recursion closes exactly on the coordinates $w = \widehat{\boldsymbol{g}}^\top \boldsymbol{W}\widehat{\boldsymbol{g}}$, $y = \widehat{\boldsymbol{h}}^\top \boldsymbol{W}\widehat{\boldsymbol{g}}$, $\omega = \widehat{\boldsymbol{h}}^\top \boldsymbol{W}\widehat{\boldsymbol{h}}$ and $w_j = \boldsymbol{v}_j^\top \boldsymbol{W}\boldsymbol{v}_j$. With*

$$\begin{aligned} m &= \lambda_r^2 w + 2\lambda_r \xi y + \xi^2 \omega, \qquad & P &= \xi^2 w + 2\lambda_h \xi y + \lambda_h^2 \omega, \quad T_\perp = \textstyle\sum_j \lambda_j^2 w_j, \\ b &= (\lambda_r w + \xi y)\xi + \lambda_h(\lambda_r y + \xi \omega), \end{aligned} \tag{85}$$

the recursion reads

$$w^+ = w - 2\eta m_2(\lambda_r w + \xi y) + \eta^2\big[m_4 m + m_2(P + T_\perp)\big], \tag{86}$$

$$y^+ = y - \eta\big[p\xi w + (p\lambda_h + m_2\lambda_r)y + m_2\xi\omega\big] + 2\eta^2 m_2 b,$$

$$\omega^+ = \omega - 2\eta p(\xi y + \lambda_h \omega) + \eta^2\big[m_2 m + 3pP + pT_\perp\big], \tag{87}$$

$$w_j^+ = w_j - 2\eta p\lambda_j w_j + \eta^2\big[m_2 m + p(P + T_\perp) + 2p\lambda_j^2 w_j\big]. \tag{88}$$

The mean-square critical step is the smallest positive root of

$$\boxed{\Xi_{\mathrm{c}}(\eta) := \frac{\eta}{2}\big(1 + r_{\mathrm{c}}(\eta)\big)\sum_j \frac{\lambda_j}{1 - \eta\lambda_j} = 1, \qquad r_{\mathrm{c}}(\eta) = \frac{m_2\widetilde{m} + p\widetilde{P}}{p},} \tag{89}$$

where $(\widetilde{w}, \widetilde{y}, \widetilde{\omega})$ solves the 3×3 system obtained by setting $T_\perp = 1$ in the stationary form of (86)–(87), and $\widetilde{m}, \widetilde{P}$ are (85) evaluated there. When $\xi = 0$ the plane splits, y decouples, ω rejoins the tail sum and (89) becomes (81).

Equation (89) has exactly the shape of (81): the transverse resolvent sum still carries no gate rate, and all gate dependence sits in one scalar attached to the plane rather than to the radial mode alone. What changed is the size of the premise. The cross block $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{M}\widehat{\boldsymbol{g}}$, which Section C.9 identified as the boundary information no diagonal functional retains, is now retained exactly along its leading direction $\widehat{\boldsymbol{h}}$.

We write $\eta_{\mathrm{c}}^\star$ for the smallest positive root of (89).

C.15 NONSYMMETRIC FO PER-SAMPLE STABILITY

The per-sample Jacobian $\boldsymbol{A} - \boldsymbol{R} = (\boldsymbol{A} - \boldsymbol{R}_s) - \boldsymbol{R}_a$ need not be symmetric, so neither its largest real part nor its largest eigenvalue modulus alone determines the stability edge. For an eigenvalue λ with $\Re\lambda > 0$, the update factor $1 - s\lambda$ is contractive exactly when $s < 2\Re\lambda/|\lambda|^2$. Thus the definition in Section 2 gives $\eta_{\mathrm{ps}}^\star = \eta^\star(\boldsymbol{A} - \boldsymbol{R})$. For a positive real spectrum this reduces to the familiar threshold $2/\lambda_1(\boldsymbol{A} - \boldsymbol{R})$.

Theorem C.24 (Gap-free localisation of the clipped spectrum)**.** *Let $\boldsymbol{A}$, $\boldsymbol{R}_s$, and $\boldsymbol{R}_a$ be as in Eq. (11), and write $S := \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$. Then every $\lambda \in \mathrm{spec}(\boldsymbol{A} - \boldsymbol{R})$ satisfies*

$$\lambda_{\min}(\boldsymbol{A} - \boldsymbol{R}_s) \;\le\; \Re\lambda \;\le\; \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s), \qquad |\Im\lambda| \;\le\; \|\boldsymbol{R}_a\|_2. \tag{90}$$

Consequently, for every eigenvalue with $\Re\lambda > 0$,

$$\frac{|\lambda|^2}{\Re\lambda} = \Re\lambda + \frac{(\Im\lambda)^2}{\Re\lambda} \;\le\; S + \frac{\|\boldsymbol{R}_a\|_2^2}{\Re\lambda}. \tag{91}$$

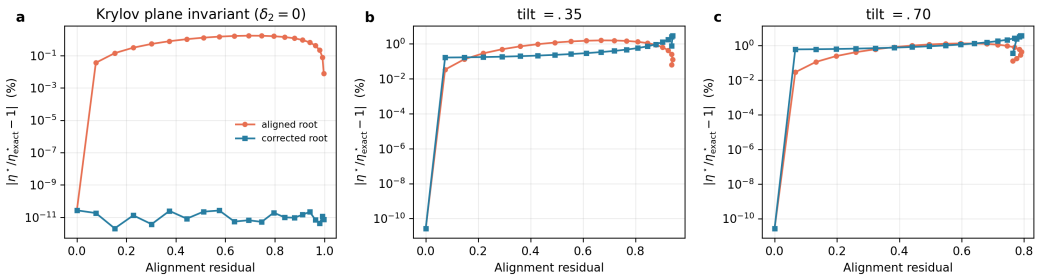

Figure 4: Synthetic validation of Theorem C.23 against the exact operator boundary, on a 16-dimensional spiked spectrum at $\kappa = 1$. **(a)** When Assumption C.22 holds, the corrected root is exact to 7×10^{-13} relative, ten orders below the aligned root, whose error grows as residual$^{1.94}$ — the quadratic scale Proposition C.21 predicts. **(b,c)** Tilting the gradient out of the invariant plane makes the second Krylov residual δ_2 nonzero and the correction loses its guarantee. Over the whole tilted family its median error is 2–4 times smaller than the aligned root's, but it is closer at only 56% of points, and the advantage is concentrated at moderate δ_2: better at 90% of points with $\delta_2 \in [0.1, 0.2)$ and at 31% once $\delta_2 > 0.3$. Note the logarithmic vertical axis.

If $\lambda^\star$ attains the minimum defining $\eta^\star_{\text{ps}}$, then

$$\frac{2}{\eta^\star_{\text{ps}}} - S \;\le\; \frac{\|\boldsymbol{R}_a\|_2^2}{\Re\lambda^\star}. \tag{92}$$

The theorem uses the numerical range rather than a perturbation expansion and therefore requires no gap between the two largest eigenvalues. It gives a one-sided quadratic bound. Proposition C.28 supplies the lower side using separation of the leading eigenvalue cluster from the rest of the spectrum.

Corollary C.25 (Batch clipping: the drift trace is not the criticality scalar)**.** *Under the hypotheses of Theorem 4.1, at criticality the drift trace satisfies*

$$\mathrm{Tr}(\boldsymbol{K}) = m_2\lambda_r + p\tau_\perp \;=\; p\left(\theta\lambda_r + \tau_\perp\right) \;\approx\; p\cdot\frac{2}{\eta} \qquad \textit{whenever } \lambda_r \ll \tau_\perp.$$

The gate-weighted effective trace therefore undershoots *$2/\eta$ by approximately the gate rate p, and it approaches $2/\eta$ only in the weak-gating limit $p \to 1$, where it coincides with the raw trace anyway.*

Corollary C.26 (Ungated limit: Song et al.)**.** *As $\kappa \to \infty$, $(p, m_2, m_4) \to (1, 1, 3)$ and $r(\eta) = \eta\lambda_r/(2 - 3\eta\lambda_r)$, in which case (81) is algebraically equivalent to $\sum_j \frac{\eta\lambda_j}{2(1-\eta\lambda_j)} = 1$, the ungated Gaussian ZO boundary of Song et al. (2026).*

Corollary C.27 (Strong-gating limit: radial deletion)**.** *As $\kappa \to 0$, $r(\eta) = \eta\lambda_r\kappa^2/6 + O(\kappa^4) \to 0$ and (81) converges to $\frac{\eta}{2}\sum_{j\neq r}\frac{\lambda_j}{1-\eta\lambda_j} = 1$: the ungated boundary of the spectrum with the gradient direction* deleted*. This is the ZO counterpart of the FO gradient-direction removal in Theorem 3.1.*

Proposition C.28 (Cluster refinement and the lower side)**.** *Let $S = \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$ and let $\boldsymbol{P}$ be the spectral projector of $\boldsymbol{A} - \boldsymbol{R}_s$ onto its eigenvalues in $[\lambda_1(\boldsymbol{A} - \boldsymbol{R}_s) - \epsilon, \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)]$ and let Γ_c be the separation of that cluster from the rest of* $\mathrm{spec}(\boldsymbol{A} - \boldsymbol{R}_s)$*. If $4\|\boldsymbol{R}_a\|_2 \le \Gamma_c$, then $\boldsymbol{A} - \boldsymbol{R}$ has exactly* $\mathrm{rank}(\boldsymbol{P})$ *eigenvalues within $\Gamma_c/2$ of the cluster, and each of them satisfies*

$$\Re\lambda \;\ge\; \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s) - \epsilon - \frac{4\|\boldsymbol{R}_a\|_2^2}{\Gamma_c}, \qquad |\Im\lambda| \;\le\; \|\boldsymbol{P}\boldsymbol{R}_a\boldsymbol{P}\|_2 + \frac{5\|\boldsymbol{R}_a\|_2^2}{\Gamma_c}. \tag{93}$$

The first-order effect of $\boldsymbol{R}_a$ inside the cluster is its compression $\boldsymbol{P}\boldsymbol{R}_a\boldsymbol{P}$, which is again skew and therefore purely rotational; the real part moves only at second order, through coupling out of the cluster. Combining this lower side with Theorem C.24 gives Eq. (13). Constants are not optimised.

Remark C.29 (Non-aligned and per-sample cases). Without Assumption C.19 the cross terms b in (77) couple the radial mode to the transverse block, and the recursion no longer closes on diagonal covariances. Two options remain, and both now use *computed* rather than assumed constants. First, Corollary C.11 bounds the deviation from the aligned model by $2\kappa\phi(\kappa)\|\boldsymbol{Q}\boldsymbol{H}\hat{\boldsymbol{g}}\|$, which is

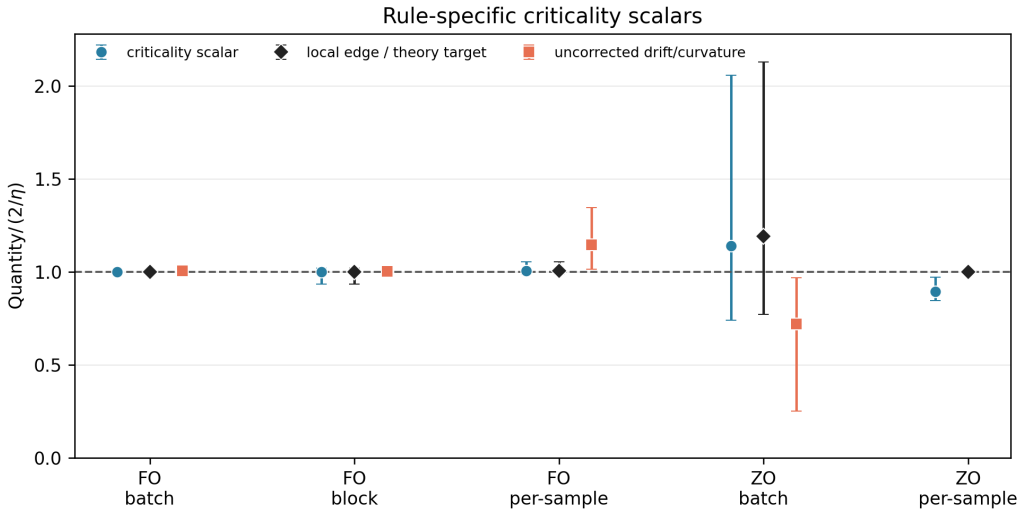

Figure 5: Five rule-specific comparisons, normalized by $2/\eta$. Markers are medians and whiskers are full late-checkpoint ranges. Blue denotes the proposed rule-specific scalar; black denotes the directly computed FO edge, the aligned-model batch-ZO root, or the per-sample co-activation target as appropriate; orange denotes the uncorrected coefficient-only FO curvature or gated batch-ZO drift. Overlap is exact up to numerical error for FO batch and block clipping, second-order accurate for FO per-sample clipping, conditional for batch ZO, and a band-level comparison for per-sample ZO.

$O((1+\kappa)e^{-\kappa^2/2})$ in the weak-gating regime of interest. Second, one may sandwich the exact operator (77) between Gaussian references $\mathcal{Q}_{\boldsymbol{G}}(\boldsymbol{W}) = 2\boldsymbol{G}\boldsymbol{W}\boldsymbol{G} + \mathrm{Tr}(\boldsymbol{G}\boldsymbol{W}\boldsymbol{G})\boldsymbol{I}$ with constants $c_\pm$ obtained in closed form from (p, m_2, m_4), recovering a band of the type used in earlier work but without an unverifiable hypothesis. Both options concern batch clipping only; the per-sample case is treated in Section 4.2.

Remark C.30 (How the per-sample moments are evaluated). For $i \neq j$ the pair $(z_i, z_j) = (\boldsymbol{u}^\top \widehat{\boldsymbol{g}}_i, \boldsymbol{u}^\top \widehat{\boldsymbol{g}}_j)$ is bivariate normal with correlation $\rho_{ij} = \widehat{\boldsymbol{g}}_i^\top \widehat{\boldsymbol{g}}_j$, and the two gates cut out the intersection of two slabs. Proposition C.18 reduces every moment $\mathbb{E}[a_i a_j\, x_1^a x_2^b]$, $a+b \leq 4$, to one Gauss–Legendre quadrature in the plane coordinate x_1, since the inner integral in x_2 is a closed-form incomplete Gaussian moment. With 192 nodes the value is converged to machine precision, matching a 384-node reference to 1.3×10^{-15}, and agrees with 4×10^6-draw Monte Carlo within 1.9 standard errors on every moment of five test triples. The co-activation matrix c_{ij} is the $a = b = 0$ member of the same family, which is why ς is exact rather than sampled. The negative consequence still stands: *no formula stated for batch clipping in Sections C.9–C.9, in particular Proposition C.16 and Theorem C.20, may be applied to per-sample runs*, since both are derived from a single shared slab; Theorem C.14 is their replacement. The cost is $O(G^2)$ quadratures for G gates, so full enumeration is used up to $G = 128$ and unbiased sampling of ordered off-diagonal pairs beyond it, the diagonal $c_{ii} = p_i$ being exact.

Remark C.31 (Bounded increments are not bounded oscillation). Clipped ZO updates satisfy $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| \leq \eta C \|\boldsymbol{u}_t\|$ regardless of curvature. This rules out an arbitrarily large *one-step* jump, but it does not imply a bounded trajectory or bounded separation of two trajectories: the bounded increments may accumulate. Local mean-square instability therefore does not by itself distinguish bounded oscillation from nonlinear escape. Establishing the former requires a separate global Lyapunov or recurrence argument, not the clipping bound alone.

C.16 Full empirical section

The following reproduces the complete experimental analysis, of which Section 5 reports only the headline findings.

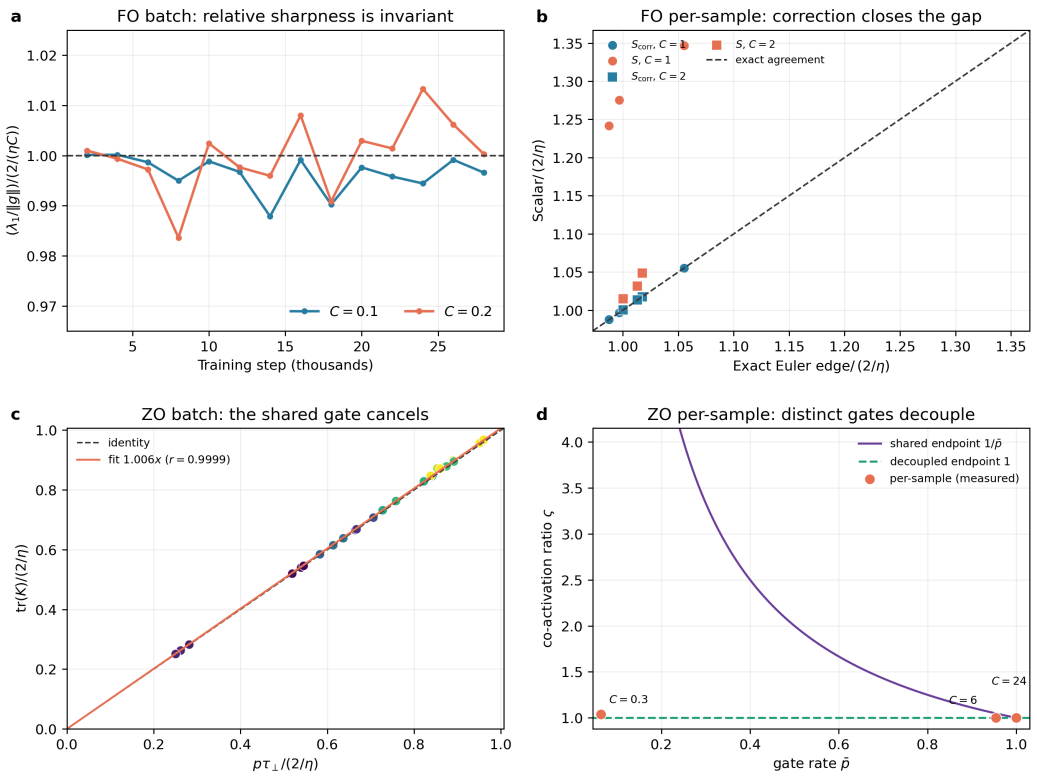

Figure 6: Four-rule diagnostic summary retained from the earlier overview. **(a)** Batch-FO relative sharpness follows its predicted invariant. **(b)** The per-sample FO correction moves $\lambda_1(\boldsymbol{A})$ to $S = \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s)$, which agrees with the computed stability threshold. **(c)** The exact batch-ZO drift trace tracks $p\tau_\perp$. **(d)** Measured per-sample gates lie near the decoupled co-activation endpoint rather than the shared-gate endpoint.

Protocol and results

Protocol. We train a 30,570-parameter two-layer convolutional network with GELU and average pooling on 1,000 CIFAR-10 training examples from four classes (Krizhevsky, 2009). The objective is full-batch mean squared error. FO uses $\eta = 0.05$; ZO uses one Gaussian two-point direction per step, $\mu = 0.001$, and $\eta \in \{0.001, 0.002\}$. We sweep global and per-sample clip thresholds and analyze the last three periodic checkpoints of every run: 36 FO (including 12 per-layer) and 72 ZO checkpoints (108 total). A further group-clipping sweep, $G \in \{1, 2, 4, 8, 20, 40, 100, 1000\}$, branches from one common well-trained checkpoint and adds 24 checkpoints; every branch uses the threshold C_G calibrated at that checkpoint to gate rate $\bar{p} = 0.6$, so the branches differ only in the number of gates. All runs use seed zero. Exact matrices are accessed through Hessian–vector products and full-reorthogonalized Lanczos or Arnoldi iterations. Appendix F gives the full grid and estimator details.

FO batch: gradient-direction removal is negligible and the coefficient does all the work. For the six active late batch checkpoints ($C = 0.1, 0.2$), the relative top gap $(\lambda_1(\boldsymbol{H}) - \lambda_2(\boldsymbol{H}))/\lambda_1(\boldsymbol{H})$ is 9.65×10^{-4}–1.11×10^{-2}, and the actual proxy–exact gap is 2.69×10^{-4}–9.76×10^{-3}. Every point satisfies Theorem 3.1; Figure 9 (left) shows the pointwise inequality. Read through Corollary C.3, this is not evidence that the direction-removal mechanism matters: it is evidence that it *cannot* matter at $k = 1$, because the top Hessian cluster is nearly degenerate, which is itself a signature of EoS training. The stabilizing effect at $k = 1$ is entirely the scalar $\min\{1, C/\|\boldsymbol{g}\|\}$, i.e. the change of invariant in Eq. (10).

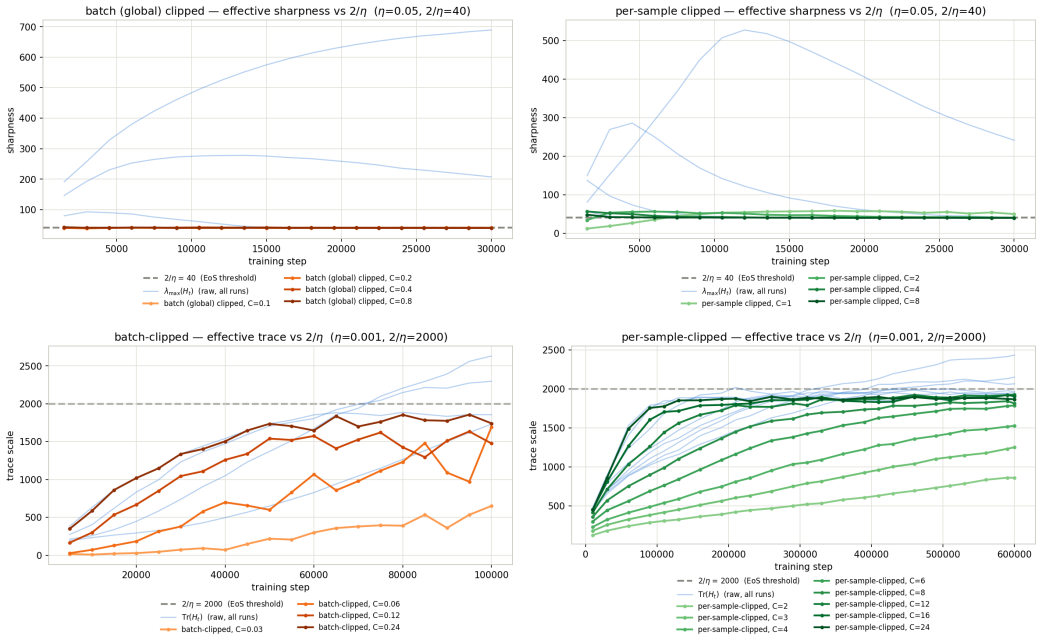

Figure 7: The empirical observation that motivates clipping-aware curvature diagnostics. Top: FO raw sharpness (light blue) substantially exceeds $2/\eta$, so raw sharpness is not the operative constraint; the coefficient-scaled curve $\min\{1, C/\|\boldsymbol{g}\|\}\lambda_1(\boldsymbol{H})$ remains near the boundary, and by Eq. (10) that condition is the relative-sharpness invariant $\lambda_1/\|\boldsymbol{g}\| = 2/(\eta C)$, which Theorem 3.1 shows to agree with the criticality scalar to within the top gap. Bottom: ZO raw trace can cross $2/\eta$, whereas the gated drift trace is controlled. For batch clipping (left), Corollary C.25 conditionally interprets the gap as the gate rate and $\tau_\perp$ as the aligned-model scalar. The per-sample panel (right) is shown only as motivation; its exact edge follows from the pair moments of Proposition C.18 and is computed in Figure 3.

The late-checkpoint comparison in Figure 1(a) and the block-level check in Figure 8 bear this out. Over the full trajectories, after the first logged 2,000-step transient, $(\lambda_1/\|\boldsymbol{g}\|)/(2/(\eta C))$ stays in $[0.988, 1.000]$ for $C = 0.1$ and $[0.984, 1.013]$ for $C = 0.2$; the medians are 0.997 and 1.001. Origin-constrained fits of λ_1 on $\|\boldsymbol{g}\|$ have slopes 398.36 and 199.91, against the predicted 400 and 200, with $R^2 = 0.9998$ and 0.9966.

For the codimension test we trained four per-layer runs, treating the two convolutional tensors and the final affine layer as $L = 3$ blocks. At the late checkpoints $C = 0.03, 0.06, 0.1, 0.2$ have respectively $k = 3, 3, 2, 1$ active blocks. Their $(\gamma_k, \text{proxy–exact gap})$ ranges are $(0.229$–$0.232, 0.0586$–$0.0638)$, $(0.148$–$0.155, 0.0117$–$0.0172)$, $(0.0118$–$0.0139, 0.00066$–$0.00218)$, and $(0.00367$–$0.00707, 0.00224$–$0.00507)$. Across all twelve checkpoints the Pearson correlation is 0.888 (Spearman 0.783), and every exact curvature also satisfies the heterogeneous-coefficient lower bound in Theorem C.8(ii). Since the block coefficients are not equal, the diagonal in Figure 8 is the common-coefficient reference from part (i), not a pointwise bound for these runs; the observed association is therefore qualitative evidence for codimension-k compression rather than a test of the equal-coefficient formula.

FO per-sample: the derived error scalings. At $C = 2$, only 3.1–6.8% of samples are clipped and $\|\boldsymbol{R}_a\|_2/S = 0.00585$–$0.00849$; the relative discrepancy $|(2/\eta^\star_{\text{ps}}) - S|/S$ is 6.46×10^{-5}–1.06×10^{-4}. At $C = 1$, where 85.0–88.1% of samples are clipped, the corresponding ranges are 0.00990–0.01044 and 5.28×10^{-5}–1.50×10^{-4}. Thus the magnitude predicted by Proposition 3.2 is reproduced. This is a scale-consistency check rather than an empirical identification of the exponent, since $\|\boldsymbol{R}_a\|_2/S$ varies over a narrow range. The simple-gap route is not merely unnecessary here, it is unavailable. Measuring $\gamma_s = \lambda_1(\boldsymbol{A} - \boldsymbol{R}_s) - \lambda_2(\boldsymbol{A} - \boldsymbol{R}_s)$ gives $4\|\boldsymbol{R}_a\|_2/\gamma_s = 2.09$–$3.53$ at $C = 1$ and 4.88–16.68 at $C = 2$, so the classical simple-eigenvalue hypothesis fails at every active checkpoint, by a

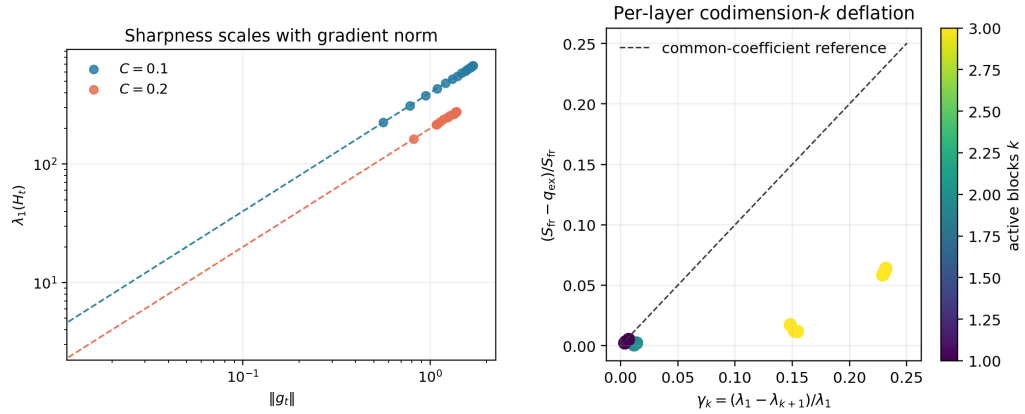

Figure 8: Complementary FO falsification checks. Left: raw sharpness is linear in gradient norm with the slope $2/(\eta C)$ predicted by the relative-sharpness invariant in Figure 7 (top left). Right: the measured per-layer proxy–exact gap grows with the codimension-k spectral gap. The dashed diagonal is the common-coefficient reference; the experiments use heterogeneous block coefficients.

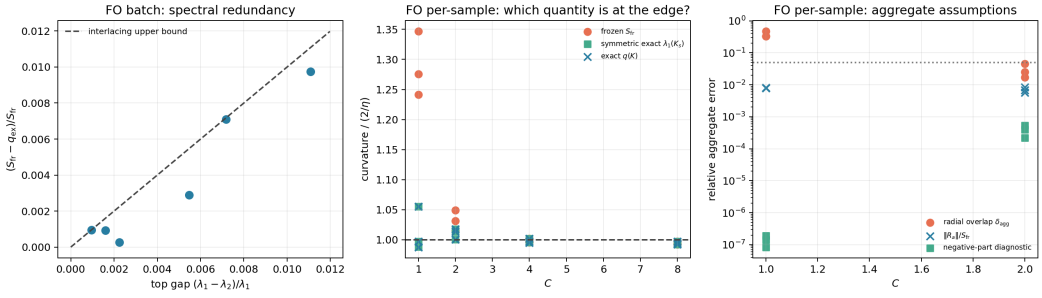

Figure 9: FO theorem checks. Left: the batch proxy–exact gap lies below the one-direction-removal bound of Theorem 3.1. Center: exact per-sample curvature remains near the edge while the uncorrected $\lambda_1(\boldsymbol{A})$ overestimates it at $C = 1$, whereas S tracks the exact edge. Right: aggregate diagnostics showing that the skew component is small.

wide margin at $C = 2$. This is the empirical content of Corollary C.3: the symmetric part inherits the near-degenerate top cluster these runs exhibit, so γ_s is small exactly where clipping is active. Because $\gamma_s < 4\|\boldsymbol{R}_a\|_2$ here, the simple-gap argument is inapplicable, whereas the gap-free localisation of (90) remains valid. The pointwise results are in Table 7. The dimensionless gap-free statement is tested directly in S units. At all six checkpoints $2/\eta^\star_{\mathrm{ps}} \le S$, so the upper-displacement bound holds with slack; the measured ratio $\Re\lambda^\star/S$ is 0.99985–0.99995. Selecting the smallest leading Ritz cluster separated by at least $4\|\boldsymbol{R}_a\|_2$ gives ranks $9, 9, 2$ at $C = 1$ and rank 2 at all $C = 2$ checkpoints. The measured $4\|\boldsymbol{R}_a\|_2/\Gamma_c = 0.127$–$0.925 < 1$, so the weaker cluster condition needed for the lower side holds everywhere even though the simple-eigenvalue condition fails.

Finally, the corrected proxy removes the radial error that is linear in $\boldsymbol{R}_s$, as predicted. Normalized by $2/\eta$, S is 0.988–1.055 at $C = 1$ and 1.000–1.018 at $C = 2$, matching the exact branch to within $1.50 \times 10^{-4} S$ at every active checkpoint. The raw 24–35% gap at $C = 1$ therefore collapses to at most 0.012%, well below 1%; all settings are reported in Table 9.

ZO: high effective rank is the regime in which aligned theory predicts leading cancellation. Across all 72 ZO checkpoints, $\mathrm{Tr}(\boldsymbol{K})/\lambda_1(\boldsymbol{K}) = 10.1$–$27.6$ (median 16.8). For the batch-clipped runs, Theorem 4.1 says that this high-effective-rank condition is what makes the transverse sum dominate, and Theorem C.20 says the transverse sum carries no gate rate. The measured gate rates in the near-edge runs are 0.988, 0.962, 0.954 and 0.9998 (Table 1), i.e. $p \approx 1$, which is precisely the

regime in which the drift trace and the gradient-orthogonal trace $\tau_\perp$ coincide; the two readings of Figure 7 are therefore *not distinguishable* on the near-edge runs alone.

The strong-gating control checkpoints do distinguish them. At batch gate rates 0.753 and 0.331 the drift trace is far below $2/\eta$, while Corollary C.25 predicts that $\tau_\perp$ is not. The factor-p attenuation behind this separation is visible directly in Figure 1(b). The per-sample controls are a different test: there Corollary C.15 predicts the opposite direction, with the drift trace approaching $(1/\varsigma)(2/\eta)$ and ς close to one when the sample gates decorrelate. The per-sample rows of Table 1 therefore cannot be read through Theorem C.20, which is derived for a single shared gate. Table 11 reports the requested batch quantities at all 24 checkpoints. At $\eta = 0.001$, $\tau_\perp/(2/\eta) = 0.741$–$1.219$ and the actual step divided by the aligned-model boundary root is 0.772–1.264; at $\eta = 0.002$ the respective ranges are 0.926–2.055 and 0.977–2.129. The paired final $C = 0.24$ values are $(0.973, 1.013)$ for $\eta = 0.001$ and $(0.926, 0.977)$ for $\eta = 0.002$. Trace estimates use 16-probe, 64-step stochastic Lanczos quadrature; their relative standard errors are 2.7–6.1%.

Crucially, these roots are not an empirical validation of the word "exact" in Theorem C.20: the theorem is conditional on radial alignment. The measured residual $\|(\boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top)\boldsymbol{H}\widehat{\boldsymbol{g}}\|/\|\boldsymbol{H}\widehat{\boldsymbol{g}}\| = 0.669$–$0.911$, so that premise fails substantially at every batch checkpoint. We therefore label the reported roots *aligned-model plug-ins*; they diagnose what the scalar equation would predict after imposing the theorem's premise, but they cannot be interpreted as exact stability boundaries of these checkpoints. In contrast, the parameter-free batch skew formula does not require this interpretation and matches the measured norm to within 9.6×10^{-6} relatively.

For per-sample clipping, the Gaussian directions used by the closure probes give $\varsigma = 1.00021$ at $C = 6$ and 1.0000002 at $C = 24$. The resulting point targets $(1/\varsigma)(2/\eta)$ are 0.99979 and 1.00000 after normalization, whereas the measured drift traces are 0.894 and 0.933. The corresponding finite-rank Gaussian-reference bands, $[0.894, 1.001]$ and $[0.933, 1.096]$, contain the targets. Thus the data support the band-level comparison but not literal equality of the trace and target at these finite checkpoints. The strongly gated $C = 0.3$ control gives $\varsigma = 1.0398$, a small but resolvable co-activation correction. Finally, common-noise nonlinear perturbation tests start two copies 10^{-4} apart and apply the same Gaussian directions for 2,048 steps. Across three replicates the final separation is amplified by 80.8–$90.6\times$ for batch $C = 0.03$, 468.8–$621.4\times$ for batch $C = 0.12$, and 5.86–$6.23\times$ for per-sample $C = 0.3$; all nine fitted second-half log slopes are positive. Thus none of the strongly gated runs is locally contracting over the tested horizon; the batch controls are included in Figure 14. The test establishes non-decay, but not bounded oscillation. We therefore extended the final batch $C = 0.03$ checkpoint to 65,536 nonlinear updates (Figure 13, left). Across three replicas the separation reaches 0.539–0.552, or 5,388–5,520 times its initial radius, and its second-half linear slope remains 1.11–1.16×10^{-5} per step. This is 0.370–0.388 times ηC *per step*, not a plateau at scale ηC. The bounded update in Remark C.31 can accumulate, and in this experiment it does.

Our previously reported fourth-moment probes are now a validation of a formula rather than a test of a hypothesis. For $\boldsymbol{W} = \boldsymbol{x}\boldsymbol{x}^\top$ and test vector $\boldsymbol{v}$, the ratio

$$r_4(\boldsymbol{x}, \boldsymbol{v}) = \frac{\mathbb{E}\left[(\boldsymbol{v}^\top \boldsymbol{u})^2 \left(n^{-1}\sum_i a_i \boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{x}\right)^2\right]}{\|\boldsymbol{v}\|^2 \|\boldsymbol{G}\boldsymbol{x}\|^2 + 2(\boldsymbol{v}^\top \boldsymbol{G}\boldsymbol{x})^2}$$

compares $\mathcal{Q}$ with the Gaussian reference $\mathcal{Q}_{\boldsymbol{G}}$; for batch clipping the numerator is available in closed form from Proposition C.16, and for per-sample clipping it is the pair-moment object of Theorem C.14, which these probes predate and which supersedes them as a test of the operator. Table 1 shows concentration around one for near-edge weak gates and broadening to 0.719–12.97 under strong clipping. For the batch rows, substituting the measured $(p, m_2, m_4, \boldsymbol{x}, \boldsymbol{v})$ into (77) predicts median $r_4 = 1.013, 1.321, 3.095$ at $C = 0.24, 0.12, 0.03$, against measured medians $1.013, 1.345, 3.340$. The median sampled-to-predicted ratios are $1.000, 1.014, 1.099$, respectively; the agreement persists into the strongly gated controls and validates the closed form within Monte Carlo error. We do not retrofit analogous per-sample predictions onto these probes; the pair moments they would need are instead used where they belong, in the operator of Theorem C.14, which is validated directly against Monte Carlo in Appendix C.16.

The corrected batch scalar and the pair-gate operator

Validating the pair moments. Proposition C.18 is implemented as a 192-node Gauss–Legendre rule in x_1 with the closed-form incomplete Gaussian moments I_b inside. Against a 384-node

Table 1: ZO closure probes and paired late-checkpoint Gaussian-reference bands. With $\boldsymbol{G} = (\boldsymbol{K} + \boldsymbol{K}^\top)/2$ and $\tau = \mathrm{Tr}(\boldsymbol{G})$ as in Eq. (64), the bands are $[\tau, \tau + 2\lambda_1(\boldsymbol{G})]/(2/\eta)$. For per-sample clipping they are reference diagnostics rather than bounds on the pair-moment operator, which Theorem C.14 now supplies directly; probe ranges are finite-sample evidence. The last column is the new quantity implied by the theory, and it is *different for the two rules*: for batch rows it is $\tau_\perp/(2/\eta)$, which Corollary C.4 predicts to be near one at criticality; for per-sample rows Theorem C.20 does not apply and the relevant quantity is the co-activation ratio ς of (58), through Corollary C.15.

Checkpoint	Gate rate	r_4 range	Paired band	New quantity
Batch $\eta = 0.001, C = 0.24$	0.988	0.917–1.164	$[0.879, 1.000]$	$\tau_\perp/(2/\eta) : 0.973$
Batch $\eta = 0.002, C = 0.24$	0.962	—	$[0.968, 1.105]$	$\tau_\perp/(2/\eta) : 0.926$
Per-sample $C = 6$	0.954	0.874–1.078	$[0.894, 1.001]$	$\varsigma : 1.00021$
Per-sample $C = 24$	0.9998	0.973–1.113	$[0.933, 1.096]$	$\varsigma : 1.000000$

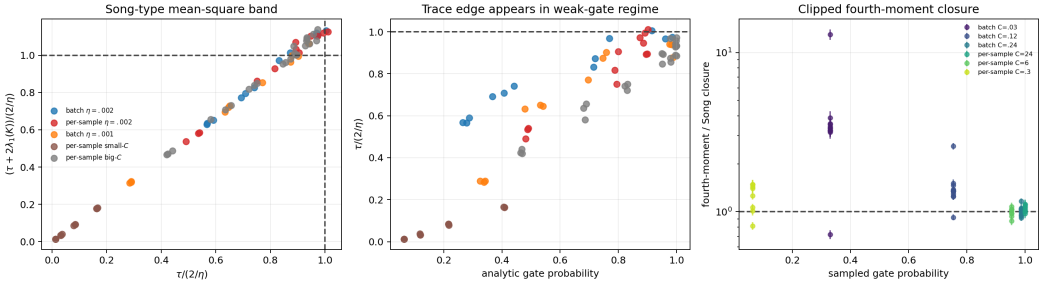

Figure 10: ZO theorem checks. Left: normalized Gaussian-reference bands. Center: the drift trace approaches $2/\eta$ as the gate becomes weak, which Corollary C.25 attributes to the factor p rather than to a change in the critical step. Right: fourth-moment ratios, now comparable against the closed form of Proposition C.16.

reference the largest absolute difference over all moments with $a + b \le 4$ is 1.3×10^{-15}, so the quadrature is at machine precision; the integrand is $\phi(x_1)$ times a smooth function on a compact interval, which is why so few nodes suffice. Against direct Monte Carlo with 4×10^6 draws on the triples $(\kappa_i, \kappa_j, \rho) \in \{(0.4, 0.9, 0), (0.7, 0.7, 0.6), (1.3, 0.5, -0.8), (2, 2, 0.95), (0.25, 3, 0.3)\}$, every moment agrees within 1.9 Monte Carlo standard errors. Comparing relative deviations would be meaningless here: the moments with $a + b$ odd vanish identically by the symmetry $(x_1, x_2) \mapsto (-x_1, -x_2)$ of the region.

Validating the assembled operator. Two checks pin Theorem C.14. With one gate the implementation must reproduce the batch operator built from (77); on random spectra, including an isotropic bulk block, the two agree to 1.1×10^{-11} on the retained block and 4.8×10^{-9} on the bulk coordinate. With four gates in general position there is no closed form to compare against, so we estimate $\mathbb{E}[\bar{a}^2(\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top]$ directly. The gap falls monotonically with the draw count — 1.4×10^{-3}, 1.1×10^{-3}, 7.5×10^{-4}, 4.0×10^{-4} relative at $N = 2, 8, 32, 128$ million — with no sign of a plateau, which is what distinguishes sampling error from a bias. The decay is somewhat slower than $N^{-1/2}$ over this range, as expected for an estimator whose integrand is quartic in a Gaussian and therefore heavy-tailed.

Locating the operator boundary. The spectral radius must be obtained by a Krylov method, not power iteration. The isotropic bulk contributes an eigenvalue $1 - 2\eta\bar{p}\lambda_{\mathrm{bulk}} + O(\eta^2)$, and with $\lambda_{\mathrm{bulk}} \approx 0.016$ and $\eta = 0.001$ that is within 3×10^{-5} of one, so a dense cluster sits just below the leading eigenvalue. Successive Rayleigh ratios then plateau far from convergence: a power iteration with a 10^{-12} relative stopping rule returned boundaries 38% too large before we replaced it with the same ARPACK path the batch controls use. We report this because the failure is silent — the iteration appears converged — and because the same near-degeneracy makes ρ itself a poor discriminator: at a late $C = 0.24$ checkpoint, ρ moves only from 0.999984 to 1.000006 as η ranges over a factor of

Table 2: Batch-ZO boundary values at all 24 late checkpoints. The reference is $\eta^\star_{\mathrm{op}}$, the bisected root of $\rho(\mathcal{T}_\eta) = 1$ for the alignment-free reduced operator; the last three columns are the relative errors of the trace-only threshold $2/\tau_\perp$, the aligned root of Theorem C.20 and the corrected root of Theorem C.23. Both premises fail here: δ_1 is the alignment residual $\|\boldsymbol{Q}\boldsymbol{H}\widehat{\boldsymbol{g}}\|/\|\boldsymbol{H}\widehat{\boldsymbol{g}}\|$ and δ_2 the second Krylov residual. The corrected root is closer at every point, and every scalar overestimates the boundary.

η	C	step	δ_1	δ_2	$10^3\eta^\star_{\mathrm{op}}$	$2/\tau_\perp$ (%)	aligned (%)	corrected (%)
0.001	0.03	90k	0.911	0.329	1.2571	+3.80	+0.112	+0.107
0.001	0.03	95k	0.882	0.406	1.2878	+4.77	+0.129	+0.124
0.001	0.03	100k	0.824	0.274	1.1085	+3.97	+0.042	+0.036
0.001	0.06	90k	0.832	0.407	0.8199	+3.65	+0.122	+0.108
0.001	0.06	95k	0.820	0.369	0.8311	+4.27	+0.065	+0.054
0.001	0.06	100k	0.789	0.348	0.7897	+3.88	+0.150	+0.126
0.001	0.12	90k	0.722	0.329	0.9216	+3.82	+0.106	+0.062
0.001	0.12	95k	0.736	0.429	0.9431	+4.32	+0.142	+0.102
0.001	0.12	100k	0.776	0.499	0.8240	+3.23	+0.135	+0.115
0.001	0.24	90k	0.804	0.458	1.1066	+5.23	+0.356	+0.258
0.001	0.24	95k	0.834	0.273	1.0840	+5.35	+0.255	+0.091
0.001	0.24	100k	0.849	0.414	0.9882	+4.05	+0.261	+0.188
0.002	0.03	70k	0.827	0.371	1.0294	+3.55	+0.116	+0.111
0.002	0.03	80k	0.832	0.284	1.0308	+4.14	+0.097	+0.089
0.002	0.03	90k	0.774	0.461	0.9365	+3.91	+0.120	+0.116
0.002	0.06	70k	0.779	0.488	1.2093	+3.22	+0.138	+0.131
0.002	0.06	80k	0.733	0.492	1.1804	+3.73	+0.199	+0.190
0.002	0.06	90k	0.669	0.431	1.0691	+3.05	+0.083	+0.076
0.002	0.12	70k	0.755	0.471	1.6288	+4.16	+0.246	+0.209
0.002	0.12	80k	0.764	0.337	1.6691	+4.82	+0.252	+0.161
0.002	0.12	90k	0.847	0.365	1.6761	+4.92	+0.259	+0.204
0.002	0.24	70k	0.735	0.223	2.0029	+7.04	+0.440	+0.093
0.002	0.24	80k	0.839	0.524	1.9705	+5.08	+0.322	+0.284
0.002	0.24	90k	0.768	0.414	2.0375	+6.01	+0.167	+0.107

two. Comparing scalars therefore means comparing their roots against the bisected $\rho(\mathcal{T}_\eta) = 1$, not comparing ρ values.

Consistency of the two roots. Theorem C.23 claims $\Xi_{\mathrm{c}} \to \Xi$ as $\xi \to 0$, with λ_h rejoining the tail. Evaluating both on a fixed spectrum while shrinking ξ confirms it and measures the rate: at $\kappa = 0.5, 1, 2$ the relative difference is 1.9×10^{-7}, 6.8×10^{-7} and 1.8×10^{-6} at $\xi = 10^{-2}$ and falls by exactly two decades per decade of ξ. That quadratic rate is (84) again, seen from the scalar side rather than from the operator.

The three batch scalars at the checkpoints. Table 2 reports, per checkpoint, the alignment residual, the second Krylov residual, and the relative error of each scalar against the bisected operator boundary on a rank-96 Lanczos reduction that additionally contains $\widehat{\boldsymbol{g}}$ and $\widehat{\boldsymbol{h}}$ exactly. The ordering is uniform: $2/\tau_\perp$ errs by $+3.05\%$ to $+7.04\%$ (median $+4.09\%$), the aligned root by $+0.042\%$ to $+0.440\%$ (median $+0.140\%$), the corrected root by $+0.036\%$ to $+0.284\%$ (median $+0.113\%$), and the corrected root is closer at all 24 points. Every scalar overestimates the boundary here. For $2/\tau_\perp$ that direction is what Theorem 4.1 gives *within the aligned model*; that it also holds against the alignment-free operator is an observation about these checkpoints, not a consequence of the corollary. The improvement from the correction is real but modest here because the second Krylov residual is 0.223–0.524: the premise of Theorem C.23 fails at these checkpoints too, one order later than the premise of Theorem C.20. The synthetic families of Figure 4 separate the two effects, since there the premise can be imposed exactly, and they also show that off the premise the correction can go the wrong way: it is closer at 56% of the tilted points overall, at 90% of those with second Krylov residual in $[0.1, 0.2)$, and at only 31% of those above 0.3. The uniform $24/24$ win at the checkpoints is thus a property of these checkpoints rather than a guarantee, and we report it as such.

Table 3: Group-clipped ZO at matched calibration. Every branch starts from one checkpoint with the threshold C_G chosen there so that $\bar{p} = 0.6$; values are means over the three late checkpoints of each branch, at which the realised gate rate has drifted and is reported. ς is exact from Proposition C.18 for $G \le 128$ and an unbiased off-diagonal estimate at $G = 1000$. The reference $\eta^{\star}_{\mathrm{op}}$ is the root of $\rho(\mathcal{T}_\eta) = 1$ for the operator of Theorem C.14; the last two columns are the relative errors of the trace-only threshold $2/\tau$ and of the co-activation surrogate $2/(\varsigma\bar{p}\tau)$ of Corollary C.15. The operator is assembled exactly only where the retained subspace can hold every gate direction, hence the omission at $G = 1000$. Values are means over the three checkpoints; the per-checkpoint spread is small, and the qualitative crossing between $G = 2$ and $G = 4$ holds at every one of them.

G	$\bar{p}$	ς	$1/\bar{p}$	$10^3\eta^{\star}_{\mathrm{op}}$	$2/\tau$ (%)	$2/(\varsigma\bar{p}\tau)$ (%)
1	0.914	1.0946	1.0946	1.0116	+2.9	+2.9
2	0.897	1.0716	1.1149	1.0556	−1.0	+3.0
4	0.782	1.0846	1.2800	1.2220	−15.3	−0.1
8	0.803	1.0349	1.2465	1.2570	−16.6	+0.4
20	0.786	1.0177	1.2733	1.2851	−20.1	−0.1
40	0.784	1.0109	1.2765	1.3269	−20.9	−0.2
100	0.779	1.0067	1.2846	1.3383	−21.6	−0.0
1000	0.797	1.0022	1.2546	—	—	—

The group sweep. As an implementation check, running (73) at $G = 1$ and at $G = n$ from the same seed reproduces the batch and per-sample branches exactly, including the logged pre- and post-clip magnitudes, while the two branches differ from each other; the group rule is therefore a genuine interpolation rather than a third rule that happens to resemble them. Thresholds are calibrated once, at the shared branch checkpoint, so that every G starts at gate rate $\bar{p} = 0.6$; the predicted co-activation ratios there are $1.667, 1.356, 1.176, 1.091, 1.040, 1.022, 1.013$ and 1.005 for $G = 1, \dots, 1000$, spanning the interval $[1, 1/\bar{p}]$ almost end to end. Gate rates drift upward during the branch runs as the gradients shrink, so the realized $\bar{p}$ at the analysis checkpoints is not exactly 0.6 and is reported alongside each ς; the comparison of ς with its own endpoint $1/\bar{p}$ uses the realized value and is unaffected. The exact operator is assembled for $G \le 128$, where the retained subspace can contain every gate direction without error; at $G = 1000$ we report the drift and the unbiased co-activation estimate only. Sampling ordered off-diagonal pairs is what makes the latter unbiased — averaging the full submatrix of a random subset of gates over-weights the diagonal $c_{ii} = p_i$ and inflates ς, which in an earlier pass produced the impossible value $\varsigma < 1$ at $G = 1000$.

Table 3 reports the outcome. Two features are worth separating. The first is that ς behaves exactly as Lemma C.6 says: at $G = 1$ the measured value equals $1/\bar{p}$ to five digits, which is the shared-gate endpoint attained rather than approached, and the normalised position $(\varsigma - 1)/(1/\bar{p} - 1)$ falls monotonically to 0.008 at $G = 1000$. The second is that the surrogate built from ς is quantitatively good away from the batch endpoint and not at it: $2/(\varsigma\bar{p}\tau)$ is within 0.4% of the exact boundary for all $G \ge 4$ — and still within 0.04% at $G = 100$, where the trace threshold is off by 21.6% — but overshoots by 3.0% at $G = 2$, exactly where $2/\tau$ is nearly exact. Both scalars come from the same isotropic model with gates independent of $\boldsymbol{u}$, so neither is guaranteed; what the exact operator adds is the location of the crossing, which no argument in Section 4.2 predicts. We note also that the two scalars err in opposite directions on either side of it, so a practitioner who picks one and stays with it is wrong by up to 22% in the step size at the far end of the sweep.

The reduction is well conditioned across the sweep: after two Gram–Schmidt passes the retained basis is orthonormal to 3.2×10^{-5}, and each gate direction is captured by that basis to within floating-point error, which is the property the plane construction of Theorem C.14 requires.

Direct tests of the proposed scalars

Figure 5 puts the five rule-specific comparisons on one scale. For all three FO rules the proposed scalar overlaps the directly computed stability edge, while the uncorrected per-sample reference $\lambda_1(\boldsymbol{A})$ has a visible first-order excess. The two ZO rows must be read differently. For batch clipping the black interval is the *aligned-model* root, not an exact edge of the measured non-aligned checkpoints; for per-sample clipping it is the isotropic co-activation target $1/\varsigma$, since the exact mean-square boundary requiring the full pair-gate fourth-moment operator has not been computed. The panel is therefore

a visual taxonomy of the five comparisons, not a claim that all five black markers have the same empirical status. In particular, the per-sample drift trace spans 0.846–0.973 over the displayed $C = 6, 24$ checkpoints while its targets are 0.99979 and 1.00000, consistent with the finite-rank bands of Table 1 but not with literal equality.

We next carried out the direct test proposed by the earlier draft rather than assuming its outcome. At all 24 batch-ZO late checkpoints we evolved three pairs of models from separation 10^{-4} for $2{,}048$ exact finite-difference updates with common Gaussian directions. All 72 second-half log slopes are positive. Checkpoint-median amplification ranges from $66.1\times$ to $5{,}759.7\times$; every gated drift trace lies below $2/\eta$ and has Spearman correlation 0.095 with amplification. The transverse scalar orders the points better (Spearman 0.751), but it does *not* separate growth at one: nine checkpoints with $\tau_\perp/(2/\eta) < 1$ still amplify by $66.1\times$–$1{,}434.8\times$. Figure 14 therefore rules out using $\tau_\perp$ as a premise-free empirical classifier for these checkpoints.

That conclusion changes once the test is aimed at the theorem's actual object. Following the positive-curvature convention of Remark C.2, at each checkpoint we formed 128-step Lanczos decompositions, retained the top $r \in \{64, 80, 96\}$ Ritz directions, appended $\widehat{\boldsymbol{g}}$, diagonalised the positive reduced block, and represented the remaining measured trace by one isotropic bulk coordinate with its true multiplicity. Equation (77) then makes $\mathcal{T}_\eta$ an explicit operator on symmetric reduced covariances plus the bulk variance, whose spectral radius follows by matrix-free Arnoldi iteration. The relative leakage $\|\boldsymbol{HV} - \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{HV})\|_F / \|\boldsymbol{HV}\|_F$ of the retained subspace is 0.0036–0.0059, the stability labels are identical at ranks 64, 80 and 96, and from rank 80 to 96 the median absolute change in ρ is 3.50×10^{-6} with maximum 3.58×10^{-4}. The rank-96 values range from 0.999991 to 1.025749: under this reduced operator, 17 of the 24 checkpoints are asymptotically mean-square unstable and seven are stable. We then used this reduction to iterate the linearized recursion

$$\boldsymbol{e}_{t+1} = \big(\boldsymbol{I} - \eta a(\boldsymbol{u}_t)\boldsymbol{u}_t\boldsymbol{u}_t^\top \boldsymbol{H}\big)\boldsymbol{e}_t \tag{94}$$

with 256 independent replicas for 1,024 steps. The retained coordinates are updated explicitly. The isotropic bulk need not be materialised: by rotational invariance its update depends only on its norm, one standard normal radial coordinate and an independent χ^2_{d-r-1} transverse norm. Alongside the replica mean and median log norm we propagated the covariance operator itself for 65,536 steps, which evaluates $\mathbb{E}\|\boldsymbol{e}_t\|^2$ without the rare-event sampling error of a finite replica mean.

Within the reduced model, this test separates asymptotic stability. The sign of the late 65,536-step mean-square slope agrees with $\log \rho(\mathcal{T}_\eta)$ at all 24 clipped checkpoints: 17 grow and seven decay. Two caveats belong here rather than in a footnote. First, this agreement validates the *linear implementation* against the operator, not the fidelity of the finite-rank reduction: both quantities are computed from the same reduced $\mathcal{T}_\eta$, so their agreement is close to an identity relating power iteration to a spectral radius. Rank stability and low leakage are evidence for the reduction, but they establish its internal convergence rather than equality with the full operator. Second, the operator ranks the checkpoints only slightly better than the aligned scalar: Spearman 0.782 against 0.751 for $\tau_\perp$ with the earlier nonlinear amplification.

Thresholding $\tau_\perp/(2/\eta)$ labels 15 of the 24 as unstable against the operator's 17, and the two disagreements are both immediately above the boundary, at $\rho = 1.00000165$ and 1.00000217. It is tempting to read this as the scalar reduction failing off its premise. That reading is wrong, and the correct one is available in Table 11 without any new computation.

The two thresholds in that table are not equivalent. Comparing $\tau_\perp/(2/\eta)$ with one uses the high-rank simplification; comparing $\eta/\eta^\star_{\mathrm{aligned}}$ with one uses the root of the scalar equation $\Xi(\eta) = 1$ itself. Counting the second column, 17 of the 24 checkpoints exceed one, and they are the same 17 that $\rho(\mathcal{T}_\eta)$ calls unstable. The two extra points are precisely the ones the $\tau_\perp$ threshold misses: $\tau_\perp/(2/\eta) = 0.973$ and 0.966 become $\eta/\eta^\star_{\mathrm{aligned}} = 1.013$ and 1.012. So the disagreement is not between a scalar and the operator; it is between the trace-only approximation and the scalar equation it approximates.

The direction and size of the correction are the ones Theorem 4.1 predicts. Since $\eta^\star_{\mathrm{ms}} \le 2/\tau_\perp$, the trace-only threshold always understates instability, so it can only miss unstable points, never invent them — which is why the count moves from 15 up to 17 and not down. Across Table 11 the ratio of the two columns is 1.037–1.049, comfortably enough to lift 0.973 and 0.966 past one and far too

small to move any of the other 22. This is the same correction that the synthetic sweep measures as a step ratio of 1.15–1.41 at effective ranks 4.50 and 10.34; the measured checkpoints, whose effective ranks are 10.1–27.6, sit at the tight end, as higher rank predicts.

We are careful about what this does and does not show. It does not show that $\eta^{\star}_{\mathrm{aligned}}$ is the true boundary at these checkpoints: the alignment premise still fails with residual 0.669–0.911, and the agreement is at the level of labels, not values. What it does show is that the structural argument of Appendix C.13 — that the cross block b of Eq. (77) carries boundary information no diagonal functional retains — has not yet been detected in these data at the resolution of a stable/unstable call. The honest summary is that off the premise we can no longer *derive* a scalar, and we should not claim one is guaranteed, but the aligned plug-in remains an accurate diagnostic here.

The complementary and more useful observation is how *well* the aligned scalar survives. The alignment premise is not mildly violated here but substantially so, with residuals 0.669–0.911, and yet $\tau_{\perp}$ orders the checkpoints almost as well as the exact operator and agrees with it on 22 of 24 labels. For practice this says the aligned model remains a serviceable approximation off its premise, with the caveat that it should not be trusted within measurement error of the boundary.

We tested reduction fidelity independently at three high-margin unstable checkpoints by iterating Eq. (94) in all 30,570 coordinates, applying the checkpoint Hessian only through Hessian–vector products. With 128 replicas over 512 steps, the reduced/full mean-square exponents are $0.01499/0.01668$, $0.02542/0.03224$, and $0.00674/0.00949$. The full values are 11.3%, 26.8%, and 40.9% larger. A replica-bootstrap 95% interval contains the reduced value at the first two points (barely at the first) but not at the third, whose interval is $[0.00899, 0.00997]$. Thus the reduction gets the sign and ordering right at these points but is not quantitatively exact; in this check it systematically underestimates growth. This is the independent fidelity evidence that the reduced covariance propagation itself could not provide.

Two features of these three numbers bound what they license. The relative bias is largest at the *smallest* exponent — 11.3%, 26.8%, 40.9% at reduced exponents 0.01499, 0.02542, 0.00674 — while the absolute bias stays in 0.0017–0.0068. And all three are high-margin points chosen precisely because their sign is unambiguous, whereas the checkpoints on which the operator and $\tau_{\perp}$ disagree have $|\log \rho| \sim 10^{-6}$, three to four orders of magnitude smaller than anything tested here. A multiplicative bias of the observed size would leave those labels intact; an additive bias of the observed size would not. We therefore report the $17/7$ split and the 22-of-24 agreement as properties of the finite-rank reduction, with independent verification only at high margin.

We ran the same full-dimensional HVP recursion at the reduced-stable $(\eta, C, t) = (0.001, 0.24, 90{,}000)$ checkpoint and at the two operator–scalar disagreements, $(0.001, 0.24, 100{,}000)$ and $(0.002, 0.24, 80{,}000)$. Over 512 steps, the first-to-last-quarter changes in $\log \mathbb{E}\|\boldsymbol{e}_t\|^2$ are respectively 0.625, 0.658, and 0.860, with replica-bootstrap 95% intervals $[0.614, 0.634]$, $[0.647, 0.669]$, and $[0.842, 0.878]$. The reduced asymptotic predictions accumulated over the same quarter separation are -0.00336, 0.000635, and 0.000835. Thus all three full-HVP curves are still in a large positive transient, including the point the reduction calls stable; even their last-quarter slopes remain positive. The experiment therefore does not determine any of the three asymptotic signs and cannot convert the $17/7$ split into a full-dimensional measurement.

We also formed the full covariance operator exactly in a 16-dimensional quadratic model. We fixed $\boldsymbol{H} = \mathrm{diag}(4, 1.5, \ldots, 0.5, 0.005)$, set $\kappa = 1$, and rotated $\widehat{\boldsymbol{g}}$ in the plane of the largest and smallest eigenvectors so that the alignment residual sweeps 0–0.998. At each angle we solved $\rho(\mathcal{T}_{\eta}) = 1$ directly, with no reduction. The full aligned-formula plug-in is exact at residual zero; its largest step error is only 1.69%, at residual 0.694, where the true operator evaluated at that step has $\rho - 1 = 0.00284$. The error then decreases even while the residual continues to grow, reaching 10^{-5} in ρ at residual 0.998.

The non-monotonicity is not a numerical curiosity; it shows that the normalised residual alone cannot bound the scalar error. With $\widehat{\boldsymbol{g}} = \cos\theta\, \boldsymbol{v}_1 + \sin\theta\, \boldsymbol{v}_{16}$ and $\lambda_1/\lambda_{16} = 800$, the residual $\|(\boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^{\top})\boldsymbol{H}\widehat{\boldsymbol{g}}\|/\|\boldsymbol{H}\widehat{\boldsymbol{g}}\|$ is maximised at $\tan^2\theta = \lambda_1/\lambda_{16}$, namely $\theta = 88.0°$, where it equals 0.998. Although $\widehat{\boldsymbol{g}}$ is then Euclidean-close to $\boldsymbol{v}_{16}$, its small $\boldsymbol{v}_1$ component is amplified 800-fold by $\boldsymbol{H}$, making $\boldsymbol{H}\widehat{\boldsymbol{g}}$ nearly orthogonal to $\widehat{\boldsymbol{g}}$; the alignment premise is therefore maximally, not minimally, violated. What has already fallen is the absolute cross-block norm $\|\boldsymbol{Q}\boldsymbol{H}\widehat{\boldsymbol{g}}\| = (\lambda_1 - \lambda_{16})|\sin\theta\cos\theta|$, from nearly 2 around 45° to 0.141 at the terminal angle. Dividing it by the simultaneously shrinking

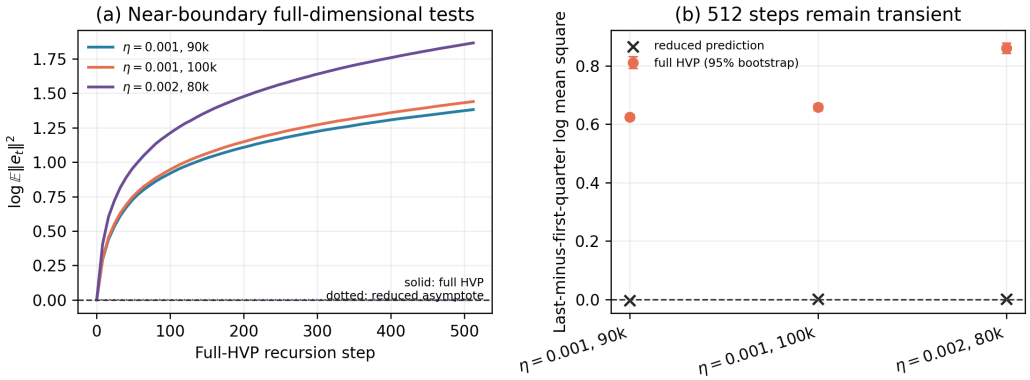

Figure 11: Near-boundary full-dimensional HVP checks with 128 replicas. **(a)** Solid curves are the observed log mean square over 512 steps and dotted lines are the reduced-operator asymptotes. **(b)** First-to-last quarter changes with 95% whole-replica bootstrap intervals; crosses are the reduced predictions over the same time separation. All three observed curves are positive and decelerating but have not reached the 10^{-6}-scale asymptotic regime, so the finite-window sign is a transient observation rather than a stability label.

$\|\boldsymbol{H}\widehat{\boldsymbol{g}}\|$ hides that scale. This is why the operator error can decrease while the normalised residual grows: the residual measures an angle, but not the magnitude of the discarded coupling relative to the transverse trace.

The trace-only step $2/\tau_\perp$ behaves differently and more informatively. In the initial spiked-spectrum, $\kappa = 1$ sweep its operator excess is 4.28%–5.01%, including at perfect alignment, so it contains a finite-rank/radial correction that is not caused by misalignment. To separate rank from gate effects, we repeated all 25 angles at $\kappa \in \{0.5, 1, 2\}$ for both this spectrum, whose trace-over-top effective rank is 4.50, and a flatter spectrum with effective rank 10.34. At every κ the flatter spectrum reduces the trace-only excess: the spiked/flat ranges are 1.77–$2.81\%/1.42$–1.64% at $\kappa = 0.5$, 4.28–$5.01\%/2.54$–3.31% at $\kappa = 1$, and 7.02–$14.40\%/3.55$–6.25% at $\kappa = 2$. The maximum operator discrepancy of the full aligned-formula plug-in likewise falls from $0.043/0.284/1.125\%$ for the spiked spectrum to $0.0057/0.037/0.141\%$ for the flat spectrum at $\kappa = 0.5, 1, 2$, respectively. The rank-10.34 case overlaps the low end of the measured effective-rank range 10.1–27.6. At the 24 network checkpoints the measured gate widths span 0.339–2.504, corresponding to gate rates 0.265–0.988. Split by threshold, the κ ranges are 0.339–0.443, 0.478–0.741, 1.029–1.198, and 1.723–2.504 for $C = 0.03, 0.06, 0.12, 0.24$, respectively; Table 11 reports every value beside p. The synthetic interval 0.5–2 covers all six $C = 0.12$ points, five of six $C = 0.06$ points, and one $C = 0.24$ point: 12 of 24 widths in total. Transfer to the seven strongest-gated and five weakest-gated points therefore requires extrapolation beyond the tested gate-width interval. These controls establish the rank dependence, but also show a substantial gate-moment dependence: spectral rank and κ jointly limit the scalar approximation, whereas the alignment residual alone does not.

The direction of the gate effect is predicted rather than incidental, but its magnitude is not $\bar{c} - 1$ alone. At the aligned endpoint and at the trace-only step, the exact scalar equation factorises as

$$\Xi(2/\tau_\perp) = \underbrace{\left[1 + r(2/\tau_\perp)\right]}_{\bar{c}} \underbrace{\frac{1}{\tau_\perp} \sum_{j \neq r} \frac{\lambda_j}{1 - 2\lambda_j/\tau_\perp}}_{R_\perp}.$$

The trace-only approximation discards both the radial correction $\bar{c} - 1$ and the transverse resolvent correction $R_\perp - 1$, which is 18.84% for the spiked spectrum and 17.00% for the flat spectrum. The second of these is not an omission of Theorem 4.1: since $\lambda_j/(1 - 2\lambda_j/\tau_\perp) \leq \lambda_j/(1 - 2\lambda^\perp_{\max}/\tau_\perp)$, we have $R_\perp \leq (1 - 2\lambda^\perp_{\max}/\tau_\perp)^{-1}$, so $R_\perp - 1$ is exactly what the corollary's slack term $2\lambda^\perp_{\max}/\tau_\perp$ already covers. The corollary's content is the combined bound $\bar{c} + 2\lambda^\perp_{\max}/\tau_\perp$, not the factor $\bar{c}$ alone.

The second and more consequential mismatch is one of units. Theorem 4.1 bounds the critical *step* $\eta^\star_{\mathrm{ms}}$, whereas the last two columns of Table 4 report a *spectral radius* excess at a fixed step. These are

Table 4: Direct synthetic test of Theorem 4.1 at $\widehat{\boldsymbol{g}} = \boldsymbol{v}_1$. The fifth column is the trace-only step divided by the exact mean-square critical step, obtained from the scalar root $\Xi(\eta) = 1$; the last column is its upper bound in Eq. (95). Only $100(\bar{c} - 1)$ is a percentage; the final two columns are dimensionless step ratios.

Spectrum	r_{eff}	κ	$100(\bar{c} - 1)$	Exact step ratio	Corollary upper bound
Spiked	4.50	0.5	2.403	1.1785	1.2382
Spiked	4.50	1	9.915	1.2355	1.3134
Spiked	4.50	2	46.306	1.4076	1.6773
Flat	10.34	0.5	0.877	1.1529	1.1801
Flat	10.34	1	3.319	1.1733	1.2046
Flat	10.34	2	10.308	1.2257	1.2744

related by a monotone but nontrivial reparameterisation, so no direct numerical agreement should be expected between $\bar{c} - 1$ and $\rho(\mathcal{T}_{2/\tau_\perp}) - 1$. In the aligned case the comparison can be made without any conversion, because Theorem C.20 makes $\Xi(\eta) = 1$ the exact boundary: solving that scalar equation for $\eta^\star_{\text{ms}}$ turns the corollary into the sharp inequality

$$\frac{2/\tau_\perp}{\eta^\star_{\text{ms}}} \;\le\; \bar{c} + \frac{2\lambda^\perp_{\max}}{\tau_\perp}, \tag{95}$$

which is a statement purely about steps and is either true or false at each of the six settings.

Table 4 makes the comparison in the units the corollary actually uses. All six exact step ratios satisfy Eq. (95); the additive slack (upper bound minus exact ratio) ranges from 0.0273 to 0.2696. The bound becomes looser at lower effective rank and larger κ, as expected when the radial and transverse resolvent corrections are both large. Thus the direct test supports the corollary, while also quantifying how much sharpness its separate bounds lose.

The relation between the two units is itself informative. Writing the scalar-unit excess at the trace-only step as $\Xi(2/\tau_\perp) - 1 = \bar{c}R_\perp - 1$, the ratio $(\bar{c}R_\perp - 1)/(\rho(\mathcal{T}_{2/\tau_\perp}) - 1)$ evaluates to 10.97 and 11.00 at $\kappa = 0.5$, to 6.32 and 6.32 at $\kappa = 1$, and to 5.13 and 4.65 at $\kappa = 2$, where each pair is (spiked, flat). Multiplying by the gate rate $p = 2\Phi(\kappa) - 1$ gives $4.20, 4.21$, then $4.32, 4.31$, then $4.90, 4.44$. These values were recomputed directly rather than inferred from rounded table entries.

An 11-point sweep over $\kappa = 0.35$–2.5, essentially the full measured checkpoint range, shows the limit of that apparent collapse. The p-scaled ratio stays in 4.165–4.497 for the flat spectrum but ranges from 4.147 to 5.752 for the spiked spectrum, rising visibly beyond $\kappa \simeq 1.5$. Over the moderate-to-strong gating regime $\kappa \le 1.5$, all values lie in 4.147–4.496, so $4.3/p$ is a useful empirical approximation there, but it is not a spectrum-independent conversion law over the full range.

This qualified $1/p$ trend is consistent with leading gate-rate cancellation. A shared gate enters drift and diffusion at the same gate order, but the normalised moments m_2/p and m_4/p, together with the spectrum, still affect the boundary and the rate at which ρ departs from one. Factoring out p captures only the leading effect. At a simple boundary the exact infinitesimal conversion is

$$\lim_{\eta \to \eta^\star_{\text{ms}}} \frac{\Xi(\eta) - 1}{\rho(\mathcal{T}_\eta) - 1} = \frac{\Xi'(\eta^\star_{\text{ms}})}{\frac{\mathrm{d}}{\mathrm{d}\eta}\rho(\mathcal{T}_\eta)\Big|_{\eta = \eta^\star_{\text{ms}}}}. \tag{96}$$

Numerical differentiation gives p times this local ratio in 5.09–5.55 for the flat spectrum and 5.12–6.68 for the spiked one. It is therefore an exact local formula, not a missing universal constant for the finite excess measured at $2/\tau_\perp$. The local values exceed the finite ones by roughly a fifth at every setting, and the sign of that offset is expected rather than residual: the finite excess is a secant taken over the interval from $\eta^\star_{\text{ms}}$ to $2/\tau_\perp$, whose width is the step ratio 1.15–1.41 of Table 4, so a convex $\rho(\eta)$ makes the secant slope fall below the tangent slope at the boundary.

We also performed the paired ungated control at every checkpoint by setting $a \equiv 1$ while holding the same reduced Hessian fixed. The two evaluation steps were $0.5\eta^\star_{\text{ungated}}$ and $1.5\eta^\star_{\text{ungated}}$; equivalently their raw-trace ratios span 0.476–0.486 and 1.428–1.458, so they bracket $2/\tau$ with a wide margin. The long covariance slopes decay at all 24 below-edge settings and grow at all 24 above-edge settings.

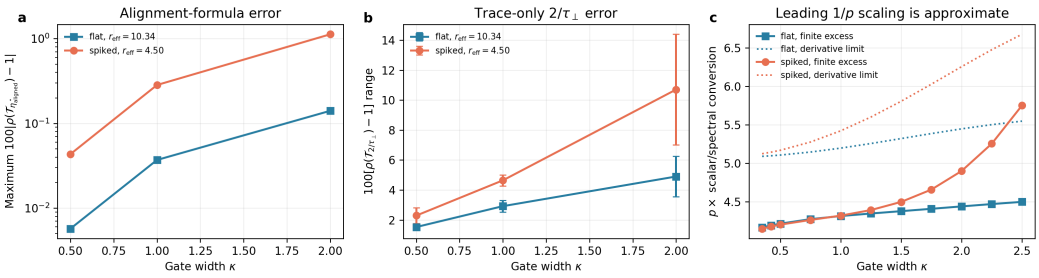

Figure 12: Exact synthetic controls across two spectra and an extended gate sweep. **(a)** Maximum operator discrepancy at the full aligned-formula step; the log scale shows an approximately eightfold gap between effective ranks 4.50 and 10.34 at each κ. **(b)** Range across gradient angles of the operator excess at the trace-only step $2/\tau_\perp$. Higher effective rank improves both approximations, while increasing κ worsens both, so neither alignment nor rank alone controls the error. **(c)** Gate-scaled conversion between the scalar residual and operator excess: solid curves use the finite excess at $2/\tau_\perp$, while dotted curves use the local derivative ratio in Eq. (96). The leading $1/p$ collapse is accurate under moderate-to-strong gating but breaks down for the low-rank spectrum as the gate becomes weak.

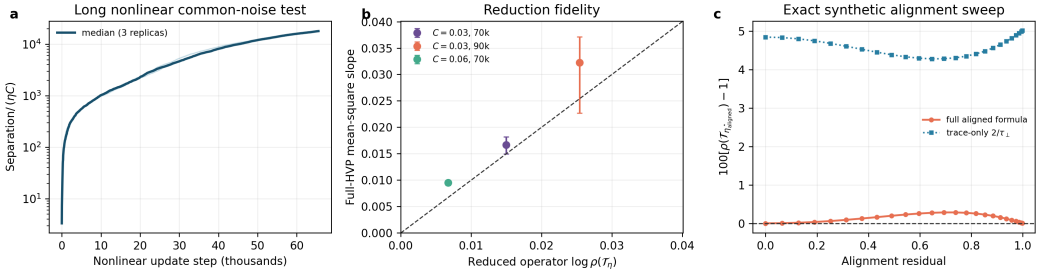

Figure 13: The three follow-up tests. **(a)** Three nonlinear common-noise pairs at the final strongly gated batch checkpoint ($\eta = 0.001, C = 0.03$), scaled by the one-step coefficient ηC; separation continues to grow over $65{,}536$ steps rather than plateauing. **(b)** Full-dimensional HVP mean-square slopes against the reduced operator exponents at three checkpoints; bars are 95% replica-bootstrap intervals. The reduction preserves sign and ordering but underestimates the measured rates. **(c)** Exact synthetic alignment sweep. The orange curve evaluates the full operator at the step from the aligned scalar equation; the blue curve uses the high-rank approximation $2/\tau_\perp$. Misalignment error is nonmonotone because the normalised residual omits the absolute cross-block scale, so it is not an error bound. The $\tau_\perp$ excess remains nonzero at perfect alignment, isolating a finite-rank/radial correction from the separate misalignment effect.

Thus the operator computation and the linear implementation pass the negative control. The control is paired: it reuses the reduced Hessian of a clipped-trained checkpoint with the gate set to one, which isolates the effect of the gate but is not a set of independently trained ungated trajectories.

The short window gives the opposite lesson. All 24 clipped replica means and medians increase over 1,024 steps, including the seven points with $\rho(\mathcal{T}_\eta) < 1$; even all 24 ungated below-edge controls initially increase. A representative stable checkpoint in Figure 15(d) peaks after roughly 10^3 steps and then decays, while the 1,024-step mean and median see only the rising part. This is nonnormal transient amplification, not asymptotic instability. It also explains the all-positive nonlinear result above without asking $\tau_\perp$ or any other scalar to classify a finite-time phenomenon. The 256-replica mean-square and median-log slopes can visibly differ in the most unstable runs, but their more important shared failure here is horizon bias: both are positive before stable recursions enter their asymptotic regime.

The structural gate-counting predictions are cleaner because they do not require radial alignment. Across the 24 batch checkpoints, $\mathrm{Tr}(\boldsymbol{K})/(2/\eta)$ regressed on $p\tau_\perp/(2/\eta)$ through the origin has slope

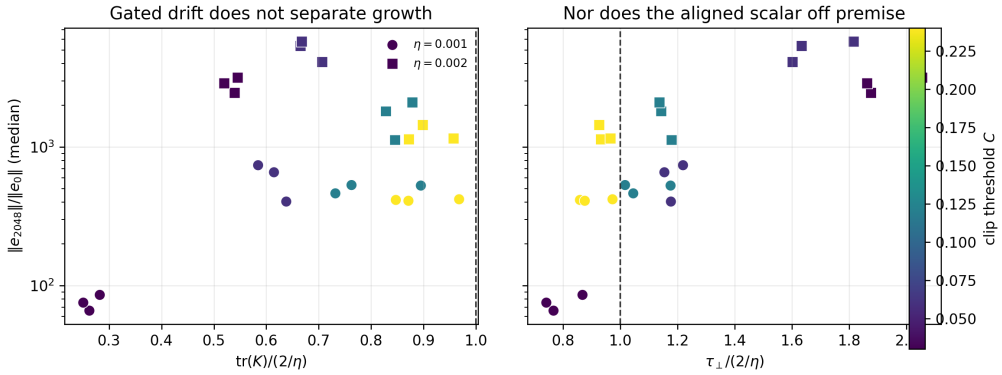

Figure 14: Common-noise nonlinear amplification at all 24 batch-ZO late checkpoints (three replicates per point; medians shown). Left: the gated drift is below the nominal line everywhere and does not rank growth. Right: $\tau_\perp$ is more strongly associated with growth (Spearman 0.751 against 0.095), but nine points below the nominal line still amplify. Every one of the 24 checkpoints amplifies, so this data contains no negatives and cannot separate any candidate scalar; the figure therefore rules out $\tau_\perp$ as a premise-free classifier here without testing it as a boundary. This sharpens rather than removes the limitation stated in Section 6.

1.006 and Pearson correlation 0.99995 (Figure 16, left). The omitted radial contribution is only 0.28% at the median and 2.12% at worst. That number is the same fact as the alignment failure, seen from another side: a radially aligned $\widehat{\boldsymbol{g}}$ would contribute $\widehat{\boldsymbol{g}}^\top \boldsymbol{H}\widehat{\boldsymbol{g}} = \lambda_1$, roughly a 6% share of the trace at the measured effective ranks, so a contribution below 2% says the gradient direction is spread across the spectrum. It also has a consequence worth stating on its own: on these checkpoints $\tau_\perp$ and the raw trace differ by at most two percent, so gradient-direction suppression — whose FO endpoint is complete direction removal — is numerically negligible on the ZO side, and the two candidate ZO scalars are not empirically distinct objects here. On the right, the shared-gate identity lies on $\varsigma = 1/\bar{p}$, whereas the three measured per-sample values 1.03975, 1.00021, and 1.0000002 lie near the decoupled endpoint $\varsigma = 1$. These are direct tests of gate order; they do not invoke the aligned scalar boundary.

The requested codimension and trajectory-level FO tests are already the two panels of Figure 8: the former gives the 0.888 Pearson association between γ_k and the proxy–exact gap, and the latter shows the predicted slope-one relation between sharpness and gradient norm on log–log axes. Table 7 gives the pointwise skew and cluster checks. The six values obey $\big(|(2/\eta^\star_{\mathrm{ps}}) - S|/S\big)/(\|\boldsymbol{R}_a\|_2/S)^2 = 0.531$–$1.910$, but the free log–log fit has slope 0.32 because the horizontal variable spans only a factor 1.78. Thus these checkpoints verify the predicted magnitude, not the exponent. The logically decisive observation is in the right panel: every simple gap falls below $4\|\boldsymbol{R}_a\|_2$, while every selected cluster separation lies above it. The classical route is unavailable and the cluster/gap-free argument is applicable.

D Exact finite-difference ZO linearization

This appendix records the exact Jacobians suppressed from the main setup. At a fixed checkpoint $\boldsymbol{x}$, differentiate the centered finite difference:

$$\boldsymbol{r}_i^\mu(\boldsymbol{u}) := \nabla_{\boldsymbol{x}} d_i^\mu(\boldsymbol{x};\boldsymbol{u}) = \frac{\boldsymbol{g}_i(\boldsymbol{x}+\mu\boldsymbol{u}) - \boldsymbol{g}_i(\boldsymbol{x}-\mu\boldsymbol{u})}{2\mu}, \qquad \boldsymbol{r}^\mu(\boldsymbol{u}) = \frac{1}{n}\sum_i \boldsymbol{r}_i^\mu(\boldsymbol{u}).$$

Away from the scalar clipping kinks, define the exact gates

$$\bar{a}^\mu_{\mathrm{batch}}(\boldsymbol{u}) = \mathbf{1}\{|d^\mu(\boldsymbol{x};\boldsymbol{u})| < C\}, \qquad \bar{a}^\mu_i(\boldsymbol{u}) = \mathbf{1}\{|d^\mu_i(\boldsymbol{x};\boldsymbol{u})| < C\}.$$

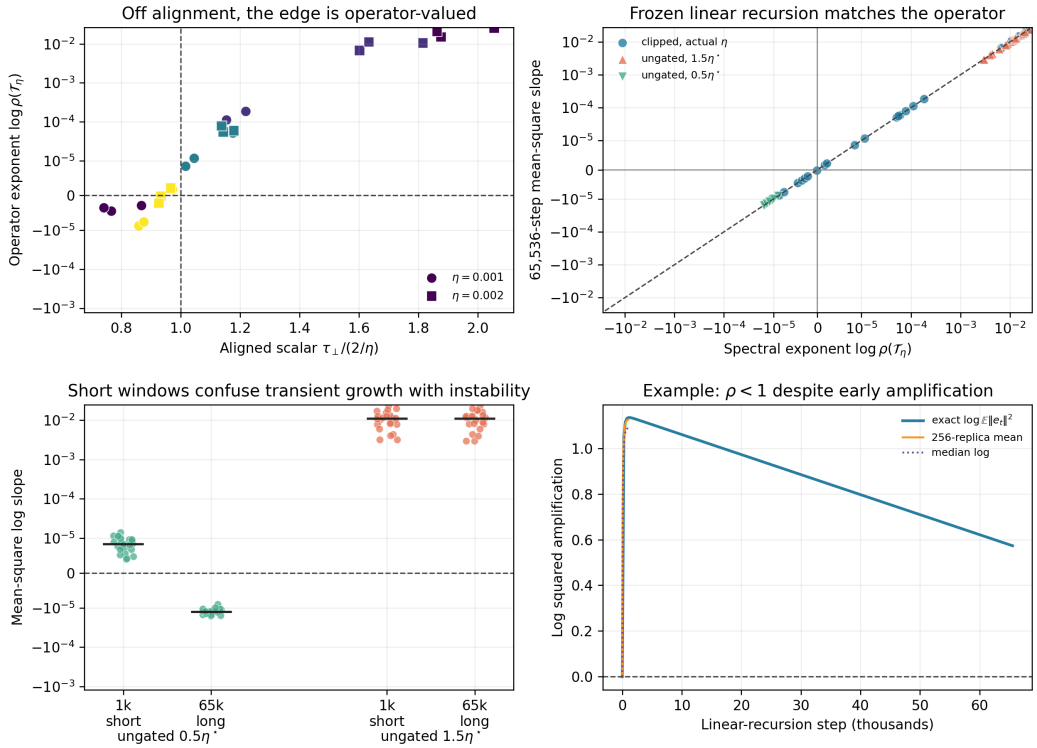

Figure 15: Non-aligned covariance-operator and linearized controls. (a) The rank-96 reduced operator labels 17 checkpoints unstable; the trace-only threshold $\tau_\perp$ labels 15 and misses the two points left of the vertical line but above the horizontal line. Those two are recovered by the root of the scalar equation $\Xi(\eta) = 1$, which matches the operator on all 24. Colour denotes C and circles/squares denote $\eta = 0.001/0.002$. (b) The 65,536-step exact covariance slope lies on $\log \rho(\mathcal{T}_\eta)$ for clipped checkpoints and paired ungated controls. (c) The ungated below-edge control appears to grow at 1,024 steps but decays at 65,536; the above-edge control grows at both horizons. Black bars are medians over 24 checkpoints. (d) At a representative stable clipped checkpoint ($\eta = 0.001, C = 0.24$, step 90,000), the exact mean square first amplifies and then decays. The 256-replica mean and median-log curves end inside the initial transient. Hence short amplification and asymptotic mean-square instability are different observables.

Applying the chain rule to the update directions in Eqs. (1)–(2) gives

$$D_{\boldsymbol{x}}[\mathrm{clip}_C(d^\mu(\boldsymbol{x};\boldsymbol{u}))\boldsymbol{u}] = \bar{a}^\mu_{\mathrm{batch}}(\boldsymbol{u})\boldsymbol{u}[\boldsymbol{r}^\mu(\boldsymbol{u})]^\top,$$

$$D_{\boldsymbol{x}}\left[\frac{1}{n}\sum_i \mathrm{clip}_C(d_i^\mu(\boldsymbol{x};\boldsymbol{u}))\boldsymbol{u}\right] = \boldsymbol{u}\left[\frac{1}{n}\sum_i \bar{a}_i^\mu(\boldsymbol{u})\boldsymbol{r}_i^\mu(\boldsymbol{u})\right]^\top.$$

Assuming additionally that each ℓ_i is four times continuously differentiable in a neighborhood of the evaluation points, the centered expansion gives $\boldsymbol{r}_i^\mu(\boldsymbol{u}) = \boldsymbol{H}_i\boldsymbol{u} + O(\mu^2)$. Replacing the exact gates by their small-μ limits in (14) recovers the ZO Jacobians analyzed in Section 4. In particular, the gate—rather than the nonlinear multiplier $\min(1, C/|d_i^\mu|)$—appears because it is the derivative of scalar clipping, and it is the idempotence of this gate that drives Lemma C.9.

E Proofs

E.1 First-order gradient-direction removal

Proof of Lemma C.7. If $\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\lambda \neq 0$, applying $\boldsymbol{Q}_k$ gives $\lambda\boldsymbol{Q}_k\boldsymbol{x} = \lambda\boldsymbol{x}$, hence $\boldsymbol{Q}_k\boldsymbol{x} = \boldsymbol{x}$. Therefore $\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k\boldsymbol{x} = \boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\boldsymbol{x} \in \mathrm{range}(\boldsymbol{Q}_k)$. Conversely, if $\boldsymbol{x} \in \mathrm{range}(\boldsymbol{Q}_k)$

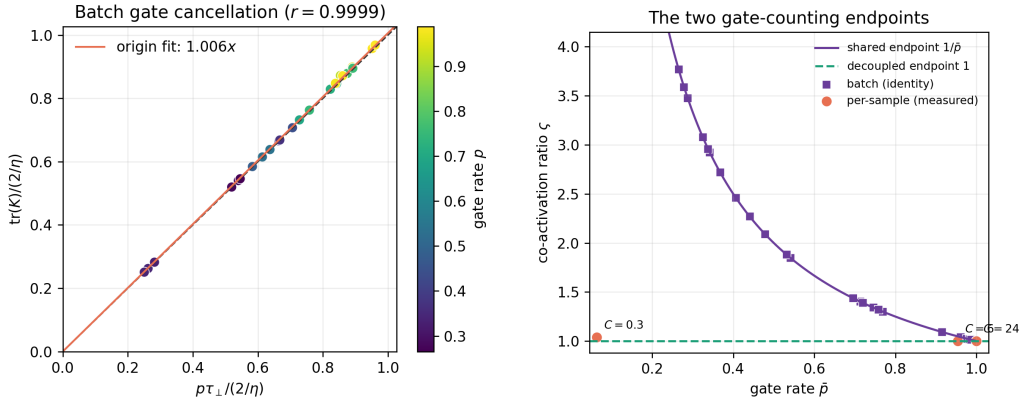

Figure 16: Direct gate-counting tests. Left: the exact batch drift is nearly $p\tau_\perp$, with the small displacement above the identity supplied by the radial term $m_2\widehat{\boldsymbol{g}}^\top \boldsymbol{H}\widehat{\boldsymbol{g}}$. Right: batch and per-sample gates occupy opposite co-activation endpoints. Batch squares are the exact shared-gate identity; per-sample circles are Monte Carlo pair-moment measurements at $C = 0.3, 6, 24$.

and $\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k\boldsymbol{x} = \lambda\boldsymbol{x}$, then $\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{x} = \lambda\boldsymbol{x}$. Algebraic multiplicity is preserved because in the decomposition $\mathrm{range}(\boldsymbol{Q}_k) \oplus \ker(\boldsymbol{Q}_k)$ the matrix of $\boldsymbol{Q}_k\boldsymbol{H}$ is $\left(\begin{smallmatrix} \boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k & \boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{\Pi}_k \\ 0 & 0 \end{smallmatrix}\right)$, which is block upper triangular. If $\boldsymbol{H} \succeq 0$, write this block matrix as $\left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B} \\ 0 & 0 \end{smallmatrix}\right)$. Positive semidefiniteness of the corresponding symmetric block matrix implies $\mathrm{range}(\boldsymbol{B}) \subseteq \mathrm{range}(\boldsymbol{A})$, so $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{X}$ for some $\boldsymbol{X}$. A block shear then makes $\boldsymbol{Q}_k\boldsymbol{H}$ similar to $\left(\begin{smallmatrix} \boldsymbol{A} & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Since $\boldsymbol{A} = \boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k$ is symmetric positive semidefinite on $\mathrm{range}(\boldsymbol{Q}_k)$, diagonalizability and semisimplicity of zero follow. □

Proof of Theorem 3.1. If $\|\boldsymbol{g}\| < C$, including $\boldsymbol{g} = 0$, then $\boldsymbol{J}_{\mathrm{batch}} = \boldsymbol{H}$. Hence $s_{\mathrm{batch}} = \lambda_1(\boldsymbol{H})$ and $\eta^\star_{\mathrm{batch}} = 2/s_{\mathrm{batch}}$, with the convention $2/0 = +\infty$, which proves the first branch of (8). Suppose now that $\|\boldsymbol{g}\| > C$. Then $\alpha = C/\|\boldsymbol{g}\|$, and (7) gives the local Jacobian $\alpha\boldsymbol{Q}\boldsymbol{H}$. Moreover, Lemma C.7 identifies its nonzero spectrum with that of $\alpha\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}$ on $\mathrm{range}(\boldsymbol{Q})$. This spectrum is nonnegative and real, so $s_{\mathrm{batch}} = \alpha\lambda_1(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q})$ and $\eta^\star_{\mathrm{batch}} = 2/s_{\mathrm{batch}}$, proving the second branch of (8). Cauchy interlacing gives

$$\alpha\lambda_2(\boldsymbol{H}) \le \alpha\lambda_1(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}) \le \alpha\lambda_1(\boldsymbol{H}).$$

Inverting the sandwich gives (9), with a zero denominator interpreted as an infinite endpoint. □

Proof of Theorem C.8. The vectors $\boldsymbol{E}_\ell\widehat{\boldsymbol{g}}^{(\ell)}$, $\ell \in \mathcal{B}$, are orthonormal because the blocks are disjoint, so $\boldsymbol{\Pi}_k$ is an orthogonal projector of rank k.

Case 1. With a common α and all blocks active, $\boldsymbol{D} = \alpha\boldsymbol{Q}_k$, so $\mathbf{J}_{\mathrm{FO,blk}} = \alpha\boldsymbol{Q}_k\boldsymbol{H}$ and Lemma C.7 gives $q^{\mathrm{ex}}_{\mathrm{blk}} = \alpha\lambda_1(\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k|_{\mathrm{range}(\boldsymbol{Q}_k)})$. Cauchy interlacing for a codimension-k compression (Horn & Johnson, 2013) gives

$$\lambda_1(\boldsymbol{H}) \;\ge\; \lambda_1(\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k|_{\mathrm{range}(\boldsymbol{Q}_k)}) \;\ge\; \lambda_{k+1}(\boldsymbol{H}),$$

and multiplying by α proves (60).

Case 2. $\boldsymbol{D}$ in (55) is symmetric PSD with $\alpha_{\min}\boldsymbol{Q}_k \preceq \boldsymbol{D} \preceq \alpha_{\max}\boldsymbol{Q}_k$, because each block contributes α_ℓ times the projector that $\boldsymbol{Q}_k$ contributes on that same block — namely $\boldsymbol{I} - \widehat{\boldsymbol{g}}^{(\ell)}\widehat{\boldsymbol{g}}^{(\ell)\top}$ if ℓ is active and the block identity if it is not — and the blocks are orthogonal. Taking $\alpha_{\min}$ and $\alpha_{\max}$ over all L blocks, with $\alpha_\ell = 1$ on the inactive ones, is what makes the sandwich hold on the inactive blocks too. For $\boldsymbol{H} \succeq 0$ the nonzero spectrum of $\boldsymbol{D}\boldsymbol{H}$ equals that of $\boldsymbol{H}^{1/2}\boldsymbol{D}\boldsymbol{H}^{1/2}$, and $\boldsymbol{D} \mapsto \lambda_1(\boldsymbol{H}^{1/2}\boldsymbol{D}\boldsymbol{H}^{1/2})$ is monotone in the Loewner order, so $\alpha_{\min}\lambda_1(\boldsymbol{H}^{1/2}\boldsymbol{Q}_k\boldsymbol{H}^{1/2}) \le q^{\mathrm{ex}}_{\mathrm{blk}} \le \alpha_{\max}\lambda_1(\boldsymbol{H}^{1/2}\boldsymbol{Q}_k\boldsymbol{H}^{1/2})$, and $\lambda_1(\boldsymbol{H}^{1/2}\boldsymbol{Q}_k\boldsymbol{H}^{1/2}) = \lambda_1(\boldsymbol{Q}_k\boldsymbol{H}\boldsymbol{Q}_k|_{\mathrm{range}(\boldsymbol{Q}_k)})$ is interlaced as in Case 1.

A tangent mode with positive curvature λ has multiplier $1 - s\lambda$ and is nonexpansive iff $0 \le s\lambda \le 2$, so under (12) the maximum stable step on the positive branch is $2/q^{\mathrm{ex}}_{\mathrm{blk}}$; inverting the sandwich gives (61). □

Proof of Corollary C.3. If $\lambda_{k'}(\boldsymbol{H}) \geq (1-\epsilon)\lambda_1(\boldsymbol{H})$ and $k < k'$ then $\lambda_{k+1}(\boldsymbol{H}) \geq \lambda_{k'}(\boldsymbol{H}) \geq (1-\epsilon)\lambda_1(\boldsymbol{H})$; apply (62). □

E.2 First-order per-sample: Theorem C.24 and Proposition C.28

Proof of Theorem C.24. Let $W(\boldsymbol{A}-\boldsymbol{R}) = \{\boldsymbol{x}^*(\boldsymbol{A}-\boldsymbol{R})\boldsymbol{x} : \boldsymbol{x} \in \mathbb{C}^d, \|\boldsymbol{x}\| = 1\}$ be the numerical range; $\mathrm{spec}(\boldsymbol{A}-\boldsymbol{R}) \subseteq W(\boldsymbol{A}-\boldsymbol{R})$. Since $\boldsymbol{A}-\boldsymbol{R} = (\boldsymbol{A}-\boldsymbol{R}_s)-\boldsymbol{R}_a$, its Hermitian part is $\boldsymbol{A}-\boldsymbol{R}_s$, so $\Re(\boldsymbol{x}^*(\boldsymbol{A}-\boldsymbol{R})\boldsymbol{x}) = \boldsymbol{x}^*(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{x} \in [\lambda_{\min}(\boldsymbol{A}-\boldsymbol{R}_s), \lambda_1(\boldsymbol{A}-\boldsymbol{R}_s)]$. Its skew-Hermitian part is $-\boldsymbol{R}_a$, so $\Im(\boldsymbol{x}^*(\boldsymbol{A}-\boldsymbol{R})\boldsymbol{x}) = \boldsymbol{x}^*(i\boldsymbol{R}_a)\boldsymbol{x}$, and $i\boldsymbol{R}_a$ is Hermitian with eigenvalues equal to the real numbers $\mp b_k$, where $\pm ib_k$ are the eigenvalues of $\boldsymbol{R}_a$; hence $|\Im(\boldsymbol{x}^*(\boldsymbol{A}-\boldsymbol{R})\boldsymbol{x})| \leq \|\boldsymbol{R}_a\|_2$. This proves (90), and (91) follows from $|\lambda|^2/\Re\lambda = \Re\lambda + (\Im\lambda)^2/\Re\lambda$. Taking the maximum over the positive-real branch and using $\Re\lambda^\star > 0$ gives (92). Note that only $\|\boldsymbol{R}_a\|_2$ enters, and that the first-order term is absent because $\boldsymbol{R}_a$ contributes nothing to the Hermitian part. □

Proof of Proposition C.28. Count. Let Γ be the boundary of the $\Gamma_c/2$-neighbourhood of the cluster, so every point of Γ is at distance at least $\Gamma_c/2$ from $\mathrm{spec}(\boldsymbol{A}-\boldsymbol{R}_s)$ and $\|\boldsymbol{R}_0(z)\|_2 \leq 2/\Gamma_c$ on Γ, where $\boldsymbol{R}_0(z) = (z-(\boldsymbol{A}-\boldsymbol{R}_s))^{-1}$. Since $z-(\boldsymbol{A}-\boldsymbol{R}) = (z-(\boldsymbol{A}-\boldsymbol{R}_s))+\boldsymbol{R}_a$, the hypothesis $4\|\boldsymbol{R}_a\|_2 \leq \Gamma_c$ gives $\|\boldsymbol{R}_0\boldsymbol{R}_a\|_2 \leq 1/2$, so $z-(\boldsymbol{A}-\boldsymbol{R}) = \big(z-(\boldsymbol{A}-\boldsymbol{R}_s)\big)(\boldsymbol{I}+\boldsymbol{R}_0\boldsymbol{R}_a)$ is invertible on Γ and the two Riesz projections depend continuously on the interpolation $\boldsymbol{R}_a \mapsto t\boldsymbol{R}_a$, $t \in [0,1]$; their ranks are integers, hence equal (Kato, 1995). So $\boldsymbol{A}-\boldsymbol{R}$ has exactly $\mathrm{rank}(\boldsymbol{P})$ eigenvalues inside Γ.

Bounds. These follow without any invariant-subspace tilt estimate. Let λ be such an eigenvalue with unit eigenvector $\boldsymbol{x}$, and split $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{r}$, $\boldsymbol{x}_0 = \boldsymbol{P}\boldsymbol{x}$, $\boldsymbol{r} = (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}$; note $\boldsymbol{P}$ is an *orthogonal* projector because $\boldsymbol{A}-\boldsymbol{R}_s$ is Hermitian. Projecting $\big((\boldsymbol{A}-\boldsymbol{R}_s)-\boldsymbol{R}_a\big)\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\boldsymbol{I}-\boldsymbol{P}$ gives $\big((\boldsymbol{A}-\boldsymbol{R}_s)-\lambda\big)\boldsymbol{r} = (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{R}_a\boldsymbol{x}$. On $\mathrm{range}(\boldsymbol{I}-\boldsymbol{P})$ the spectrum of $\boldsymbol{A}-\boldsymbol{R}_s$ is at distance at least $\Gamma_c/2$ from λ, because λ lies within $\Gamma_c/2$ of the cluster and the cluster is separated from the rest by Γ_c; hence

$$\|\boldsymbol{r}\| \;\leq\; \frac{2\|\boldsymbol{R}_a\|_2}{\Gamma_c} \;\leq\; \frac{1}{2}. \tag{97}$$

For the real part, $\Re\lambda = \boldsymbol{x}^*(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{x}$ because $\boldsymbol{R}_a$ is skew, and the cross terms vanish since $(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{x}_0 \in \mathrm{range}(\boldsymbol{P}) \perp \boldsymbol{r}$: $\Re\lambda = \boldsymbol{x}_0^*(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{x}_0 + \boldsymbol{r}^*(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{r}$. Substituting $\boldsymbol{r}^*(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{r} = \lambda\|\boldsymbol{r}\|^2 + \boldsymbol{r}^*\boldsymbol{R}_a\boldsymbol{x}$, taken from the projected equation, and using $\boldsymbol{x}_0^*(\boldsymbol{A}-\boldsymbol{R}_s)\boldsymbol{x}_0 \geq (\lambda_1(\boldsymbol{A}-\boldsymbol{R}_s)-\epsilon)\|\boldsymbol{x}_0\|^2$,

$$\Re\lambda\big(1-\|\boldsymbol{r}\|^2\big) \;\geq\; \big(\lambda_1(\boldsymbol{A}-\boldsymbol{R}_s)-\epsilon\big)\big(1-\|\boldsymbol{r}\|^2\big) - \|\boldsymbol{r}\|\,\|\boldsymbol{R}_a\|_2,$$

so with (97) and $1-\|\boldsymbol{r}\|^2 \geq 3/4$, $\Re\lambda \geq \lambda_1(\boldsymbol{A}-\boldsymbol{R}_s) - \epsilon - \frac{8}{3}\|\boldsymbol{R}_a\|_2^2/\Gamma_c$, which implies the first half of (93). For the imaginary part, $\Im\lambda = -\Im(\boldsymbol{x}^*\boldsymbol{R}_a\boldsymbol{x})$ and $\boldsymbol{x}^*\boldsymbol{R}_a\boldsymbol{x} - \boldsymbol{x}_0^*\boldsymbol{R}_a\boldsymbol{x}_0 = \boldsymbol{x}_0^*\boldsymbol{R}_a\boldsymbol{r} + \boldsymbol{r}^*\boldsymbol{R}_a\boldsymbol{x}_0 + \boldsymbol{r}^*\boldsymbol{R}_a\boldsymbol{r}$, whence $|\Im\lambda| \leq \|\boldsymbol{P}\boldsymbol{R}_a\boldsymbol{P}\|_2 + \big(2\|\boldsymbol{r}\| + \|\boldsymbol{r}\|^2\big)\|\boldsymbol{R}_a\|_2 \leq \|\boldsymbol{P}\boldsymbol{R}_a\boldsymbol{P}\|_2 + 5\|\boldsymbol{R}_a\|_2^2/\Gamma_c$, using $\|\boldsymbol{r}\|^2 \leq \|\boldsymbol{r}\|/2$. Combining the lower bound on $\Re\lambda$ with (91) yields the second term in Δ_c and hence (13). □

E.3 Zeroth order: Propositions C.10–C.16, Theorems C.20, C.23 and C.14

Proof of Proposition C.10. Condition on $\boldsymbol{e}_t$ and expand $(\boldsymbol{I}-\eta\boldsymbol{J}(\boldsymbol{u}_t))\boldsymbol{e}_t\boldsymbol{e}_t^\top(\boldsymbol{I}-\eta\boldsymbol{J}(\boldsymbol{u}_t))^\top$. Independence of $\boldsymbol{u}_t$ and $\boldsymbol{e}_t$ yields (63) after expectation. Since $\mathcal{T}_\eta$ preserves the PSD cone and $\mathbb{E}\|\boldsymbol{e}_t\|^2 = \mathrm{Tr}(\boldsymbol{V}_t)$, decay of second moments for every initial covariance is equivalent to spectral radius below one on the symmetric covariance space. □

Proof of Lemma C.6. $n^{-2}\sum_{ij} c_{ij} = \mathbb{E}[(n^{-1}\sum_i a_i)^2] \geq (\mathbb{E}[n^{-1}\sum_i a_i])^2 = \bar{p}^{\,2}$ by Jensen, giving $\varsigma \geq 1$. For the upper endpoint, $c_{ij} \leq \sqrt{p_i p_j}$, so $n^{-2}\sum_{ij} c_{ij} \leq (n^{-1}\sum_i \sqrt{p_i})^2 \leq \bar{p}$ by Cauchy–Schwarz, i.e. $\varsigma \leq 1/\bar{p}$; batch clipping has $c_{ij} = \mathbb{E}[a] = \bar{p}$ for all i,j and attains it. For the last claim, put $\bar{a} = n^{-1}\sum_i a_i$, so $\boldsymbol{B} = \bar{a}\boldsymbol{H}$, $\boldsymbol{K} = \bar{p}\boldsymbol{H}$ and, by independence of $\bar{a}$ and $\boldsymbol{u}$, $\mathcal{Q}(\boldsymbol{W}) = \mathbb{E}[\bar{a}^2](2\boldsymbol{H}\boldsymbol{W}\boldsymbol{H} + \mathrm{Tr}(\boldsymbol{H}\boldsymbol{W}\boldsymbol{H})\boldsymbol{I})$ with $\mathbb{E}[\bar{a}^2] = \varsigma\bar{p}^{\,2}$. Diagonalizing $\boldsymbol{H}$ and writing $c = \varsigma\bar{p}^{\,2}$, the fixed-point condition of $\mathcal{T}_\eta$ gives $w_j = \eta c T/[2\lambda_j(\bar{p}-\eta c\lambda_j)]$ with $T = \sum_j \lambda_j^2 w_j$, hence $\frac{\eta c}{2}\sum_j \frac{\lambda_j}{\bar{p}-\eta c\lambda_j} = 1$. When $\eta c\lambda_j \ll \bar{p}$ for all j, this is $\eta c\tau/(2\bar{p}) = 1$, i.e. (59). Setting $\varsigma = 1/\bar{p}$ gives $\eta^\star \simeq 2/\tau$, free of $\bar{p}$. □

Proof of Proposition C.5. Apply the first identity of (54) together with the second-moment formula of Proposition C.16 applied to each slab $\{|\boldsymbol{u}^\top \boldsymbol{g}_i| < C\}$ separately, which is legitimate because that formula involves one gate only. □

Proof of Corollary C.15. In the isotropic model $\operatorname{Tr} \boldsymbol{K} = \bar{p}\tau$, and (59) gives $\tau \simeq 2/(\eta\varsigma\bar{p})$ at $\eta = \eta^\star$. Multiply. □

Proof of Proposition C.16. Decompose $\boldsymbol{u} = z\widehat{\boldsymbol{g}} + \boldsymbol{u}_\perp$ with $z = \boldsymbol{u}^\top \widehat{\boldsymbol{g}} \sim \mathcal{N}(0,1)$ independent of $\boldsymbol{u}_\perp \sim \mathcal{N}(0, \boldsymbol{Q})$ supported on $\widehat{\boldsymbol{g}}^\perp$. The gate depends only on z. For the second moment, $\mathbb{E}[a\,\boldsymbol{u}\boldsymbol{u}^\top] = \mathbb{E}[az^2]\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + \mathbb{E}[a]\,\mathbb{E}[\boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top] = m_2 \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + p\boldsymbol{Q}$, the cross terms vanishing by symmetry of $\boldsymbol{u}_\perp$.

For the fourth moment, use $a^2 = a$ so that $\mathcal{Q}(\boldsymbol{W}) = \mathbb{E}[a\,(\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top]$ with $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}$. Expanding $\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u} = z^2 m + 2z\boldsymbol{b}^\top \boldsymbol{u}_\perp + \boldsymbol{u}_\perp^\top \boldsymbol{M}_\perp \boldsymbol{u}_\perp$ and $\boldsymbol{u}\boldsymbol{u}^\top = z^2 \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + z(\widehat{\boldsymbol{g}}\boldsymbol{u}_\perp^\top + \boldsymbol{u}_\perp \widehat{\boldsymbol{g}}^\top) + \boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top$, every product with an odd power of z or an odd power of $\boldsymbol{u}_\perp$ has zero expectation. The five surviving products are

$$\begin{aligned}
\mathbb{E}[az^4]\, m\, \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top &= m_4 m\, \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top, \\
\mathbb{E}[az^2]\, m\, \mathbb{E}[\boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top] &= m_2 m \boldsymbol{Q}, \\
2\mathbb{E}[az^2]\, \mathbb{E}\big[\boldsymbol{b}^\top \boldsymbol{u}_\perp (\widehat{\boldsymbol{g}}\boldsymbol{u}_\perp^\top + \boldsymbol{u}_\perp \widehat{\boldsymbol{g}}^\top)\big] &= 2m_2(\widehat{\boldsymbol{g}}\boldsymbol{b}^\top + \boldsymbol{b}\widehat{\boldsymbol{g}}^\top), \\
\mathbb{E}[az^2]\, \mathbb{E}[\boldsymbol{u}_\perp^\top \boldsymbol{M}_\perp \boldsymbol{u}_\perp]\, \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top &= m_2 \operatorname{Tr}(\boldsymbol{M}_\perp) \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top,
\end{aligned}$$

and $\mathbb{E}[a]\,\mathbb{E}[(\boldsymbol{u}_\perp^\top \boldsymbol{M}_\perp \boldsymbol{u}_\perp)\boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top] = p(2\boldsymbol{M}_\perp + \operatorname{Tr}(\boldsymbol{M}_\perp)\boldsymbol{Q})$ by the Gaussian fourth-moment identity applied on $\widehat{\boldsymbol{g}}^\perp$. Summing gives (77). The moment formulas (75) follow from $\phi'(z) = -z\phi(z)$ by two integrations by parts. Setting $(p, m_2, m_4) = (1, 1, 3)$ and using $\boldsymbol{M} = m\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + \widehat{\boldsymbol{g}}\boldsymbol{b}^\top + \boldsymbol{b}\widehat{\boldsymbol{g}}^\top + \boldsymbol{M}_\perp$, $\operatorname{Tr} \boldsymbol{M} = m + \operatorname{Tr} \boldsymbol{M}_\perp$, reproduces $2\boldsymbol{M} + \operatorname{Tr}(\boldsymbol{M})\boldsymbol{I}$ term by term. □

Proof of Corollary C.11. $\boldsymbol{K} = \boldsymbol{\Sigma}_a \boldsymbol{H}$ with $\boldsymbol{\Sigma}_a = p\boldsymbol{I} + (m_2 - p)\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$, so $\boldsymbol{K} - \boldsymbol{K}^\top = (m_2 - p)[\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top \boldsymbol{H} - \boldsymbol{H}\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top]$. Writing $\boldsymbol{h} = \boldsymbol{H}\widehat{\boldsymbol{g}}$ and $\boldsymbol{h}_\perp = \boldsymbol{Q}\boldsymbol{h}$, the bracket equals $\widehat{\boldsymbol{g}}\boldsymbol{h}_\perp^\top - \boldsymbol{h}_\perp \widehat{\boldsymbol{g}}^\top$, a skew rank-two matrix with both nonzero singular values equal to $\|\boldsymbol{h}_\perp\|$. Finally $p - m_2 = 2\kappa\phi(\kappa)$ by (75). □

Proof of Theorem C.20. Under Assumption C.19, $\widehat{\boldsymbol{g}} = \boldsymbol{v}_r$ is an eigenvector of $\boldsymbol{H}$, so $\boldsymbol{\Sigma}_a$ and $\boldsymbol{H}$ commute and $\boldsymbol{K} = m_2 \lambda_r \boldsymbol{v}_r \boldsymbol{v}_r^\top + p\sum_{j \neq r} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$ is symmetric. For $\boldsymbol{W}$ diagonal in $\{\boldsymbol{v}_j\}$ we get $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{W}\boldsymbol{H} = \sum_j \lambda_j^2 w_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$, hence $m = \lambda_r^2 w_r$, $b = 0$ and $\boldsymbol{M}_\perp = \sum_{j \neq r} \lambda_j^2 w_j \boldsymbol{v}_j \boldsymbol{v}_j^\top$ with $\operatorname{Tr} \boldsymbol{M}_\perp = T_\perp$. Substituting into (77) shows that $\mathcal{Q}(\boldsymbol{W})$ is again diagonal, so $\mathcal{T}_\eta$ preserves the diagonal subspace and (63) reduces to (79)–(80).

It remains to justify that this invariant diagonal subspace determines the spectral radius of the full covariance operator. The operator $\mathcal{T}_\eta$ preserves the PSD cone, so its spectral radius has a nonzero PSD eigenmatrix $\boldsymbol{V} \succeq 0$. For every diagonal sign matrix $\boldsymbol{S}$ in the eigenbasis $\{\boldsymbol{v}_j\}$, Gaussian sign symmetry and the fact that the gate depends only on $|\boldsymbol{u}^\top \boldsymbol{v}_r|$ give $\mathcal{T}_\eta(\boldsymbol{S}\boldsymbol{V}\boldsymbol{S}) = \boldsymbol{S}\mathcal{T}_\eta(\boldsymbol{V})\boldsymbol{S}$. Averaging $\boldsymbol{S}\boldsymbol{V}\boldsymbol{S}$ over all sign matrices removes every off-diagonal entry while preserving its positive trace. The resulting nonzero diagonal PSD matrix is therefore an eigenmatrix at the same spectral radius. Thus the diagonal recursion captures the full Perron root on $\operatorname{range}(\boldsymbol{H})$.

At the mean-square boundary the Perron vector is a fixed point. Put $F = m_2 \lambda_r^2 w_r + pT_\perp$. From (80), $2\eta p \lambda_j (1 - \eta\lambda_j) w_j = \eta^2 F$, so $w_j = \eta F/[2p\lambda_j(1 - \eta\lambda_j)]$ and

$$T_\perp = \sum_{j \neq r} \lambda_j^2 w_j = \frac{\eta F}{2p} \sum_{j \neq r} \frac{\lambda_j}{1 - \eta\lambda_j}. \tag{98}$$

From (79), $\eta\lambda_r(2m_2 - \eta m_4 \lambda_r) w_r = \eta^2 m_2 T_\perp$, so $w_r = \eta m_2 T_\perp/[\lambda_r(2m_2 - \eta m_4 \lambda_r)]$. Substituting into F gives $F = T_\perp[p + \eta m_2^2 \lambda_r/(2m_2 - \eta m_4 \lambda_r)]$, and inserting this in (98) and cancelling $T_\perp > 0$ yields (81). Since $\tau_\perp > 0$, the left-hand side starts at zero and diverges before the first transverse or radial admissibility pole, so a root exists. On the admissible range each summand $\lambda_j/(1 - \eta\lambda_j)$ is positive and increasing in η, and $r(\eta)$ has increasing numerator and decreasing positive denominator, so Ξ is strictly increasing and the root is unique. □

Proof of Proposition C.21. Take $\boldsymbol{W} \in \mathcal{D}$, so $\boldsymbol{W} = w\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + \boldsymbol{W}_\perp$ with $\boldsymbol{W}_\perp = \boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}$. With $\boldsymbol{H}_\mathrm{a}$ block diagonal, $\boldsymbol{H}_\mathrm{a}\boldsymbol{W}\boldsymbol{H}_\mathrm{a}$ and $\boldsymbol{\Sigma}_a\boldsymbol{H}_\mathrm{a}\boldsymbol{W} + \boldsymbol{W}\boldsymbol{H}_\mathrm{a}\boldsymbol{\Sigma}_a$ are block diagonal, and in (77) the cross block $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{M}\widehat{\boldsymbol{g}}$ then vanishes, so $\mathcal{T}_\eta[\boldsymbol{H}_\mathrm{a}]\boldsymbol{W} \in \mathcal{D}$; the same computation on $\boldsymbol{W} \in \mathcal{D}^\perp$ (that is, $\boldsymbol{W} = \widehat{\boldsymbol{g}}\boldsymbol{c}^\top + \boldsymbol{c}\widehat{\boldsymbol{g}}^\top$ with $\boldsymbol{c} \perp \widehat{\boldsymbol{g}}$) gives a matrix in $\mathcal{D}^\perp$. For the linear term, note $\boldsymbol{\Sigma}_a\boldsymbol{E} = m_2\widehat{\boldsymbol{g}}\boldsymbol{h}_\perp^\top + p\boldsymbol{h}_\perp\widehat{\boldsymbol{g}}^\top$, so for $\boldsymbol{W} \in \mathcal{D}$

$$(\boldsymbol{\Sigma}_a E)\boldsymbol{W} = m_2\widehat{\boldsymbol{g}}(\boldsymbol{W}_\perp\boldsymbol{h}_\perp)^\top + p\,w\,\boldsymbol{h}_\perp\widehat{\boldsymbol{g}}^\top,$$

which lies in $\mathcal{D}^\perp$, and likewise for $\boldsymbol{W}\boldsymbol{E}\boldsymbol{\Sigma}_a$. The first-order change of $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}$ is $\delta\boldsymbol{M} = \boldsymbol{E}\boldsymbol{W}\boldsymbol{H}_\mathrm{a} + \boldsymbol{H}_\mathrm{a}\boldsymbol{W}\boldsymbol{E}$, which is again purely off block diagonal; hence in (77) the scalars $m = \widehat{\boldsymbol{g}}^\top\boldsymbol{M}\widehat{\boldsymbol{g}}$ and $\operatorname{Tr}\boldsymbol{M}_\perp$ are unchanged at this order while $\boldsymbol{b}$ becomes nonzero, so the $O(\boldsymbol{E})$ image of $\mathcal{D}$ is spanned by $\widehat{\boldsymbol{g}}\boldsymbol{b}^\top + \boldsymbol{b}\widehat{\boldsymbol{g}}^\top \in \mathcal{D}^\perp$. The same argument with the roles exchanged sends $\mathcal{D}^\perp$ into $\mathcal{D}$. The unperturbed leading eigenvector lies in $\mathcal{D}$: $\mathcal{T}_\eta[\boldsymbol{H}_\mathrm{a}]$ preserves the positive semidefinite cone, so by Krein–Rutman its Perron vector is positive semidefinite, and no nonzero positive semidefinite matrix lies in $\mathcal{D}^\perp$, whose elements have vanishing diagonal blocks. Hence the first-order Rayleigh–Schrödinger correction $\langle \boldsymbol{L}, (D\mathcal{T}\cdot\boldsymbol{E})\boldsymbol{R}\rangle / \langle \boldsymbol{L}, \boldsymbol{R}\rangle$ pairs a matrix in $\mathcal{D}$ with a matrix in $\mathcal{D}^\perp$ and vanishes, giving (84). □

Proof of Theorem C.23. Under Assumption C.22, $\widehat{\boldsymbol{h}}$ is an eigenvector of $\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}$ with eigenvalue λ_h and $\boldsymbol{h}_\perp$ has no component on $\mathcal{P}^\perp$. Block $\boldsymbol{W}$ as $w\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + (\widehat{\boldsymbol{g}}\boldsymbol{c}^\top + \boldsymbol{c}\widehat{\boldsymbol{g}}^\top) + \boldsymbol{W}_\perp$ with $\boldsymbol{c} = \boldsymbol{Q}\boldsymbol{W}\widehat{\boldsymbol{g}}$. Using $\boldsymbol{\Sigma}_a\boldsymbol{H} = m_2\lambda_r\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + m_2\widehat{\boldsymbol{g}}\boldsymbol{h}_\perp^\top + p\boldsymbol{h}_\perp\widehat{\boldsymbol{g}}^\top + p\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}$, the drift $\boldsymbol{\Sigma}_a\boldsymbol{H}\boldsymbol{W} + \boldsymbol{W}\boldsymbol{H}\boldsymbol{\Sigma}_a$ has blocks

$$2m_2(\lambda_r w + \boldsymbol{h}_\perp^\top\boldsymbol{c}), \qquad p\,w\,\boldsymbol{h}_\perp + p\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}\,\boldsymbol{c} + m_2\lambda_r\boldsymbol{c} + m_2\boldsymbol{W}_\perp\boldsymbol{h}_\perp, \qquad p(\boldsymbol{h}_\perp\boldsymbol{c}^\top + \boldsymbol{c}\boldsymbol{h}_\perp^\top) + p(\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}\,\boldsymbol{W}_\perp + \boldsymbol{W}_\perp\boldsymbol{Q}\boldsymbol{H}\boldsymbol{Q}),$$

and $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}$ has m, b and $\boldsymbol{M}_\perp$ as in (85) after substituting $\boldsymbol{c} = y\widehat{\boldsymbol{h}}$, $\boldsymbol{W}_\perp\widehat{\boldsymbol{h}} = \omega\widehat{\boldsymbol{h}}$. Substituting these into (77) and reading off the $\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{h}}\widehat{\boldsymbol{g}}$, $\widehat{\boldsymbol{h}}\widehat{\boldsymbol{h}}$ and $\boldsymbol{v}_j\boldsymbol{v}_j$ entries gives (86)–(88); the subspace spanned by these coordinates is invariant because every term generated is proportional to $\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$, $\widehat{\boldsymbol{h}}\widehat{\boldsymbol{g}}^\top$, $\widehat{\boldsymbol{h}}\widehat{\boldsymbol{h}}^\top$ or $\boldsymbol{v}_j\boldsymbol{v}_j^\top$.

That subspace also carries the spectral radius, by the sign-symmetry argument used for Theorem C.20 with the plane held fixed. Let $\boldsymbol{S}$ act as the identity on $\mathcal{P}$ and as ± 1 on each $\boldsymbol{v}_j$. Then $\boldsymbol{S}$ commutes with $\boldsymbol{H}$ and with $\boldsymbol{\Sigma}_a$, because both are block diagonal for $\mathcal{P} \oplus \mathcal{P}^\perp$ and diagonal in $\{\boldsymbol{v}_j\}$ on $\mathcal{P}^\perp$, and the gate depends on $\boldsymbol{u}$ only through $\boldsymbol{u}^\top\widehat{\boldsymbol{g}}$, which $\boldsymbol{S}$ fixes; hence $\mathcal{T}_\eta(\boldsymbol{S}\boldsymbol{V}\boldsymbol{S}) = \boldsymbol{S}\mathcal{T}_\eta(\boldsymbol{V})\boldsymbol{S}$. Averaging $\boldsymbol{S}\boldsymbol{V}\boldsymbol{S}$ over all such $\boldsymbol{S}$ removes every entry coupling $\mathcal{P}$ to $\mathcal{P}^\perp$ and every entry coupling two distinct $\boldsymbol{v}_j$, while leaving the plane block and the tail diagonal — that is, exactly the coordinates (w, y, ω, w_j) — and preserving the trace. Since $\mathcal{T}_\eta$ preserves the positive semidefinite cone its spectral radius is attained at a nonzero $\boldsymbol{V} \succeq 0$, whose positive trace survives the averaging, so the average is a nonzero eigenmatrix at the same spectral radius supported on those coordinates.

At the boundary the Perron vector is a fixed point (Krein–Rutman applies since $\mathcal{T}_\eta$ preserves the positive semidefinite cone). Put $F = m_2 m + p(P + T_\perp)$. From (88), $2\eta p\lambda_j(1 - \eta\lambda_j)w_j = \eta^2 F$, so $T_\perp = \frac{\eta F}{2p}\sum_j \frac{\lambda_j}{1-\eta\lambda_j}$ exactly as in (98). The remaining three stationary equations are linear and homogeneous in (w, y, ω) with the tail entering only through $T_\perp$, so scaling $T_\perp = 1$ and solving the 3×3 system gives $(\widetilde{w}, \widetilde{y}, \widetilde{\omega})$ and $F = p(1 + r_\mathrm{c}(\eta))T_\perp$ with r_c as stated. Substituting into the expression for $T_\perp$ and cancelling $T_\perp > 0$ yields (89). If $\xi = 0$ the y equation decouples and forces $\widetilde{y} = 0$, the ω equation becomes the tail equation with $\lambda = \lambda_h$, and r_c collapses to r of (81) with λ_h restored to the sum. Unlike Ξ, Ξ_c need not be monotone, since the 3×3 determinant can vanish; the boundary is the smallest positive root. □

Proof of Proposition C.18. (x_1, x_2) are standard normal in the stated basis and $\boldsymbol{u}^\top\widehat{\boldsymbol{g}}_i = x_1$, $\boldsymbol{u}^\top\widehat{\boldsymbol{g}}_j = \rho x_1 + s x_2$, so R is as claimed and depends on $\boldsymbol{u}$ only through (x_1, x_2). Fubini in x_2 at fixed x_1 gives (78). The recursion for I_b follows from $\phi'(x) = -x\phi(x)$: integrating $\frac{d}{dx}(-x^{b-1}\phi) = -(b-1)x^{b-2}\phi + x^b\phi$ over $[\ell, \upsilon]$ rearranges to the stated identity. If $|\rho| = 1$ the two slabs are nested, $R = \{|x_1| < \min(\kappa_i, \kappa_j)\}$, and x_2 is unconstrained, so the moments factor into I_a times the free Gaussian moments $1, 0, 1, 0, 3$. □

Proof of Theorem C.14. By Lemma C.1 with $\boldsymbol{H}_i \equiv \boldsymbol{H}$, $\mathcal{Q}_\mathrm{ps}(\boldsymbol{W}) = n^{-2}\sum_{ij}\mathbb{E}[a_i a_j(\boldsymbol{u}^\top\boldsymbol{M}\boldsymbol{u})\boldsymbol{u}\boldsymbol{u}^\top]$, so it suffices to evaluate one pair. Split $\boldsymbol{u} = \boldsymbol{u}_\mathcal{P} + \boldsymbol{u}_\perp$ with $\boldsymbol{u}_\mathcal{P}$ standard normal on $\mathcal{P}_{ij}$ and $\boldsymbol{u}_\perp$ independent standard normal on $\mathcal{P}_{ij}^\perp$; the gates

depend only on $\boldsymbol{u}_{\mathcal{P}}$. Expanding $\boldsymbol{u}^\top \boldsymbol{M}\boldsymbol{u} = \boldsymbol{u}_{\mathcal{P}}^\top \boldsymbol{M}_{\mathcal{P}}\boldsymbol{u}_{\mathcal{P}} + 2\boldsymbol{u}_{\mathcal{P}}^\top \boldsymbol{M}_\times \boldsymbol{u}_\perp + \boldsymbol{u}_\perp^\top \boldsymbol{M}_\perp \boldsymbol{u}_\perp$ and $\boldsymbol{u}\boldsymbol{u}^\top = \boldsymbol{u}_{\mathcal{P}}\boldsymbol{u}_{\mathcal{P}}^\top + \boldsymbol{u}_{\mathcal{P}}\boldsymbol{u}_\perp^\top + \boldsymbol{u}_\perp \boldsymbol{u}_{\mathcal{P}}^\top + \boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top$, every product with an odd power of $\boldsymbol{u}_\perp$ has zero expectation. The five surviving products are $\mathbb{E}[1_R(\boldsymbol{u}_{\mathcal{P}}^\top \boldsymbol{M}_{\mathcal{P}}\boldsymbol{u}_{\mathcal{P}})\boldsymbol{u}_{\mathcal{P}}\boldsymbol{u}_{\mathcal{P}}^\top] = T^4_{ij} \cdot \boldsymbol{M}_{\mathcal{P}}$ on $\mathcal{P}$; $\mathbb{E}[1_R(\boldsymbol{u}_\perp^\top \boldsymbol{M}_\perp \boldsymbol{u}_\perp)\boldsymbol{u}_{\mathcal{P}}\boldsymbol{u}_{\mathcal{P}}^\top] = \operatorname{Tr}(\boldsymbol{M}_\perp)\boldsymbol{S}_{ij}$ on $\mathcal{P}$; $2\mathbb{E}[1_R(\boldsymbol{u}_{\mathcal{P}}^\top \boldsymbol{M}_\times \boldsymbol{u}_\perp)\boldsymbol{u}_{\mathcal{P}}\boldsymbol{u}_\perp^\top] = 2\boldsymbol{S}_{ij}\boldsymbol{M}_\times$ on the cross block; $\mathbb{E}[1_R(\boldsymbol{u}_{\mathcal{P}}^\top \boldsymbol{M}_{\mathcal{P}}\boldsymbol{u}_{\mathcal{P}})]\mathbb{E}[\boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top] = \langle \boldsymbol{M}_{\mathcal{P}}, \boldsymbol{S}_{ij}\rangle$ on $\mathcal{P}^\perp$; and $G_{ij}\mathbb{E}[(\boldsymbol{u}_\perp^\top \boldsymbol{M}_\perp \boldsymbol{u}_\perp)\boldsymbol{u}_\perp \boldsymbol{u}_\perp^\top] = G_{ij}(2\boldsymbol{M}_\perp + \operatorname{Tr}(\boldsymbol{M}_\perp)\boldsymbol{\Pi}_{ij})$ on $\mathcal{P}^\perp$ by the Gaussian fourth-moment identity. Writing $\operatorname{Tr}\boldsymbol{M}_\perp = \operatorname{Tr}\boldsymbol{M} - \operatorname{Tr}\boldsymbol{M}_{\mathcal{P}}$ and embedding each block with $\boldsymbol{E}_{ij}$ and $\boldsymbol{\Pi}_{ij}$ gives (66). For $i = j$ the plane is spanned by $\widehat{\boldsymbol{g}}_i$ and an arbitrary orthogonal direction whose Gaussian coordinate is unconstrained, so $G = p$, $\boldsymbol{S} = \operatorname{diag}(m_2, p)$ and T^4 has entries m_4, m_2, p according to the number of second-coordinate factors; substituting reproduces (77) term by term. □

Proof of Corollary C.26. Write $\Lambda_\eta^\perp = \sum_{j\neq r} \lambda_j/(1-\eta\lambda_j)$ for the transverse counterpart of the unclipped sum in Eq. (22). With $(p, m_2, m_4) = (1, 1, 3)$, $\Xi(\eta) = 1$ reads $\frac{\eta\Lambda_\eta^\perp}{2} \cdot \frac{2-2\eta\lambda_r}{2-3\eta\lambda_r} = 1$, i.e. $\eta\Lambda_\eta^\perp = (2-3\eta\lambda_r)/(1-\eta\lambda_r)$. The ungated boundary $\frac{\eta}{2}\sum_j \lambda_j/(1-\eta\lambda_j) = 1$ is $\frac{\eta\Lambda_\eta^\perp}{2} = 1 - \frac{\eta\lambda_r}{2(1-\eta\lambda_r)} = \frac{2-3\eta\lambda_r}{2(1-\eta\lambda_r)}$, the same equation. □

Proof of Corollary C.27. As $\kappa \to 0$, $p = 2\kappa\phi(0) + O(\kappa^3)$, $m_2 = \frac{2}{3}\kappa^3\phi(0) + O(\kappa^5)$ and $m_4 = \frac{2}{5}\kappa^5\phi(0) + O(\kappa^7)$, so $m_2/p = \kappa^2/3 + O(\kappa^4)$ and $m_4/m_2 = \frac{3}{5}\kappa^2 + O(\kappa^4)$. Hence $r(\eta) = \frac{\eta\lambda_r(m_2/p)}{2-\eta\lambda_r m_4/m_2} = \frac{\eta\lambda_r\kappa^2}{6} + O(\kappa^4)$. □

Proof of Theorem 4.1 and Corollary C.4. Since $r \geq 0$ and $\lambda_j/(1-\eta\lambda_j) \geq \lambda_j$, the boundary equation gives $1 \geq \frac{\eta}{2}\tau_\perp$ at $\eta = \eta^\star_{\mathrm{ms}}$, i.e. $\eta^\star_{\mathrm{ms}} \leq 2/\tau_\perp$. Conversely $\lambda_j/(1-\eta\lambda_j) \leq \lambda_j/(1-\eta\lambda^\perp_{\max})$ and $1 + r(\eta) \leq \bar{c}$ on $\eta \leq 2/\tau_\perp$ by monotonicity of r, so $1 \leq \frac{\eta\bar{c}\tau_\perp}{2(1-\eta\lambda^\perp_{\max})}$, i.e. $2 - 2\eta\lambda^\perp_{\max} \leq \eta\bar{c}\tau_\perp$, which is the lower bound in (16). Inverting (16) at $\eta = \eta^\star_{\mathrm{ms}}$ and using $\lambda^\perp_{\max} \leq \varepsilon_{\mathrm{rank}}\tau_\perp$ gives (56).

For (82), Assumption C.19 has $\boldsymbol{H} \succeq 0$, so $0 \leq \widehat{\boldsymbol{g}}^\top \boldsymbol{H}\widehat{\boldsymbol{g}} \leq \lambda_1$ and therefore $\operatorname{Tr}\boldsymbol{H} - \lambda_1 \leq \tau_\perp \leq \operatorname{Tr}\boldsymbol{H}$; and $\lambda^\perp_{\max}$ is the largest eigenvalue of $\boldsymbol{H}$ restricted to $\widehat{\boldsymbol{g}}^\perp$, hence at most λ_1. Substituting the upper inclusion into the upper bound of (16) and the lower inclusions into its lower bound gives (82), and rearranging $\eta = \eta^\star_{\mathrm{ms}}$ gives (83). Both substitutions only widen the band, so (82) is implied by, and never tighter than, (16). Writing $\varepsilon = \lambda_1/\operatorname{Tr}(\boldsymbol{H})$, the ratio of the two sides of (82) is $(\bar{c} + 2\varepsilon)/(1-\varepsilon) = \bar{c} + (2+\bar{c})\varepsilon + O(\varepsilon^2)$, which is $1 + 3\varepsilon + O(\varepsilon^2)$ at $\bar{c} = 1$. □

Derivation of Eqs. (18)–(19). For (18), $z = \boldsymbol{g}^\top \boldsymbol{u} \sim \mathcal{N}(0, \|\boldsymbol{g}\|^2)$ and clip_C is Lipschitz, so Stein's lemma gives, coordinatewise, $\mathbb{E}[\mathrm{clip}_C(z)u_i] = \mathbb{E}[\partial_{u_i}\mathrm{clip}_C(z)] = \mathbb{E}[\mathrm{clip}'_C(z)][\boldsymbol{g}]_i$, and $\mathbb{E}[\mathrm{clip}'_C(z)] = \Pr(|z| < C) = p(\kappa)$. For (19), split $\boldsymbol{u} = \zeta\widehat{\boldsymbol{g}} + \boldsymbol{u}_\perp$ with $\zeta = \widehat{\boldsymbol{g}}^\top \boldsymbol{u} \sim \mathcal{N}(0,1)$ and $\boldsymbol{u}_\perp = (\boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top)\boldsymbol{u}$ independent of ζ, so $z = \|\boldsymbol{g}\|\zeta$ and $\mathrm{clip}_C(z)^2 = \|\boldsymbol{g}\|^2\min(\zeta^2, \kappa^2)$. The cross terms are odd in $\boldsymbol{u}_\perp$ and vanish, leaving $\mathbb{E}[\min(\zeta^2,\kappa^2)\zeta^2]\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top + \mathbb{E}[\min(\zeta^2,\kappa^2)]\mathbb{E}[\boldsymbol{u}_\perp\boldsymbol{u}_\perp^\top]$, which is the claim since $\mathbb{E}[\boldsymbol{u}_\perp\boldsymbol{u}_\perp^\top] = \boldsymbol{I} - \widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$. The closed forms follow by splitting $\min(\zeta^2,\kappa^2)$ at $|\zeta| = \kappa$ and using (75).

Proof of Theorem 4.3. Write $z = \boldsymbol{u}^\top \boldsymbol{H}\boldsymbol{\delta}$ and $\boldsymbol{W}_t = \mathbb{E}[\boldsymbol{\delta}_t\boldsymbol{\delta}_t^\top]$. Squaring (17),

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta\,\mathbb{E}\big[\mathrm{clip}_C(z)(\boldsymbol{u}\boldsymbol{\delta}^\top + \boldsymbol{\delta}\boldsymbol{u}^\top)\big] + \eta^2\,\mathbb{E}\big[\mathrm{clip}_C(z)^2\boldsymbol{u}\boldsymbol{u}^\top\big].$$

Condition on $\boldsymbol{\delta}$ and apply Eqs. (18)–(19) with $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{\delta}$, so $\|\boldsymbol{g}\| = r$ and $\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top r^2 = \boldsymbol{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{H}$; the conditional diffusion is $\mu_2(r)r^2\boldsymbol{I} + (\mu_4 - \mu_2)(r)\boldsymbol{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top\boldsymbol{H}$ and the conditional drift carries $p(r)\boldsymbol{H}\boldsymbol{\delta}\boldsymbol{\delta}^\top$. Applying the three identities of (27) term by term,

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta p\,(\boldsymbol{H}\boldsymbol{W}_t + \boldsymbol{W}_t\boldsymbol{H}) + \eta^2\big[\mu_2\operatorname{Tr}(\boldsymbol{H}^2\boldsymbol{W}_t)\,\boldsymbol{I} + (\mu_4 - \mu_2)\,\boldsymbol{H}\boldsymbol{W}_t\boldsymbol{H}\big].$$

Diagonalise in the eigenbasis of $\boldsymbol{H}$. The entries of $\boldsymbol{W}_t$ off the diagonal decouple, $\boldsymbol{W}_{jk}^{t+1} = f(\lambda_j,\lambda_k)\boldsymbol{W}_{jk}^t$ with $f(x,y) = 1 - \eta p(x+y) + \eta^2(\mu_4-\mu_2)xy$, and $f(x,y) \leq \frac{1}{2}\big(f(x,x) + f(y,y)\big)$ because $f(x,y) - \frac{1}{2}(f(x,x)+f(y,y)) = -\frac{1}{2}\eta^2(\mu_4-\mu_2)(x-y)^2 \leq 0$; on the positive-real branch

of Appendix C.2 the diagonal block therefore decides criticality, and we work with it. Writing $w_j = \boldsymbol{W}_{jj}$ and $T = \sum_k \lambda_k^2 w_k$,

$$w_j^{t+1} = w_j^t\big(1 - 2\eta p\lambda_j + \eta^2(\mu_4 - \mu_2)\lambda_j^2\big) + \eta^2\mu_2 T.$$

No term is discarded. A stationary $w \succeq 0$, $w \neq 0$, requires

$$w_j\lambda_j\big(2\eta p - \eta^2(\mu_4-\mu_2)\lambda_j\big) = \eta^2\mu_2 T, \qquad \text{i.e.} \qquad w_j = \frac{\eta\mu_2 T}{2p\,\lambda_j\big(1-\beta\lambda_j\big)}, \qquad \beta = \frac{\eta(\mu_4-\mu_2)}{2p},$$

which is positive exactly on the admissible range $\beta\lambda_1 < 1$. Substituting into $T = \sum_j \lambda_j^2 w_j$ and cancelling $T > 0$ gives

$$1 = \frac{\eta\mu_2}{2p}\sum_j \frac{\lambda_j}{1-\beta\lambda_j}, \qquad \text{i.e.} \qquad \eta\,\tau = 2. \tag{99}$$

Since $\beta \geq 0$ and $\boldsymbol{H} \succeq 0$, each summand is at least λ_j and at most $\lambda_j/(1-\beta\lambda_1)$, which is the sandwich $\gamma\operatorname{Tr}(\boldsymbol{H}) \leq \tau \leq \gamma\operatorname{Tr}(\boldsymbol{H})/(1-\beta\lambda_1)$ and hence $0 \leq (\tau - \gamma\operatorname{Tr}(\boldsymbol{H}))/\tau \leq \beta\lambda_1$. The closed forms of the moments give $\mu_4 - 3\mu_2 = -2\kappa^2(1-p) \leq 0$, hence $(\mu_4-\mu_2)/(2\mu_2) \leq 1$. Therefore

$$\beta\lambda_1 = \eta\gamma\lambda_1\frac{\mu_4-\mu_2}{2\mu_2} \leq \eta\gamma\lambda_1 \leq 2\frac{\lambda_1}{\operatorname{Tr}(\boldsymbol{H})},$$

where the last step uses $\eta\gamma\operatorname{Tr}(\boldsymbol{H}) \leq \eta\tau = 2$. Combining the last two displays proves (23).

For ψ, differentiating $\mu_4 - \mu_2 = \mathbb{E}[\min(\zeta^2,\kappa^2)(\zeta^2-1)]$ in κ gives $\frac{d}{d\kappa}(\mu_4-\mu_2) = 2\kappa\,\mathbb{E}[(\zeta^2-1)\mathbf{1}\{|\zeta|>\kappa\}] = 4\kappa^2\phi(\kappa)$, using $\mathbb{E}[\zeta^2\mathbf{1}\{|\zeta|<\kappa\}] = p - 2\kappa\phi(\kappa)$. Both $\mu_4-\mu_2$ and p vanish at $\kappa = 0$, so

$$\mu_4(\kappa) - \mu_2(\kappa) = \int_0^\kappa 4s^2\phi(s)\,ds, \qquad p(\kappa) = \int_0^\kappa 2\phi(s)\,ds, \tag{100}$$

whose integrand ratio $2s^2$ is strictly increasing; by the monotone form of l'Hôpital's rule $\psi = (\mu_4-\mu_2)/p$ is strictly increasing, with $\psi(0^+) = 0$ and $\psi(\infty) = (3-1)/1 = 2$. Hence $\beta = \eta\psi/2$ increases strictly from 0 to η, and $\psi(\kappa) = \frac{2}{3}\kappa^2 + O(\kappa^4)$ from (100). Because γ and β are both strictly increasing and $\lambda_j \geq 0$, $\kappa \mapsto \eta\tau(\kappa)$ is strictly increasing on the admissible interval $\beta(\kappa)\lambda_1 < 1$. If $\eta\lambda_1 < 1$, this interval is $(0,\infty)$ and the endpoint value is $\eta\sum_j \lambda_j/(1-\eta\lambda_j)$. If $\eta\lambda_1 \geq 1$, the endpoint instead has $\beta\lambda_1 \uparrow 1$ and $\eta\tau \to \infty$. Since $\eta_0^\star$ is the unique root of (22), existence and uniqueness of $\kappa^\star$ in case (ii), and its absence in case (i), follow. For the stability claim in (i), work on $\operatorname{range}(\boldsymbol{H})$, where $\lambda_j > 0$. The diagonal block at $\boldsymbol{W} = 0$ is the $\kappa \to \infty$ limit $(p,\mu_2,\mu_4) = (1,1,3)$, a linear map with nonnegative matrix $\boldsymbol{M}_{jk} = \big(1 - 2\eta\lambda_j + 2\eta^2\lambda_j^2\big)\delta_{jk} + \eta^2\lambda_k^2$ (nonnegative because $1 - 2x + 3x^2 > 0$ for all real x). Write $\boldsymbol{I} - \boldsymbol{M} = \boldsymbol{D} - \eta^2\mathbf{1}\boldsymbol{v}^\top$ with $\boldsymbol{D} = \operatorname{diag}\big(2\eta\lambda_j(1-\eta\lambda_j)\big) \succ 0$ and $\boldsymbol{v}_k = \lambda_k^2$. Since $\boldsymbol{v}^\top\boldsymbol{D}^{-1}\mathbf{1} = \Lambda_\eta/(2\eta)$, where $\Lambda_\eta := \sum_j \lambda_j/(1-\eta\lambda_j)$, Sherman–Morrison gives

$$(\boldsymbol{I}-\boldsymbol{M})^{-1} = \boldsymbol{D}^{-1} + \frac{\eta^2\,\boldsymbol{D}^{-1}\mathbf{1}\,\boldsymbol{v}^\top\boldsymbol{D}^{-1}}{1-\eta\Lambda_\eta/2},$$

whose entries are all nonnegative exactly when $\eta\sum_j \lambda_j/(1-\eta\lambda_j) < 2$. For a nonnegative matrix, $\rho(\boldsymbol{M}) < 1$ iff $(\boldsymbol{I}-\boldsymbol{M})^{-1}$ exists and is entrywise nonnegative, so the second moment decays precisely on $\eta < \eta_0^\star$, and $\rho(\boldsymbol{M}) = 1$ at the mean-square stability boundary $\eta = \eta_0^\star$. For $\eta\lambda_1 \geq 1$, its $(1,1)$ entry $1 - 2\eta\lambda_1 + 3\eta^2\lambda_1^2$ exceeds one, so the origin is also mean-square unstable on that branch.

For the monotonicity of γ, write both moments as integrals from 0. Since $\min(\zeta^2,\kappa^2) = \int_0^{\kappa^2}\mathbf{1}\{|\zeta| > \sqrt{t}\}\,dt$, taking expectations and substituting $t = s^2$ gives

$$\mu_2(\kappa) = \int_0^\kappa 2s\big(1-p(s)\big)\,ds, \qquad p(\kappa) = \int_0^\kappa 2\phi(s)\,ds,$$

so in particular $\mu_2'(\kappa) = 2\kappa(1-p(\kappa))$. Writing $Q = 1-\Phi$ and $R = Q/\phi$ for the Mills ratio, $1 - p(s) = 2Q(s)$ and the ratio of the two integrands is

$$\frac{2s\big(1-p(s)\big)}{2\phi(s)} = 2s\,R(s), \qquad \frac{d}{ds}\big[sR(s)\big] = R(s)\big(1+s^2\big) - s > 0,$$

the last inequality being the standard Gaussian tail bound $Q(s) > s\phi(s)/(1+s^2)$ for $s > 0$; we used $R' = sR - 1$. The integrand ratio is therefore strictly increasing, and by the monotone form of l'Hôpital's rule the ratio of the integrals, $\gamma = \mu_2/p$, is strictly increasing on $(0, \infty)$. Its endpoints are direct: as $\kappa \to \infty$, $(p, \mu_2) \to (1, 1)$ so $\gamma \to 1$. As $\kappa \to 0$, write $a = \sqrt{2/\pi}$; then $p = a\big(\kappa - \frac{1}{6}\kappa^3\big) + O(\kappa^5)$ from $\phi(s) = \phi(0)(1 - s^2/2 + O(s^4))$, and $\mu_2 = \int_0^\kappa 2s(1 - p(s))\, ds = \kappa^2 - \frac{2a}{3}\kappa^3 + O(\kappa^5)$. Dividing,

$$\gamma(\kappa) = \frac{\kappa}{a} \cdot \frac{1 - \frac{2a}{3}\kappa + O(\kappa^3)}{1 - \frac{1}{6}\kappa^2 + O(\kappa^4)} = \sqrt{\pi/2}\,\kappa - \tfrac{2}{3}\kappa^2 + O(\kappa^3), \tag{101}$$

which is (28); the $O(\kappa^2)$ term is not optional, and dropping it costs $O(\kappa)$ relative accuracy, not $O(\kappa^2)$. Hence γ is a strictly increasing bijection onto $(0, 1)$. In case (i) no $\kappa^\star$ exists, so no nonzero stationary state does; $\boldsymbol{W}_t \to 0$ locally, whence $\kappa \to \infty$, $\gamma \to 1$, $\beta \to \eta$, and the condition reduces to the ungated critical step (22). Finally, at the fixed point $\mathrm{Tr}(\boldsymbol{H}) = \gamma(\kappa^\star)^{-1}(2/\eta)(1 + O(\beta\lambda_1))$, and inverting (101), $\gamma^{-1} = \sqrt{2/\pi}\,\kappa^{-1}\big(1 + \frac{2}{3}\sqrt{2/\pi}\,\kappa + O(\kappa^2)\big) = \sqrt{2/\pi}\,\big(\|\boldsymbol{g}\|/C\big)\,\big(1 + O(\kappa)\big)$. □

Proof of Theorem 4.4. Work in an eigenbasis of $\boldsymbol{H}$ and set $x_j = \eta\lambda_j$. For a diagonal covariance write $w_j = \boldsymbol{W}_{jj}$, $y_j = \lambda_j^2 w_j$ and $T = \sum_j y_j$. Equation (20) becomes

$$y_j^+ = D_j(T) y_j + \mu_2(T) x_j^2 T, \qquad D_j(T) := 1 - 2p(T)x_j + (\mu_4 - \mu_2)(T)x_j^2. \tag{102}$$

Here and below a moment written as a function of T is evaluated at $\kappa(T) = C/\sqrt{T}$. Define the stationary energy shares

$$s_j(\kappa) := \frac{\mu_2 x_j^2}{1 - D_j} = \frac{\gamma(\kappa) x_j}{2\big(1 - \psi(\kappa)x_j/2\big)}, \qquad G(\kappa) := \sum_j s_j(\kappa). \tag{103}$$

The fixed-point equation is $G(\kappa^\star) = 1$, and its diagonal is $y_j^\star = T^\star s_j(\kappa^\star)$ with $T^\star = C^2/(\kappa^\star)^2$. Dividing by λ_j^2 gives

$$\boldsymbol{W}^\star = \frac{\eta\mu_2(\kappa^\star)T^\star}{2p(\kappa^\star)} \boldsymbol{H}^{-1}(\boldsymbol{I} - \beta(\kappa^\star)\boldsymbol{H})^{-1}.$$

This is the coordinate-free stationary covariance, while $G(\kappa^\star) = 1$ verifies $T(\boldsymbol{W}^\star) = T^\star$.

First consider an off-diagonal perturbation. Since $T(\boldsymbol{W}) = \mathrm{Tr}(\boldsymbol{H}^2\boldsymbol{W})$ sees no off-diagonal entry and $\boldsymbol{W}^\star_{jk} = 0$, the (j, k) mode, $j \neq k$, has derivative multiplier

$$a_{jk} = 1 - p(x_j + x_k) + p\psi x_j x_k,$$

with all moments evaluated at $\kappa^\star$. Now

$$1 - a_{jk} = p(x_j + x_k - \psi x_j x_k) > 0, \qquad 1 + a_{jk} = 2 - p(x_j + x_k) + p\psi x_j x_k > 0,$$

where we used $0 < \psi < 2$, $p \leq 1$ and $0 < x_j \leq x_1 \leq 1/2$. Thus $|a_{jk}| < 1$ for every off-diagonal mode.

It remains to treat the diagonal block. Differentiating (102) at $\boldsymbol{y}^\star$ and using $\mu_2 x_j^2 = (1 - D_j)s_j$ gives

$$\boldsymbol{J}_{\mathrm{diag}} = \boldsymbol{D} + (\boldsymbol{I} - \boldsymbol{D})\boldsymbol{\ell}\mathbf{1}^\top, \qquad \ell_j := s_j + T^\star \frac{ds_j}{dT} = s_j - \frac{\kappa^\star}{2}\frac{ds_j}{d\kappa},$$

where $\boldsymbol{D} = \mathrm{diag}(D_j(T^\star))$. Logarithmic differentiation of (103) yields

$$\ell_j = s_j \left[1 - \frac{\kappa^\star}{2}\frac{\gamma'}{\gamma} - \frac{\kappa^\star\psi'}{4}\frac{x_j}{1 - \psi x_j/2}\right]_{\kappa=\kappa^\star}.$$

Condition (29) implies $\ell_j \geq 0$ for every j, because $x/(1 - \psi x/2)$ is increasing in x. Moreover $0 \leq D_j < 1$: the lower bound follows from $D_j \geq 1 - 2px_j \geq 0$, and $1 - D_j = 2px_j(1 - \psi x_j/2) > 0$. Hence $\boldsymbol{J}_{\mathrm{diag}}$ is nonnegative.

Both γ and ψ are strictly increasing in κ, so $G'(\kappa) > 0$. Since $d\kappa/dT = -\kappa/(2T)$ and $G(\kappa^\star) = 1$,

$$\mathbf{1}^\top\boldsymbol{\ell} = G(\kappa^\star) - \frac{\kappa^\star}{2}G'(\kappa^\star) < 1.$$

Consequently

$$\boldsymbol{I} - \boldsymbol{J}_{\mathrm{diag}} = (\boldsymbol{I} - \boldsymbol{D})(\boldsymbol{I} - \boldsymbol{\ell}\mathbf{1}^\top), \qquad (\boldsymbol{I} - \boldsymbol{\ell}\mathbf{1}^\top)^{-1} = \boldsymbol{I} + \frac{\boldsymbol{\ell}\mathbf{1}^\top}{1 - \mathbf{1}^\top\boldsymbol{\ell}} \geq 0.$$

Thus $(\boldsymbol{I} - \boldsymbol{J}_{\mathrm{diag}})^{-1} \geq 0$. A nonnegative matrix has a nonnegative resolvent at one exactly when its Perron root is below one, so $\rho(\boldsymbol{J}_{\mathrm{diag}}) < 1$. Together with the off-diagonal multipliers this gives $\rho(D\mathcal{F}_{\mathrm{batch}}(\boldsymbol{W}^\star)) < 1$. The map is continuously differentiable near $\boldsymbol{W}^\star$, and the discrete-time linearisation theorem then gives (24). □

Proof of Corollary C.25. By (76), $\mathrm{Tr}(\boldsymbol{K}) = m_2\lambda_r + p\tau_\perp = p(\theta\lambda_r + \tau_\perp)$ with $\theta = m_2/p \leq 1$. If $\lambda_r \ll \tau_\perp$ then $\mathrm{Tr}(\boldsymbol{K}) \approx p\tau_\perp$, and $\tau_\perp \approx 2/\eta$ at criticality by (56). □

F Experimental details

F.1 Data, model, and optimization

We filter CIFAR-10 training data to classes 0–3, shuffle, and retain 1,000 examples. Images are centered and standardized featurewise using this subset. The network consists of two bias-free 3×3 convolutions with 32 channels, each followed by GELU and 2×2 average pooling, then a linear map from $32 \times 8 \times 8$ features to ten outputs. Including the final bias, it has 30,570 parameters. We use one-hot targets and the mean of per-example half-squared errors. Training is deterministic full-batch FO or one-direction Gaussian ZO, without momentum or weight decay.

The ZO step evaluates the centered difference at $\boldsymbol{x} \pm \mu\boldsymbol{u}$ with $\mu = 0.001$. Batch clipping clips the scalar full-batch directional derivative before multiplication by $\boldsymbol{u}$; per-sample clipping clips each sample directional derivative, averages the clipped scalars, and then multiplies by $\boldsymbol{u}$. It does not clip the norm of the final vector estimator.

Table 5: Run grid used for late-checkpoint diagnostics. Three checkpoints per run give 108 evaluations.

Family	Rule	(η, C) grid	Checkpoints
FO	global	$0.05 \times \{0.1, 0.2, 0.4, 0.8\}$	12
FO	per-layer	$0.05 \times \{0.03, 0.06, 0.1, 0.2\}$	12
FO	per-sample	$0.05 \times \{1, 2, 4, 8\}$	12
ZO	global	$0.001 \times \{0.03, 0.06, 0.12, 0.24\}$	12
ZO	global	$0.002 \times \{0.03, 0.06, 0.12, 0.24\}$	12
ZO	per-sample	$0.001 \times \{0.3, 0.6, 1.2, 2.4\}$	12
ZO	per-sample	$0.001 \times \{2, 3, 4, 6, 8, 12, 16, 24\}$	24
ZO	per-sample	$0.002 \times \{2, 4, 8, 16\}$	12

FO runs use 30,000 steps and checkpoints every 1,500 steps. The $\eta = 0.001$ ZO global/small-C runs use 100,000 steps and checkpoints every 5,000; the large-C per-sample runs use a ceiling of 600,000 and checkpoints every 10,000. The $\eta = 0.002$ runs use a ceiling of 400,000 and checkpoints every 10,000. The diagnostic takes the last three available periodic checkpoints of each run.

F.2 Curvature and closure diagnostics

Hessian and aggregate-Jacobian products are computed by automatic differentiation without materializing a $30{,}570 \times 30{,}570$ matrix. Top and bottom symmetric eigenvalues use 24-step full-reorthogonalized Lanczos; nonnormal per-sample operators use 24-step Arnoldi. Twenty Hutchinson probes estimate traces. The gate moments (p, m_2, m_4) in (75) are evaluated analytically from $\kappa = C/\|\boldsymbol{g}\|$; the same identities are applied samplewise for per-sample gates, in which case the second moment $\mathbb{E}[\mathbf{1}\{|\boldsymbol{u}^\top\boldsymbol{g}| < C\}\boldsymbol{u}\boldsymbol{u}^\top] = p\boldsymbol{I} + (m_2 - p)\widehat{\boldsymbol{g}}\widehat{\boldsymbol{g}}^\top$ is summed over samples.

The focused FO checks use 64-step Lanczos for each per-layer operator and 80-step Lanczos/Arnoldi plus 100 power iterations for the active per-sample operators. The spectrum of $\boldsymbol{\Gamma}$ is obtained from its at-most-n dimensional dual Gram matrix, so no 30,570-dimensional matrix is formed. For batch ZO, 16-probe 64-step stochastic Lanczos quadrature estimates $\mathrm{Tr}\,\boldsymbol{H}$ and the spectral function in (81);

the latter is reported only as an aligned-model plug-in because the alignment residual is measured separately. The nonlinear perturbation experiment evolves two copies from radius 10^{-4} for 2,048 steps with identical Gaussian directions, using three independent initial perturbations at each of all 24 batch-ZO late checkpoints. The strongly gated per-sample $C = 0.3$ final checkpoint is evaluated with the same protocol. For the long nonlinear control we extend the final batch $\eta = 0.001, C = 0.03$ checkpoint to 65,536 steps, record every 32 steps, and use three independent perturbations. The reduction-fidelity test instead freezes $\boldsymbol{H}$ and iterates Eq. (94) in all 30,570 coordinates using batched Hessian–vector products: 128 replicas and 512 steps at three high-margin unstable checkpoints, one reduced-stable checkpoint, and the two operator–scalar disagreements. Its confidence intervals resample whole replicas 1,000 times. Finally, the alignment sweep forms $\mathcal{T}_\eta$ explicitly on the symmetric covariance space of a 16-dimensional quadratic and solves its spectral boundary by bisection at 25 gradient directions, for $\kappa = 0.5, 1, 2$ and two spectra with trace-over-top effective ranks 4.50 and 10.34. At the aligned direction, a separate 11-point sweep over $\kappa = 0.35$–2.5 evaluates the finite scalar-to-spectral conversion and its local counterpart in Eq. (96); both derivatives use a centred difference with relative increment $10^{-5}\eta^\star_{\mathrm{ms}}$.

For the fourth-moment comparison, we evaluate ten rank-one bilinear probes at each of six representative checkpoints, with expectations over 4,096 Gaussian directions for batch and 2,048 for per-sample clipping. Since Proposition C.16 gives $\mathcal{Q}$ in closed form for batch clipping, these probes now serve as a numerical validation of the closed form rather than as evidence for an assumption. For per-sample clipping no closed form is available (Remark C.30), so the per-sample probes remain descriptive. The same Monte Carlo directions also give the co-activation matrix $c_{ij} = \mathbb{E}[a_i a_j]$ as a co-occurrence frequency, and hence ς, at no extra model cost.

F.3 Additional numerical results

Table 6: Per-layer FO codimension test over the three late checkpoints. The block coefficients are heterogeneous; γ_k is therefore a structural covariate and the applicable theorem check is the general lower bound in Theorem C.8(ii), which holds in all twelve cases.

C	Active k	γ_k	Linearized–exact gap	Exact/$(2/\eta)$
0.03	3	0.229–0.232	0.0586–0.0638	0.934–0.941
0.06	3	0.148–0.155	0.0117–0.0172	0.984–0.999
0.10	2	0.0118–0.0139	0.00066–0.00218	0.998–1.008
0.20	1	0.00367–0.00707	0.00224–0.00507	1.005–1.019

Table 7: Pointwise active per-sample FO skew and cluster diagnostics. Here $d = ((2/\eta^\star_{\mathrm{ps}}) - S)/S$ is shown in units of 10^{-4}, and $(k, \epsilon/S)$ records the selected leading cluster. The gap-free bound controls d by $\|\boldsymbol{R}_a\|_2^2/(S\Re\lambda^\star)$, and the cluster condition $4\|\boldsymbol{R}_a\|_2/\Gamma_c \le 1$ holds in every row.

C	Step	$\|\boldsymbol{R}_a\|_2/S$	$\Re\lambda^\star/S$	$10^4 d$	$(k, \epsilon/S)$	$4\|\boldsymbol{R}_a\|_2/\Gamma_c$
1	27000	0.00997	0.999947	−0.528	(9, 0.0514)	0.127
1	28500	0.01044	0.999919	−0.811	(9, 0.0634)	0.141
1	30000	0.00990	0.999850	−1.499	(2, 0.0112)	0.925
2	27000	0.00700	0.999906	−0.936	(2, 0.0056)	0.247
2	28500	0.00849	0.999894	−1.058	(2, 0.0070)	0.429
2	30000	0.00585	0.999935	−0.646	(2, 0.0014)	0.387

For the six active batch checkpoints, the median top gap is 0.00386 and the median proxy–exact gap is 0.00192, both consistent with Theorem 3.1 and with Corollary C.3. For per-sample $C = 2$, $\|\boldsymbol{R}_a\|_2/S = 0.00585$–$0.00849$; across all active per-sample checkpoints the relative discrepancy $|(2/\eta^\star_{\mathrm{ps}}) - S|/S$ is at most 1.50×10^{-4}.

The batch high-C skew norm is 1.6–1.8% of the leading scale, and the per-sample aggregate nonnormality diagnostic falls to about 0.5% at $C = 6$ and below 0.1% at $C = 24$. Corollary C.11 predicts the batch value exactly as $2\kappa\phi(\kappa)\|\boldsymbol{Q}\boldsymbol{H}\widehat{\boldsymbol{g}}\|$. Across all 24 batch checkpoints, the measured-to-predicted ratio is 0.9999946–1.0000096, a maximum relative discrepancy below 9.6×10^{-6}. This parameter-free comparison directly validates the nonnormality identity in Proposition C.16.

Table 8: Leading spectrum and stable rank of the clipped-gradient Gram operator $\mathbf{\Gamma}$. The last column is the squared alignment of the proxy's top vector $\boldsymbol{v_A}$ with the top eigenvector of $\mathbf{\Gamma}$.

C	Step	Leading three eigenvalues	$\operatorname{Tr}\mathbf{\Gamma}$	k_{eff}	Alignment^2
1	27000	$0.0361, 0.0138, 0.0119$	0.2552	7.07	0.407
1	28500	$0.0393, 0.0158, 0.0114$	0.2829	7.19	0.864
1	30000	$0.0346, 0.0145, 0.0124$	0.2756	7.96	0.470
2	27000	$0.00753, 0.00504, 0.00261$	0.0465	6.18	0.137
2	28500	$0.00731, 0.00352, 0.00263$	0.0343	4.69	0.031
2	30000	$0.00372, 0.00252, 0.00235$	0.0208	5.61	0.009

Table 9: FO late-checkpoint diagnostics. The exact column is $[2/\eta^\star(\boldsymbol{J}_m)]/(2/\eta)$, read off the Arnoldi spectrum on the positive-real branch. Theorem C.24 bounds the imaginary part at first order, $|\Im\lambda| \le \|\boldsymbol{R}_a\|_2$, but its effect on $2/\eta^\star(\boldsymbol{J}_m)$ only at second order through $(\Im\lambda)^2/\Re\lambda$, which is why the corrected and exact columns agree to the digits shown.

Setting	Clip fraction	Proxy$/(2/\eta)$	Corrected$/(2/\eta)$	Exact$/(2/\eta)$
Batch $C = 0.1, 0.2$	1.000	1.001–1.008	0.995–1.007	0.995–1.007
Per-sample $C = 1$	0.850–0.881	1.242–1.347	0.988–1.055	0.988–1.055
Per-sample $C = 2$	0.031–0.068	1.015–1.049	1.000–1.018	1.000–1.018
Per-sample $C = 4, 8$	0	0.992–1.002	0.992–1.002	0.992–1.002

G Interpretation guide

Five distinctions prevent common overclaims.

First, the fixed-checkpoint coefficient $\min(1, C/\|\boldsymbol{g}\|)$ is not the derivative of active clipping; the exact derivative additionally projects away the radial direction. The deleted direction is the gradient direction, which is why the surviving invariant couples curvature to gradient norm.

Second, ZO threshold-clipping linearization uses the gate $\mathbf{1}\{|d^\mu(\boldsymbol{x};\boldsymbol{u})| < C\}$, whereas writing the nonlinear update as a multiplier uses $\min(1, C/|d^\mu(\boldsymbol{x};\boldsymbol{u})|)$. They coincide only when clipping is inactive, and the difference is not cosmetic: the gate is idempotent and the multiplier is not. Idempotence collapses pair-gate to single-gate order; it yields exact cancellation for an independent shared gate (Lemma C.9) and in the aligned transverse slab balance, but not a premise-free scalar boundary for a direction-dependent gate. It says nothing about distinct per-sample gates, whose pair moments remain (Lemma C.1); those moments are computable (Proposition C.18), so their presence is a change in which operator applies, not a barrier.

Third, "clipping-aware effective curvature" is not one quantity. Scaling curvature by the clipping coefficient omits the radial deletion that differentiation produces, and weighting curvature by the gate omits the matching attenuation of the diffusion. The two omissions have opposite signs: the FO version overestimates curvature and so understates the stable step, while the batch ZO version understates curvature and overstates the margin. A single curve labelled "effective curvature" across rules conflates statements that differ in form and in direction of error.

Fourth, bounded updates do not imply stability. Clipped updates are norm-bounded by construction, so a run can be locally mean-square unstable and still fail to diverge (Remark C.31). Distinguishing the two requires measuring perturbation growth, not iterate norms.

Fifth, a premise failing is not a formula failing, and a formula surviving is not a premise holding. Radial alignment fails at every checkpoint here, yet the aligned root stays within half a percent of the operator boundary, because Proposition C.21 makes the residual enter only at second order. The right reading is neither "the theorem applies" nor "the scalar is worthless" but the measured one: report the scalar against the alignment-free operator and state the error. Conversely, the corrected root's advantage at 24 of 24 checkpoints is not evidence that Assumption C.22 holds there — it does not — and the synthetic sweep shows the correction can go the wrong way when the second Krylov residual is large.

Table 10: Fourth-moment controls. For batch rows, Proposition C.16 turns the deviation of r_4 from one into a prediction: it grows as the gate strengthens because the radial moments (m_2, m_4) fall away from $(1, 3)$ faster than p falls away from 1. Per-sample rows are marked "pair moments": their diffusion is not covered by (77) and requires the bivariate slab integrals of Remark C.30.

Checkpoint	Gate rate	r_4 range	Median	Predicted median
Batch $C = 0.24$	0.987	0.917–1.164	1.013	1.013
Per-sample $C = 6$	0.954	0.874–1.078	0.996	pair moments
Per-sample $C = 24$	0.9998	0.973–1.113	1.046	pair moments
Batch $C = 0.12$ control	0.753	0.922–2.580	1.345	1.321
Batch $C = 0.03$ control	0.331	0.719–12.97	3.340	3.095
Per-sample $C = 0.3$ control	0.063	0.815–1.484	1.330	pair moments

Table 11: All batch-ZO late checkpoints. Slash-separated entries follow the step order shown. Gate widths are $\kappa = C/\|\boldsymbol{g}\|$ and $p = 2\Phi(\kappa) - 1$. The root column is $\eta/\eta^\star_{\text{aligned}}$, where $\eta^\star_{\text{aligned}}$ is obtained by plugging the measured spectrum and moments into (81); because the last column is large, these are aligned-model plug-ins rather than exact empirical boundaries.

η	C	Steps (10^3)	Gate p	Width κ	$\tau_\perp/(2/\eta)$	$\eta/\eta^\star_{\text{aligned}}$	Align. residual
0.001	0.03	90/95/100	0.342/0.338/0.325	0.443/0.437/0.419	0.766/0.741/0.868	0.797/0.772/0.902	0.911/0.882/0.824
0.001	0.06	90/95/100	0.541/0.532/0.487	0.741/0.726/0.641	1.177/1.154/1.219	1.222/1.196/1.264	0.832/0.820/0.789
0.001	0.12	90/95/100	0.696/0.746/0.719	1.029/1.141/1.172	1.045/1.016/1.176	1.084/1.053/1.215	0.722/0.736/0.776
0.001	0.24	90/95/100	0.976/0.983/0.988	2.263/2.383/2.504	0.859/0.876/0.973	0.901/0.918/1.013	0.804/0.834/0.849
0.002	0.03	70/80/90	0.288/0.279/0.265	0.369/0.357/0.339	1.876/1.863/2.055	1.949/1.931/2.129	0.827/0.832/0.774
0.002	0.06	70/80/90	0.441/0.407/0.368	0.584/0.534/0.478	1.602/1.633/1.815	1.655/1.686/1.872	0.779/0.733/0.669
0.002	0.12	70/80/90	0.714/0.720/0.769	1.067/1.080/1.198	1.179/1.143/1.137	1.235/1.199/1.184	0.755/0.764/0.847
0.002	0.24	70/80/90	0.915/0.985/0.962	1.723/2.437/2.073	0.933/0.966/0.926	0.995/1.012/0.977	0.735/0.839/0.768