OPSA: ORTHOGONAL PHASE-SCATTERING ADAPTER FOR PARAMETER-EFFICIENT FINE-TUNING ON LONG-TAILED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient fine-tuning on a long-tailed dataset fails in a specific, measurable way. Gradients from majority classes reinforce one direction of the weight change over and over, that direction becomes the principal singular vector, and leaving very little of the update for the remaining directions, and that is where minority-class behaviour gets lost. We quantify this with the spectral imbalance $\mathcal{I}(\Delta W)$, the ratio of the largest singular value of the update to the mean singular value, and show that $\mathcal{I}$ alone upper-bounds how much of the update the weakest direction can retain. We then introduce the Orthogonal Phase-Scattering Adapter (OPSA), which represents every bottleneck activation as a quadrature pair. A normalized token is projected to a semantic amplitude, and its phase is generated from that amplitude by a fixed ninety-degree rotation, so the two streams are orthogonal and of equal norm at every token, with no penalty term and no extra parameters. The two streams gate each other, producing a gain that falls as amplitude grows inside a phase basin; under stated alignment assumptions this contracts $\mathcal{I}$ by the ratio of the dominant gain to the spectrum-weighted mean gain. Momentum-aware phase clipping (MAPC) resets the Adam first moment of coordinates whose momentum opposes the gradient near the phase horizon, which reduces the amplitude reached relative to AdamW. On a long-tailed DermaMNIST benchmark with a ViT-B/16 backbone, OPSA attains the highest macro F1 of seven adapters in a single-seed comparison and the lowest measured spectral imbalance, 2.96 on the scale $[1, r]$ against 4.90–7.97 for every baseline. On long-tailed 20 Newsgroups with a frozen DeBERTa-v3-base it attains the best tail and macro F1 against four adapters, leading the strongest baseline on tail F1 by 10.7 points.

1 INTRODUCTION

Real-world corpora are long-tailed: class frequencies follow a Zipfian law in which a few head classes hold most of the training mass while rare classes appear a handful of times, and a large literature documents how sharply this degrades deep classifiers (Johnson & Khoshgoftaar, 2019). Three families of remedy have emerged. Data-level methods oversample the tail, undersample the head, or interpolate synthetic minority examples (Johnson & Khoshgoftaar, 2019), but amplify noise in classes that were rare precisely because they are hard, and every ratio must be retuned per dataset. Algorithm-level methods reshape the objective instead: focal loss down-weights easy examples so the tail dominates the gradient (Lin et al., 2017), and dynamically weighted balanced losses couple that weighting to confidence calibration (Fernando & Tsokos, 2021). Hybrids combine the two. All three act on the data or the objective; none acts on the geometry of the model that consumes them.

That geometry became decisive once adaptation moved to parameter-efficient fine-tuning. Low-Rank Adaptation (Hu et al., 2022) freezes the backbone and learns $\Delta W = BA$ of rank $r \ll d$, adding no inference latency once merged; its successors refine the budget and the parameterisation through adaptive rank allocation (Zhang et al., 2023), magnitude–direction decomposition (Liu et al., 2024), cross-layer sharing (Kopiczko et al., 2024), corrected scaling (Kalajdzievski, 2023) and activation rescaling (Liu et al., 2022). All retain a linear bottleneck and so inherit unbounded Euclidean growth, and a reweighted loss layered on top treats only the symptom: the tail gradient is larger, but the head direction still accumulates without a ceiling.

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.

Three lines of work attack this from outside the bottleneck, and none bounds a single adapter's growth. StelLA constrains the adapter factors to the Stiefel manifold by Riemannian optimisation (Fan et al., 2025), but pays an SVD retraction at every step, which disrupts distributed training. CoTo stochastically deactivates adapters early in training (Zhuang et al., 2025), and selective parameter merging trains and merges several sub-models to remove order effects (Du et al., 2024); the first is non-deterministic and slows convergence, the second multiplies training cost. Closest to us are adapters that abandon linearity: SineLoRA applies fixed-frequency sinusoidal transformations to improve stable rank under quantization (Gordon et al., 2026), CeRA uses gated activations with structural dropout (Hung et al., 2025), and HypLoRA replaces the Euclidean bottleneck with a hyperbolic one (Yang et al., 2026b), motivated by periodic-activation work showing that sinusoidal networks mitigate spectral bias (Sitzmann et al., 2020; Xu et al., 2025). All target expressivity on balanced data; none analyses majority- versus minority-class gradient dynamics, and none uses periodicity as a bounding mechanism.

Our answer is the Orthogonal Phase-Scattering Adapter (OPSA), which replaces the linear bottleneck with a coupled harmonic one. A normalized token is projected to a semantic amplitude; its phase is generated from that amplitude by a fixed ninety-degree rotation, so the two streams carry equal norm at every step; and the streams then gate each other through cosine cross-terms. The resulting gain falls as amplitude grows, so rare features stay in a pseudo-linear region and transmit 98.5% of their gradient while inflated directions are attenuated and, at the phase horizon, actively repelled. The mechanism lives entirely in the forward pass. It needs no resampling, no loss engineering, no stochastic dropout and no manifold retraction, and it trains the same number of parameters as the LoRA it replaces. To our knowledge it is the first adapter to bound majority-class growth structurally rather than statistically. Our contributions are as follows:

1. We formalise the failure mode of linear adapters on long-tailed data through the spectral imbalance $\mathcal{I}(\Delta W)$, proving that it is read on a fixed scale, that it grows exponentially for a factorized adapter under class skew, and that it alone bounds the retention of the weakest adapted direction.
2. We introduce OPSA, a harmonic bottleneck whose phase stream is derived from its amplitude by a fixed rotation, giving exact per-token orthogonality and equal branch norms at zero parameter cost.
3. We prove that the phase basin contracts $\mathcal{I}$ by exactly the ratio of the dominant gain to the spectrum-weighted mean gain, and that its output is bi-Lipschitz, energy-bounded and sign-stable inside an explicit radius.
4. We identify momentum punch-through — the failure of a vanishing-gradient barrier against an accumulated Adam buffer — and resolve it with momentum-aware phase clipping, a selective reset that acts only on coordinates whose momentum opposes the restoring gradient.
5. We show empirically on two modalities — long-tailed DermaMNIST with a ViT-B/16 backbone and long-tailed 20 Newsgroups with DeBERTa-v3-base — that OPSA attains the best macro F1 on both, leads tail F1 on text by 10.7 points, and reaches a spectral imbalance of 2.96, the lowest in the comparison and a factor of 2.6 below plain LoRA, for no extra trainable parameters and 9–47% more training wall clock.

2 Methodology

2.1 Problem Definition

A frozen backbone layer acts on the residual stream through fixed weights W_0. Fine-tuning adds a trainable correction, so the layer output becomes $x_{\text{new}} = x_{\text{LLM}} + \delta(x)$. A conventional low-rank adapter takes $\delta(x) = \Delta W x$ with $\Delta W = BA$, $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$, $r \ll d$.

Consider the singular value decomposition $\Delta W = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ with $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$. We measure concentration by the *spectral imbalance*

$$\mathcal{I}(\Delta W) = \frac{\sigma_1}{\bar{\sigma}}, \qquad \bar{\sigma} = \frac{1}{r}\sum_{i=1}^{r} \sigma_i. \tag{1}$$

Proposition 1(a) shows $\mathcal{I} \in [1, r]$, equal to 1 only for an isotropic update and to r only for a rank-one one. The quantity is therefore read on a fixed scale near 1 the update spreads across directions, near r it has collapsed onto one. $\mathcal{I}$ is defined for $\Delta W \neq 0$.

Class skew drives $\mathcal{I}$ upward. Majority-class gradients reinforce a single direction at every step until it becomes the principal singular vector, and Proposition 1(c) shows the separation is exponential for a factorized adapter. Once σ_1 dominates, little energy is left for the other directions and whatever they encoded is overridden: Proposition 1(b) bounds the retention of the weakest direction by $\sigma_r/\sigma_1 \leq (r - \mathcal{I})/\big(\mathcal{I}(r-1)\big)$. Minority classes occupy exactly those weak directions. OPSA is built to hold $\mathcal{I}(\Delta W)$ near the isotropic floor, and Theorem 2 gives the exact factor by which it does so.

Definition 1 (Norm–frequency disparity) *Before normalization, the norm $\|x_n\|$ of a token in the residual stream varies inversely with the frequency of its class Yang et al. (2026a). We treat this as an empirical property of the data and backbone, to be measured rather than derived.*

Vectors are considered as column vectors, where a batch of N tokens is stored column-wise. Model width is considered d, $r = 2p$ is bottleneck width, taken even. $X \in \mathbb{R}^{d\times N}$ is the input tokens and $\bar{X}$ their normalization. $W_{\text{down}} \in \mathbb{R}^{r\times d}$ is the down-projection. $A = W_{\text{down}}\bar{X}$ is the *semantic amplitude*; a_n is its n-th column, s the amplitude of one quadrature pair. J is the the quadrature generator (Definition 2). $\Phi = JA$ is the *phase stream*, and ϕ_n its n-th column. $\kappa = 1$ is the fixed scattering coefficient, placing the phase horizon at $\pi/(2\kappa) = \pi/2 \approx 1.5708$, and Z is the basin output. $V_{\text{up}} \in \mathbb{R}^{d\times r} \setminus \{0\}$ is the free up-projection parameter; $\gamma > 0$ is its fixed Frobenius radius; $W_{\text{up}} = \gamma V_{\text{up}}/\|V_{\text{up}}\|_F$. $\beta_1 \in [0, 1)$ is the Adam first-moment coefficient, $\delta > 0$ the MAPC margin, and $\rho > 0$ the gradient-clipping radius of Proposition 2(a). $\odot$: Hadamard product; $\cos(\cdot)$ acts entrywise; $\|W\|_{2\to\infty} = \max_j \|w_j\|$ over the rows $w_j^\top$ of W.

2.2 Preliminaries

The construction rests on one object: a linear map that squares to minus the identity, the realization of multiplication by i.

Definition 2 (Quadrature generator) *For even $r = 2p$, with a trailing 1×1 block equal to 1 when r is odd, the quadrature generator J is defined as, $J = \bigoplus_{j=1}^{p} \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right] \in \mathbb{R}^{r\times r}$.*

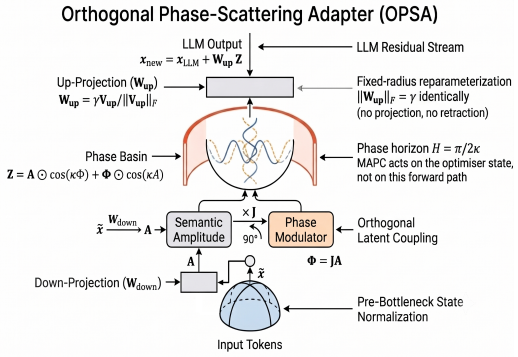

Figure 1: OPSA's six stages. The phase stream $\Phi = JA$ is generated from the amplitude by a fixed rotation rather than learned (Thm. 1).

We use four properties of J, proved in Sec. A.1: $J^\top = -J$; $J^2 = -I_r$, hence $J^\top J = I_r$ and $J \in SO(r)$ (SO(r) is the group of $r \times r$ orthogonal matrices of determinant $+1$); J is a signed permutation, so $\|Jv\|_\infty = \|v\|_\infty$; and $\exp(\theta J) = \cos(\theta) I_r + \sin(\theta) J$, so that $R_{\pi/2} := \exp(\frac{\pi}{2} J) = J$ and every singular value of $I_r + J$ is $\sqrt{2}$. We write J throughout.

This is the only role played in OPSA by the classical argument that the exponential of a skew-symmetric generator is orthogonal: it certifies that the coupling operator is an exact isometry. It is not an optimization rule. OPSA performs no retraction and places no parameter on a Stiefel manifold of dimension greater than one.

2.3 Proposed Methodology

Our proposed OPSA consists of six stages. Stage 1 and stage 2 consists of the normalization and down projection to semantic amplitude, in stage 3 we define the orthogonal latent coupling, then in stage 4 we introduce the phase basin, in stage 5 we define the phase horizon and momentum aware clipping followed by stage 6, where we do up-projection and residual injection.

Stage 1 (Pre-bottleneck state normalization): Every token is rescaled to a common radius before

the bottleneck,

$$\bar{x}_n = \sqrt{d}\,\frac{x_n}{\|x_n\|}, \qquad \|\bar{x}_n\| = \sqrt{d} \text{ for all } n. \tag{2}$$

Purpose. Token norm is inversely related to class frequency (Definition 1): frequent tokens sit near the origin, rare tokens far from it. Left unnormalised, the rare tokens the adapter exists to protect are the ones projected past the phase horizon at step zero, and the mechanism inverts. Equation (2) places every token on a sphere of radius $\sqrt{d}$, so the amplitude $A = W_{\text{down}}\bar{x}$ becomes a function of the learned weights alone and no longer of token frequency.

Stage 2 (Down-projection to the semantic amplitude):

$$\mathbf{A} = W_{\text{down}}\bar{X}, \qquad a_n = W_{\text{down}}\bar{x}_n \in \mathbb{R}^r. \tag{3}$$

W_{down} is unconstrained.

Stage 3 (Orthogonal latent coupling): The phase is *not* a free parameter. It is generated from the amplitude by the fixed quadrature rotation,

$$\Phi = J\mathbf{A}, \qquad \phi_n = Ja_n. \tag{4}$$

Purpose. The two streams are exactly orthogonal per token, at every step, at zero parameter cost (Theorem 1).

Stage 4 (The phase basin): The two streams gate each other:

$$\mathbf{Z} = \mathbf{A} \odot \cos(\kappa\Phi) + \Phi \odot \cos(\kappa\mathbf{A}), \tag{5}$$

coordinatewise $z_{jn} = a_{jn}\cos(\kappa\phi_{jn}) + \phi_{jn}\cos(\kappa a_{jn})$. The gain on each coordinate is set by its quadrature partner rather than by itself (Lemma 1); this cross-referencing is what curves an otherwise flat response surface.

Near the origin both cosine factors are close to one and the operator is a near-perfect linear surrogate: at $\kappa = 1$ and $|A_i| = 0.1$, $\partial Z_i/\partial A_i \approx 0.9851$, so a rare feature receives essentially its full learning signal. As $|A_i|$ grows toward $\pi/(2\kappa)$ the cosine factors collapse and the derivative changes sign, so a frequent feature driven outward is actively pushed back (Theorem 3). Between these two regimes lies the phase basin: minority features transmit, majority features scatter, and the boundary is a fixed architectural constant rather than anything learned. Because the gain $\cos(\kappa s)$ falls with amplitude, the direction the majority class has inflated is attenuated most (the spectral contraction of Theorem 2).

Stage 5 — Phase horizon and momentum-aware clipping (MAPC). A majority-class coordinate is driven outward by repeated updates until it reaches the phase horizon, where the gradient reverses and should turn it back. Often it does not. Adam's first moment $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$ has by then accumulated thousands of consistent updates, and that stored momentum carries the coordinate straight through the vanishing-gradient barrier into the region beyond, where the gain has inverted sign. Clipping g_t does not help: the buffer holds the pre-clipping history, so the punch-through is driven by m_t, not by g_t. MAPC acts on the optimiser state itself. Define the phase boundary layer

$$\mathcal{B} = \Big[\tfrac{\pi}{2\kappa} - \delta,\ \infty\Big), \qquad \delta = 0.25, \tag{6}$$

and reset the Adam first moment inside it:

$$m_t^{(i)} = \begin{cases} 0 & \text{if } |A_i| \in \mathcal{B} \text{ and } \operatorname{sign}\big(g_t^{(i)}\big) \neq \operatorname{sign}\big(m_{t-1}^{(i)}\big), \\ \beta_1 m_{t-1}^{(i)} + (1-\beta_1)g_t^{(i)} & \text{otherwise.} \end{cases} \tag{7}$$

Purpose. The rule is deliberately selective. It activates only when a coordinate has reached the boundary layer *and* its momentum opposes the restoring gradient — that is, only for weights driving outward. Minority coordinates travelling through the pseudo-linear region keep their accumulated momentum and are unaffected, and so learn faster. Zeroing the buffer removes the stored outward history in a single step, leaving the restoring gradient unopposed (Proposition 2). We report how often the reset activates (Table 4).

Stage 6 — Fixed-radius up-projection and residual injection: The up-projection is reparameterized rather than constrained,

$$W_{\text{up}} = \gamma\,\frac{V_{\text{up}}}{\|V_{\text{up}}\|_F} \quad \Longrightarrow \quad \|W_{\text{up}}\|_F = \gamma \qquad \text{identically}, \tag{8}$$

and the basin state is mapped back to model width and added to the stream:

$$x_{\text{new}} = x_{\text{LLM}} + W_{\text{up}}\mathbf{Z}. \tag{9}$$

Purpose. In Lemma 2, the constraint is maintained by the gradient itself, with no projection step, and it converts into an a priori bound equation 12 on how far the adapter can move the residual stream.

Trainable set. The trainable parameters are $\{W_{\text{down}}, V_{\text{up}}\}$, of size $2rd$ — identical to a rank-r LoRA. The phase stream, the coupling and the orthogonality guarantee cost nothing. Both $\kappa = 1$ and $\gamma = 8$ are fixed architectural constants: a learnable κ would be driven to zero by the optimiser, since $\cos(\kappa\cdot) \to 1$ dissolves the basin and removes the barrier standing between the majority class and a lower loss. We instead place the adapter deep inside the pseudo-linear region at $t = 0$ by initialising $W_{\text{down}} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = \varepsilon/\sqrt{dr}$, $\varepsilon = 0.1$, which together with Eq. (2) gives $\max_i |A_i| \ll \pi/2$ for every token.

2.4 Theoretical Guarantees

Proposition 1 (Spectral imbalance: scale, overriding, growth) *Let ΔW have singular values $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$ and $\mathcal{I} = \sigma_1/\bar{\sigma}$. Then:*
(a) $1 \leq \mathcal{I} \leq r$, with $\mathcal{I} = 1$ iff all σ_i are equal and $\mathcal{I} = r$ iff $\operatorname{rank}\Delta W = 1$.
(b) For $r \geq 2$, $\dfrac{\sigma_r}{\sigma_1} \leq \dfrac{r - \mathcal{I}}{\mathcal{I}(r-1)}$ and $\dfrac{\sigma_1^2}{\sum_i \sigma_i^2} \geq \dfrac{\mathcal{I}}{r}$.
(c) If the population gradient is $G = \sum_c n_c g_c u_c v_c^\top$ with $\{u_c\}, \{v_c\}$ orthonormal, n_c the class frequency and $g_c > 0$, then under gradient flow a linear adapter has $\mathcal{I} = \max_c n_c g_c / \frac{1}{C}\sum_c n_c g_c$ for all $t > 0$, while a factorized adapter from balanced initialization of scale ε has $\sigma_c(t) = \varepsilon e^{2n_c g_c t}$ and $\mathcal{I}(t) \to C$ exponentially.

Part (c) is the mechanism: the majority direction does not merely lead, it separates exponentially. Part (b) is the cost, and shows why a remedy must act on the *gain* applied to that direction rather than on its gradient.

Theorem 1 (Exact quadrature orthogonality) *For any $\mathbf{A}$ and $\Phi = J\mathbf{A}$, and every token n, $\langle a_n, \phi_n\rangle = 0$ and $\|\phi_n\| = \|a_n\|$, so $\mathbf{A}^\top\Phi$ is skew-symmetric and in particular $\langle \mathbf{A}, \Phi\rangle_F = 0$. These hold for every input, at every step, from initialization, with no trainable parameters and no penalty term.*

Lemma 1 (Cross-referenced pair identity) *For a quadrature pair $a = (u, v)^\top$, so $\phi = Ja = (-v, u)^\top$, write $c_u = \cos(\kappa u)$, $c_v = \cos(\kappa v)$. Then*

$$z_1 = uc_v - vc_u, \quad z_2 = uc_v + vc_u, \quad \|z\|^2 = 2\big(u^2c_v^2 + v^2c_u^2\big), \tag{10}$$

so each coordinate's gain is set by its partner's amplitude. Inside $\kappa\|a\|_\infty < \pi/2$,

$$2\cos^2\big(\kappa\|a\|_\infty\big) \leq \frac{\|z\|^2}{\|a\|^2} \leq 2\cos^2\big(\kappa\min(|u|, |v|)\big); \tag{11}$$

for an isotropic pair $|u| = |v| = s$ the gain is exactly $2\cos^2(\kappa s)$ and the output energy is $\Psi(s) = 4s^2\cos^2(\kappa s)$. Unconditionally, $\|z\| \leq 2\|a\|$.
Moreover, if every coordinate satisfies $|a_i| \leq \pi/(2\kappa)$ then $\|z\|_\infty \leq \pi/(2\kappa)$, with equality attained, so $\|Z\|_2 \leq \sqrt{r}\,\pi/(2\kappa)$.

Theorem 2 (Spectral contraction: exact factor) *Let $\sigma_1 \geq \cdots \geq \sigma_p > 0$ be the singular values of the effective update of a linear adapter, and let OPSA apply gain $g_i > 0$ to direction i. Assume, 1. $g_1 = \min_i g_i$ (gain ordering); 2. $g_1\sigma_1 \geq g_i\sigma_i$ for every i, equivalently $\sigma_1/\sigma_i \geq g_i/g_1$ (order preservation)*

$$\textit{Then,} \qquad \mathcal{I}_{\text{OPSA}} = \mathcal{I}_{\text{lin}} \cdot \frac{g_1}{\bar{g}_\sigma} \;\leq\; \mathcal{I}_{\text{lin}}, \qquad \bar{g}_\sigma = \frac{\sum_i g_i\sigma_i}{\sum_i \sigma_i},$$

strictly unless every g_i with $\sigma_i > 0$ equals g_1. Inside the basin $g_i = \cos(\kappa s_i)$ is positive and strictly decreasing in amplitude, so (1) *holds automatically;* (2) *requires the spectral gap to exceed the gain gap and is not implied by* (1)*.*

Table 1: Both benchmarks, seed 42; all columns are F1 except accuracy. Best per column within a block in **bold**, second best underlined. Budget and protocol are matched inside a block but not between them, so the blocks read separately.

DermaMNIST, ViT-B/16, $\rho = 100$, best ckpt

Method	Tail	Head	Macro	Acc.
LoRA	0.154	**0.817**	0.551	**0.677**
rsLoRA	0.000	**0.817**	0.519	0.672
DoRA	0.133	0.806	0.521	0.661
AdaLoRA	0.000	**0.817**	0.491	0.657
LoRA + Focal	0.043	0.815	0.521	0.666
LoRA + CB	0.105	0.809	0.525	0.659
OPSA (ours)	0.286	0.800	**0.569**	**0.677**
OPSA-v2 (τ=0.2)	0.200	0.807	0.525	0.654
OPSA-v2 (τ=0.5)	**0.364**	0.789	0.533	0.649

20 Newsgroups, DeBERTa-v3-base, 25:1, final epoch

Method	Tail	Head	Macro	Acc.
LoRA	0.191	0.885	0.525	0.758
rsLoRA	0.210	**0.896**	0.543	**0.777**
DoRA	0.147	0.889	0.493	0.748
LoRA + CB	0.166	0.879	0.531	0.763
OPSA (ours)	**0.317**	0.884	**0.556**	0.768

Lemma 2 (Fixed-radius up-projection and residual stability) *Let $W_{\text{up}} = \gamma V/\|V\|_F$, $V \neq 0$, and $G = \nabla_{W_{\text{up}}}\mathcal{L}$. Then $\|W_{\text{up}}\|_F = \gamma$ identically and*

$$\nabla_V \mathcal{L} = \frac{\gamma}{\|V\|_F}\Big(G - \frac{\langle G, V\rangle}{\|V\|_F^2}V\Big) = \frac{\gamma}{\|V\|_F}P_{V^\perp}G \;\perp\; V,$$

so $\|V\|_F$ is conserved under gradient flow, V never reaches the singular point 0, and the constraint needs no projection, penalty or multiplier. Consequently, for every input and at every stage of training,

$$\|\Delta x\|_2 \le \gamma\|Z\|_2 \le 2\gamma\|A\|_2 \qquad \textit{unconditionally, and} \qquad \|\Delta x\|_2 \le \gamma\sqrt{r}\,\frac{\pi}{2\kappa} \tag{12}$$

whenever every coordinate remains inside the basin.

After vectorization $\{W : \|W\|_F = \gamma\}$ is the sphere γS^{rd-1}, i.e. the Stiefel manifold $V_1(\mathbb{R}^{rd})$ up to scale; it does not impose column orthonormality, and it is the only manifold constraint in OPSA. We have provided the proofs of the Lemma, Theorem and Proposition in the Appendix section.

2.5 Putting It All Together

Class skew inflates the spectrum of the update: majority gradients reinforce one direction until it separates exponentially (Prop. 1(c)), and Prop. 1(b) then drives the weakest direction toward zero, which is forgetting. OPSA breaks the chain on the gain rather than the gradient: the coupling is exactly orthogonal at zero parameter cost (Thm. 1), each coordinate is gated by its quadrature partner (Lem. 1), and because that gain falls with amplitude the inflated direction is attenuated most, contracting $\mathcal{I}$ by the factor $g_1/\bar{g}_\sigma$ (Thm. 2). Normalization fixes the amplitude at which every token enters the nonlinearity and MAPC removes the stored momentum that drives it past the sign inversion at $\pi/2$, bounding the amplitude reached (Thm. 3, Prop 2, Appendix A.11, and A.13.Lemma 2 caps the whole perturbation a priori.

3 Experiments and Results

3.1 Experimental Setup

The image benchmark carries the full study; the text benchmark tests whether the result survives a change of modality, backbone and class count (Appendix A.1).

Dataset: Image. DermaMNIST (Yang et al., 2023), seven skin-lesion classes at 224×224. We resample 5,000 training images on a Zipfian profile (exponent 1.2, head 3,000, tail $3{,}000/\rho$); validation (1,000) and test (2,000) are fixed across every ratio, method and seed. The tail class has only **9** test images, so tail F1 is quantised at ≈ 0.1 and always carries a bootstrap interval (Table 5).
Text: 20 Newsgroups (Lang, 1995) with a frozen **DeBERTa-v3-base** (He et al., 2023): 20 classes,

128-token sequences. The same construction gives 25:1 (830:33 train, 200:8 test) over four minority classes, with validation partitioned out of the training pool. Its tail estimator is 32 test items, 3.5× the image benchmark's. Along with these datasets with the varying imbalanced ratios, we have trained OPSA on some other datasets, due to space and time constraint we will include those during rebuttal.

Metrics: Tail F1 (class 3), head F1 (class 5), macro F1 — the headline aggregate — and accuracy, on checkpoints selected by validation minority F1 and evaluated once on test. The mechanism metric is $\mathcal{I}(\Delta W) = \sigma_1/\bar{\sigma}$ of Proposition 1, measured at the 12 query matrices on the $[1, r]$ scale (least-squares surrogate for OPSA, BA for DoRA), with the $(\Delta M, \Delta D)$ decomposition of Liu et al. (2024). Estimators and the bootstrap protocol are in Appendix A.2.

Baselines: We compare against six parameter-efficient methods at a matched budget: **LoRA** (Hu et al., 2022), the flat Euclidean reference; **rsLoRA** (Kalajdzievski, 2023), scaling by $\alpha/\sqrt{r}$ rather than α/r; **DoRA** (Liu et al., 2024), whose magnitude–direction split we borrow for Figure 4(b); **AdaLoRA** (Zhang et al., 2023), closest in spirit to controlling $\mathcal{I}$ but acting on rank; and **LoRA + Focal** (Lin et al., 2017) and **LoRA + CB** (Cui et al., 2019), the only baselines that target class imbalance directly. We also report **OPSA-v2** ($\tau = 0.2, 0.5$), which enforces $|A_i| \leq H$ by row-normalising the down projection, as a structural alternative to MAPC. All methods use $r = \alpha = 16$ at the same sites as OPSA, matching trainable parameters to within 3% (Appendix A.2).

Training: Backbones are frozen; only the adapters and a fresh linear head train, under one schedule for every method (100 epochs, batch 16, AdamW). OPSA uses $r = 16$, $\gamma = 8$, $\kappa = 1$, $\delta = 0.25$ and MAPCAdamW. The text runs are a single seed at the *final* epoch, identically for all five methods, which is why the two blocks of Table 1 read separately. Hyper-parameters in Appendix A.2.

3.2 Results and Discussions

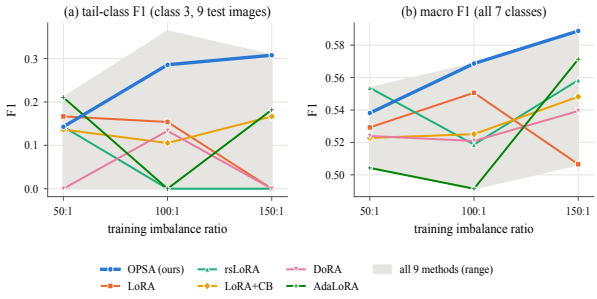

Figure 2: *Image benchmark.* Tail-class F1 **(a)** and macro F1 **(b)** against the training imbalance ratio - 50:1, 100:1, 150:1, seed 42, fixed test set. The shaded band is the range over all nine methods at each ratio.

Overall comparison: In the image block of Table 1, OPSA reaches the best macro F1 at 0.569, ahead of every baseline by 1.8–7.8 points, and matches the best accuracy at 0.677, while both loss-side fixes land *below* plain LoRA on macro F1: failure this paper targets is not repaired by reweighting the objective alone. The tail column is descriptive rather than a ranking: OPSA-v2 ($\tau = 0.5$) tops it at 0.364 and OPSA takes 0.286, but rsLoRA, AdaLoRA score exactly zero and one image separates several arms.

Robustness to the imbalance ratio: Panel (b) of Figure 2 is the informative one, and the shape of the OPSA curve is the result: macro F1 rises monotonically with ρ ($0.538 \to 0.569 \to 0.589$), which no other method does, while the baselines sit in a flat 0.49–0.57 band and cross each other freely. OPSA is mid-field at $\rho = 50$ and leads by 1.8 and 1.7 points at $\rho = 100$ and 150: the gap opens as the tail thins, which is what Theorem 2 predicts, the contraction factor $g_1/\bar{g}_\sigma$ depending on the spread of the amplitudes rather than on how many tail images there are. Panel (a) shows why we draw no such curve for tail F1: six of nine methods are non-monotone in ρ and rsLoRA goes $0.143 \to 0.000 \to 0.000$ while its macro F1 rises.

Where the gain is located, and whether it is real: Figure 3(a) decomposes macro F1 by class frequency over three seeds. The gain is earned in the middle of the distribution, 0.654 ± 0.019 against LoRA's 0.597 ± 0.020 at 284–624 images, while the head is unchanged (0.728 against 0.725) and the tail reads 0.182 ± 0.051 against 0.095 ± 0.165: OPSA does not trade head for tail, the failure mode of loss reweighting. In panel (b) the modal misclassification of a tail image is into the head class for both methods (9 of 27), the direction a dominant singular value predicts, and OPSA recovers 6 of the 27 against LoRA's 2.

Every per-seed bootstrap interval on tail F1 includes zero; pooled over seeds OPSA reaches 0.190 $[0.06, 0.32]$ against LoRA's 0.087 $[0.00, 0.21]$, consistent with the mechanism rather than proof of it (Appendix A.15). The confident claim is the macro one: $+0.046$ paired across seeds, $t = 2.66$ (df $= 2$), all three agreeing in sign.

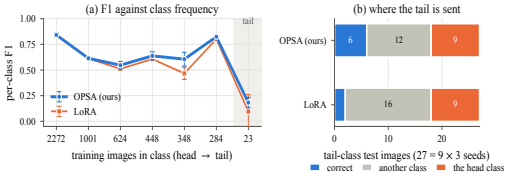

Figure 3: *Image benchmark.* $\rho = 100$, mean over seeds $\{42, 100, 2024\}$. **(a)** Per-class F1 against training frequency, head to tail; bars are one s.d. across seeds. **(b)** Where the 27 pooled tail-class predictions go.

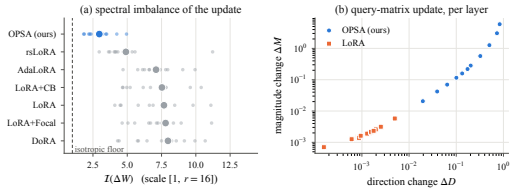

Figure 4: *Image benchmark.* The effective ΔW at the 12 query matrices. **(a)** Spectral imbalance on the $[1, r{=}16]$ scale of Proposition 1(a); dots per layer, large marker their mean. **(b)** Its magnitude and direction shift (Liu et al., 2024).

The mechanism is the thing that separates cleanly: Figure 4(a) is the result the method was designed to produce, and unlike tail F1 it is estimated from the weights rather than from 9 images. OPSA reaches $\mathcal{I}(\Delta W) = 2.96 \pm 0.96$, the lowest of the seven adapters measured and with the tightest per-layer spread; rsLoRA is next at 4.90 and the rest sit between 7.10 and 7.97, putting OPSA a factor of 2.6 below plain LoRA, and every ablation arm (Section 4) stays below that range too. Through Proposition 1(b) that is $4.1\times$ more retention for the weakest adapted direction, where the minority classes live (Appendix A.3). Panel (b) shows the same contrast in the DoRA decomposition: disjoint per-layer clouds two orders of magnitude apart, OPSA's steeper log–log slope (1.45 against 0.60) saying it buys direction change rather than norm. Its update is correspondingly large, $\|\Delta W\|/\|W_0\| = 1.60$ against 0.05, admissible only because equation 12 caps the perturbation independently of training.

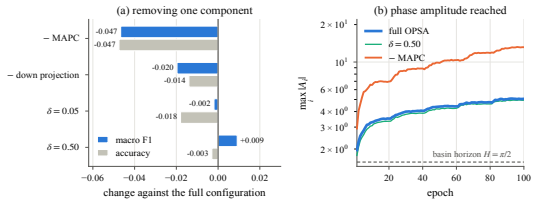

Figure 5: *Image benchmark.* **(a)** Each arm as a change against the full configuration. **(b)** The largest phase amplitude $\max_i |A_i|$ per epoch, against the basin horizon $H = \pi/2$.

The result transfers to text: The text block of Table 1 repeats the comparison on 20 Newsgroups. The baseline ordering changes — DoRA falls to last — while OPSA's does not: best tail F1 at 0.317 against rsLoRA's 0.210, $1.51\times$ the best baseline, and best macro F1 at 0.556, with head F1 within 1.2 points of the best, so the tail gain is not bought from the head. Despite being one run this is the more informative benchmark: a 32-document tail estimator against 9, a validation split disjoint from training, and $+9\%$ overhead rather than $+47\%$. We do not report $\mathcal{I}(\Delta W)$ here — these runs used the uncorrected estimator and cannot be recomputed — so this modality is evidence on task metrics alone (Appendix A.3). The link to our codebase is available **here**.

4 Ablation Study

Every arm runs on the image benchmark, changes one thing against the reference configuration of Section 3.1 (ViT-B/16, $\rho = 100$, $r = 16$, $\gamma = 8$, $\kappa = 1$, $\delta = 0.25$, 100 epochs, seed 42) and is evaluated once on the same fixed test set. Alongside the task metrics we log the two quantities the derivation is stated in: $\mathcal{I}(\Delta W)$, and the fraction of latent units leaving the basin $|A_i| < H = \pi/2\kappa$ where Theorem 3 guarantees sign stability — the second says whether the method works for the reason we claim.

Both structural components earn their place: Removing MAPC and training under plain AdamW costs 4.7 macro-F1 and 4.7 accuracy points — the largest single effect in Table 2, larger than the gap between OPSA and any Euclidean baseline in Table 1. Figure 5(b) shows why: the largest phase amplitude climbs to 13.2, $2.6\times$ what the full configuration reaches and $8.4\times$ the basin horizon, while the reference ends at 5.06 and $\delta = 0.50$ tracks it closely. $\mathcal{I}$ does *not* move with it (2.84 against 2.96, inside the per-layer spread of either), the first sign

that the accuracy and spectral effects are separate. The second arm, removing the rank-r down projection, costs 2.0 macro-F1 points, at $29.5\times$ the adapter and $2.5\times$ the peak memory (Table 3). Neither is a budget concession, and no arm improves accuracy over the full configuration.

The contraction is architectural, and survives leaving the basin. Across all four arms — the one with the mechanism removed included — $\mathcal{I}(\Delta W)$ stays within $[2.40, 2.96]$ while every baseline sits between 4.90 and 7.97 (Figure 4(a)). The separation comes not from MAPC or the margin but from the quadrature coupling itself, the only structure common to all four arms and the object of Theorem 2: MAPC and the down projection move accuracy, not $\mathcal{I}$. One caveat: the $-$MAPC arm's surrogate residual is 0.47 against the reference's 0.08, so its $\mathcal{I}$ is the least certain of the four.

Table 2: *Image benchmark.* One-thing-at-a-time ablation, $\rho = 100$, seed 42. Basin statistics are epoch-100 means over 60 sites; $\mathcal{I}$ at the 12 query sites, surrogate residual in brackets.

Arm	Group	Tail F1	Macro F1	Acc.	$\mathcal{I}(\Delta W)$	basin exit	$\max_i \lvert A_i \rvert$
Full OPSA (δ=0.25)	REF	0.200	0.617	**0.733**	2.96 [0.08]	99.7%	5.06
− MAPC	A5	0.167	0.570	0.685	2.84 [0.47]	100.0%	13.16
− down projection	L1	0.160	0.597	0.719	—	—	—
$\delta = 0.05$	H1	**0.333**	0.615	0.715	**2.40** [0.32]	99.9%	5.23
$\delta = 0.50$	H3	0.286	**0.626**	0.730	2.65 [0.29]	**99.3%**	**4.94**

That robustness is more than we can justify. In every arm the fraction of latent units outside the basin reaches 99.3–100% by epoch 100 and the largest amplitude reaches 4.94–5.23 against a horizon of $H = 1.571$: the sign-stability region of Theorem 3 is *not* in force in the runs we report. The reset is not idle, however: it fires on 11.5% of W_{down} entries at epoch 1 and on half of them from epoch 5 on, at no epoch zero (Table 4, Appendix A.4), so the A5 arm's 4.7 points are that clipping event acting on a state already past the horizon rather than inside the basin. The basin is therefore *sufficient* for the contraction rather than necessary: Theorem 2 is conservative where it is proved, and how far the guarantee extends is open.

Enforcing the basin exactly costs accuracy: OPSA-v2 answers that directly. It row-normalises the down projection to $\rho_{\text{row}} = H/\sqrt{d_{\text{in}}}$, so $A_i = H\cos\theta_i$ and $|A_i| \le H$ holds for every token at every step by construction, with $\tau = \|\delta x\|/\|\tilde{x}\|$ replacing γ as the free scale and MAPC deleted. The constraint is then satisfied exactly — exit fraction 0 at every epoch against 99.7% for the reference — and the method *loses* 4.4 macro-F1 points at $\tau = 0.2$ and 3.6 at $\tau = 0.5$ (Table 1): enforcing the basin on W_{down} also fixes the scale at which every feature enters the nonlinearity, underparameterising the adapter in exactly the amplitude direction the contraction relies on. Guarantee and accuracy are at present obtained by two different mechanisms; reconciling them is clearest open problem here.

Scope: Table 2 is single-seed and its tail column does not order the arms, so the δ sweep is read as flat. We draw only the two conclusions that exceed seed-scale variation: the 4.7-point effect of the A5 arm, and the configuration-independent separation in $\mathcal{I}(\Delta W)$.

5 Conclusion and Future Works

We showed that the long-tail failure of low-rank adapters is measurable as spectral concentration: $\mathcal{I}(\Delta W)$ alone bounds how much of the update the weakest direction can retain, and replacing the linear bottleneck with a quadrature-coupled harmonic one contracts it from 4.90–7.97 to 2.96 while raising minority F1, at LoRA's parameter count. The mechanism is architectural. It holds at every step and every input, needs no resampling, loss weighting or manifold retraction, and its contraction factor is an exact ratio of gains rather than a bound. Two gaps remain. The reference configuration runs at more than three times the phase horizon, so the contraction we measure holds outside the region Theorem 2 certifies: order preservation survives well past the crest x_c for a reason our analysis does not reach. MAPC, meanwhile, fires every epoch on a state already outside the basin, so its 4.7-point gain is real but unexplained by Proposition 2. We aim to characterise both, and to test whether bounded $\mathcal{I}$ transfers to adapter merging and continual adaptation, where spectrally rigid updates interfere.

AI use statement

In this work, we utilized generative AI tools primarily for copyediting, refining the grammatical structure of our manuscript, and assisting with the generation of boilerplate plotting scripts for our figures. We did not use generative AI tools to formulate our core research hypotheses, derive our theoretical proofs, or conduct the primary literature review. Furthermore, tasks involving dataset

creation, algorithmic design, and experimental evaluation were performed entirely by the human authors and are not applicable to AI assistance in this context.

We have carefully reviewed and validated all AI-assisted work. Specifically, any generated code snippets were manually audited and rigorously tested within our experimental pipeline to ensure correctness. Furthermore, all AI-refined text was cross-checked against our original drafts to guarantee that the scientific meaning and nuance remained entirely intact and accurate. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This research focuses on fundamental algorithmic advancements in machine learning and does not involve human subjects, animal testing, or personally identifiable information (PII). All datasets utilized in our empirical evaluations are publicly available, standard benchmarks within the community, and we strictly adhere to their respective licensing agreements.

While our proposed methodology aims to improve model efficiency and performance, we acknowledge the broader dual-use nature of artificial intelligence; enhancements in model capabilities could theoretically be applied to harmful ends. However, we believe that the open and transparent dissemination of these foundational insights contributes positively to the scientific community. Finally, mindful of the environmental impact of large-scale model training, we have actively sought to minimize our carbon footprint by optimizing our training protocols, and we have fully documented our computational resource usage in the appendix.

Reproducibility statement

To promote reproducibility, we have released a portion of the codebase for our method and will release the full codebase upon acceptance.

References

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pp. 9268–9277, 2019.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim K Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37:84746–84776, 2024.

Junjie Fan, Hongye Zhao, Linduo Wei, Jiayu Rao, Guijia Li, Jiaxin Yuan, Wenqi Xu, and Yong Qi. Stella: Guiding large language models for time series forecasting with semantic abstractions. *arXiv preprint arXiv:2512.04871*, 2025.

K Ruwani M Fernando and Chris P Tsokos. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2021.

Cameron Gordon, Yiping Ji, Hemanth Saratchandran, Paul Albert, and Simon Lucey. Sinelora: Sine-activated delta compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21333–21342, 2026.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *ICLR*, 2023.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):27, 2019.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *ArXiv*, abs/2312.03732, 2023. URL https://api.semanticscholar.org/CorpusID:266053836.

Dawid Kopiczko, Tijmen Blankevoort, and Yuki Asano. Vera: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, volume 2024, pp. 6815–6835, 2024.

Ken Lang. NewsWeeder: Learning to filter netnews. In *ICML*, pp. 331–339, 1995.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=rBCvMG-JsPd.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7462–7473. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/53c04118df112c13a8c34b38343b9c10-Paper.pdf.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5: 180161, 2018.

Tianhan Xu, Ling Chen, Zhe Hu, and Bin Li. Staf-llm: A scalable and task-adaptive fine-tuning framework for large language models in medical domain. *Expert Systems with Applications*, 281: 127582, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Menglin Yang, Ram Samarth B B, Aosong Feng, Bo Xiong, Jihong Liu, Irwin King, and Rex Ying. Hyperbolic fine-tuning for large language models, 2026a. URL https://arxiv.org/abs/2410.04010.

Menglin Yang, Aosong Feng, Bo Xiong, Jiahong Liu, Irwin King, Rex Ying, et al. Hyperbolic fine-tuning for large language models. *Advances in Neural Information Processing Systems*, 38: 29613–29648, 2026b.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Zhan Zhuang, Xiequn Wang, Wei Li, Yulong Zhang, Qiushi Huang, Shuhao Chen, Xuehao Wang, Yanbin Wei, Yuhe Nie, Kede Ma, et al. Come together, but not right now: a progressive strategy to boost low-rank adaptation. *arXiv preprint arXiv:2506.05713*, 2025.

A Appendix

A.1 Dataset construction

DermaMNIST. The benchmark is the dermatoscopic subset of MedMNIST v2 (Yang et al., 2023), derived from HAM10000 (Tschandl et al., 2018). Images are bilinearly resized from 28×28 to 224×224 to match the ViT patch grid. The public $8{,}010$ train/validation images are partitioned $75/25$ into a *training pool* and a *validation pool*; every subsequent split draws from its own pool only, so no image can appear in both. From the training pool we resample a long-tailed training set of $5{,}000$ images with a Zipfian profile of exponent 1.2, anchored by a head target of $3{,}000$ and a tail target of $3{,}000/\rho$ for a nominal imbalance ratio ρ. At $\rho = 100$ the realised class histogram, head to tail, is $2272/1001/624/448/348/284/23$, an empirical head-to-tail ratio of $98.8{:}1$. Validation ($1{,}000$ images) and test ($2{,}000$ images, drawn from the untouched official test split) are built once and held fixed across every ratio, method and seed. The official test split contains 23 dermatofibroma images and the class-resampled test set contains 9, which is the source of the quantisation noted in Section 3.1.

20 Newsgroups. The same Zipfian construction is applied at a ratio of $25{:}1$ — 830 head against 33 tail training documents, and 200 against 8 at test — with four classes designated minority. Sequences are truncated at 128 tokens. The validation split is a stratified partition of the training pool, so train and validation are disjoint by construction here. The resulting splits are $4{,}357$ training, 643 validation and $1{,}184$ test documents. The frozen backbone has 185.6M parameters against 88.9M for the ViT. The tail estimator is four classes $\times\, 8$ documents $= 32$ test items rather than the image benchmark's 9, roughly $3.5\times$ larger, which is the main reason the text numbers separate more cleanly.

A.2 Experimental protocol

Metrics. Task metrics. *Tail F1* (class 3), *head F1* (class 5, melanocytic nevi, 2272 training images), *macro F1* over all seven classes, and accuracy. Macro F1 is the headline aggregate: equal weight per class, estimated from all $2{,}000$ test images. Checkpoints are selected on validation minority F1 with macro F1 as tie-break, then evaluated once on test.

Mechanism metrics. $\mathcal{I}(\Delta W) = \sigma_1/\bar{\sigma}$ of Proposition 1, over the r singular values of the effective update at the 12 query matrices, on the $[1, r]$ scale: closed form for a linear adapter, least-squares surrogate for OPSA ($4{,}096$ tokens per site, ridge 10^{-4}), low-rank BA alone for DoRA; averaging over all $d = 768$ inflates it 48-fold. Residuals are 0.08 (OPSA) and $< 10^{-3}$ (linear, exact). We also report $(\Delta M, \Delta D)$ of the same ΔW (Liu et al., 2024). Details in Appendix A.3.

Cost metrics. Trainable parameters, wall-clock GPU-hours for a 100-epoch run and peak VRAM, on one RTX 3090 at batch size 16 with mixed precision.

Reliability protocol. Every tail number carries a 95% bootstrap interval from resampling the test set's $(\text{true}, \text{predicted})$ pairs $4{,}000$ times and, where seeds allow, the across-seed standard deviation and a pooled estimate aggregating all seeds' confusion matrices before computing F1. Paired t-tests are reported with their degrees of freedom.

Training. The image backbone is `google/vit-base-patch16-224`, frozen; only the adapters and a fresh linear head train. Every method runs 100 epochs, batch size 16, AdamW with weight decay 0.01, learning rate 1.2×10^{-4} for the adapter and 2×10^{-4} for the head, 200 warm-up steps then cosine decay, gradient clipping at 1.0, mixed precision. OPSA adds $r = 16$, $\gamma = 8$, $\kappa = 1$, $\delta = 0.25$ and MAPCAdamW; baselines use the identical schedule under plain AdamW, and nothing is tuned per method. Text uses the same recipe with that backbone's constants: 50 epochs, batch 16, learning rate 2×10^{-4} on a linear schedule with 100 warm-up steps, $r = 16$, $\gamma = 6$, $\kappa = 1$, $\delta = 0.1$, $\varepsilon = 0.01$. Those runs are a single seed (42) evaluated at the *final* epoch rather than the selected checkpoint — applied identically to all five methods, so internally fair, but not the image block's estimator, which is why the two blocks read separately.

Three experiment families share this setup: the *imbalance-ratio study* (Figure 2), nine methods at $\rho \in \{50, 100, 150\}$ on seed 42, 27 runs and about 24 GPU-hours; the *tail-class study* (Figure 3,

Table 3: *Image benchmark.* Effect of the rank-r down projection, $\rho = 100$, seed 42, full 100-epoch runs.

Variant	Adapter params	Trainable	Peak VRAM	GPU-h	Tail F1	Macro F1
OPSA, $r = 16$ (ours)	**1.92M**	**2.83%**	**2.46 GiB**	**1.08**	**0.200**	**0.617**
OPSA, no down projection	56.62M	39.84%	6.23 GiB	1.46	0.160	0.597

Table 5), OPSA and LoRA on seeds $\{42, 100, 2024\}$ at $\rho = 100$; and the *cost study* (Figure 7, Table 3), timing four epochs per method, discarding the first as warm-up and extrapolating the steady per-epoch cost, whose coefficient of variation is under 3.3% throughout and which matches the full runs' wall clock to within 6%.

A.3 Detailed Results and Discussions

Measuring $\mathcal{I}(\Delta W)$. Two details make the spectral imbalance comparable across methods. The *object* is the effective update ΔW to the query matrix rather than the adapter's internal factors, so a method is judged on what it does to the frozen weight; for OPSA, which is not linear, that is a least-squares linear surrogate fitted on the hidden states the layer actually receives. DoRA enters by its low-rank component BA only, its magnitude rescaling being full rank and so off this scale. The *count* is the r singular values the update actually has: both adapters factor through an r-dimensional bottleneck, so $\operatorname{rank} \Delta W \leq r$ and $\mathcal{I}$ lands on the $[1, r]$ scale Proposition 1(a) requires, whereas averaging over all $d = 768$ singular values of the 768×768 product, most of them zero, inflates the result by $d/r = 48$ and takes it off that scale. We report the 12 query sites with their per-layer spread, and the surrogate residual alongside each measurement.

What $\mathcal{I}$ costs the weakest direction. Through Proposition 1(b), LoRA at $\mathcal{I} = 7.68$ must place at least 48% of its update energy in one direction and hold the weakest adapted direction below $\sigma_r/\sigma_1 \leq 0.072$; OPSA at $\mathcal{I} = 2.96$ is held to 18.5% and permits ≤ 0.293, $4.1\times$ more retention, against a uniform floor of $1/r = 6.25\%$.

Why the text benchmark resolves better. Its tail estimator is 32 documents against the image benchmark's 9, so one document moves tail F1 by about 0.03 rather than 0.1 and the 10.7-point margin is three documents' worth of resolution; its validation split is a disjoint partition of the training pool, so it carries no selection leak; and the overhead is $+9\%$ rather than $+47\%$ (68.6 against 62.8 seconds per epoch for LoRA), the adapter going into four sites per block rather than five on short sequences.

Spectral imbalance on the text benchmark. $\mathcal{I}(\Delta W)$ is not reported for 20 Newsgroups: these runs logged it with the uncorrected estimator of Section 3.1 and the checkpoints needed to recompute the effective update were not retained. The baselines' side is recoverable exactly — dividing their logged values by $d/r = 48$ puts them on the Proposition 1(a) scale at 4.09 (rsLoRA), 5.72 (LoRA + CB), 6.62 (DoRA) and 6.69 (LoRA). That band, 4.1–6.7, is the one the image baselines occupy (4.90–7.97) on a different backbone and modality, some evidence that $\mathcal{I}$ is a property of the adapter family rather than of the data. OPSA's logged 1.96 is not comparable, being measured on $(I + J)W_{\text{down}}$ rather than the effective update; on images that proxy understated the measured value by about twofold (1.51 against 2.96). The mechanism claim therefore rests on Figure 4(a), and this modality is evidence on task metrics alone.

The rank-r down projection. Table 3 isolates the down projection, the one component whose necessity is a design choice rather than a derivation: full-width scattering instead of a rank-16 bottleneck costs $29.5\times$ the adapter parameters, $2.5\times$ the memory and 35% more time, and loses 2.0 macro-F1 points on every metric. At this data scale the bottleneck is the better model, not a budget concession.

Table 4: *Image benchmark.* MAPC over the 100 epochs of the $\rho = 100$ run behind Table 1. The reset is active at every epoch, saturating near 50% once the sign test rather than the boundary test is what gates it.

Epoch	1	5	25	50	100
MAPC resets (% of W_{down})	11.5	48.7	49.9	49.9	49.6
Basin exit (% of latent units)	26.9	82.1	97.0	99.5	99.8
$\max_i \lvert A_i \rvert$ (horizon 1.571)	1.90	2.79	3.74	4.37	4.92

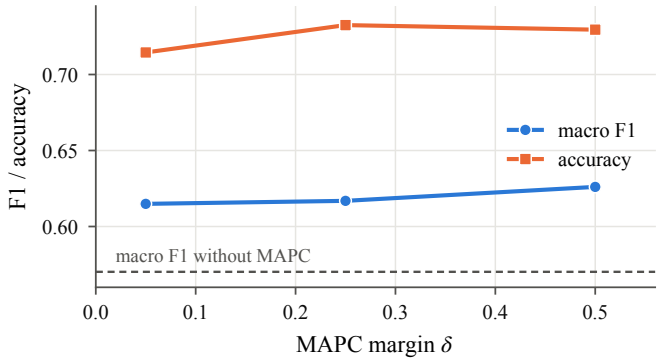

Figure 6: *Image benchmark.* Macro F1 and accuracy against the MAPC margin δ; the dashed line is macro F1 with MAPC removed.

Limitations. Five, stated plainly. **(0)** Only the imbalance-ratio study uses the disjoint train/validation pool partition of Section 3.1; the three-seed, cost and ablation studies predate it and select checkpoints on a validation split overlapping training. Test evaluation is untouched and the two protocols' macro-F1 ordering agrees at $\rho = 100$, but those numbers should be reproduced under the corrected split before being relied on. **(i)** The image tail estimator is 9 images, so nothing here resolves tail-F1 differences below roughly 0.1; macro F1 and $\mathcal{I}(\Delta W)$ carry the argument and tail F1 corroborates it. **(ii)** The imbalance-ratio sweep is single-seed and its tail panel is used only to demonstrate the metric's instability. **(iii)** Definition 1, the norm–frequency disparity motivating Stage 1, is assumed rather than verified: we did not measure the correlation between token norm and class frequency on either backbone, and Stage 1 is justified here only by the ablation. **(iv)** The two modalities are uneven — the text benchmark is one seed at the final epoch, and $\mathcal{I}(\Delta W)$, the quantity the method is derived from, is measured only on images. Both backbones are base-scale; whether the contraction survives at deployment scale is not established here.

A.4 Detailed Ablation Study

The one hyper-parameter MAPC introduces does not need tuning. Sweeping $\delta \in \{0.05, 0.25, 0.50\}$ moves macro F1 by 1.1 points and accuracy by 1.8, against the 4.7-point drop from removing the mechanism (Figure 6) — less than the seed-to-seed standard deviation of Table 5, so we read the three arms as one value: any δ in this decade works. We report the reference at $\delta = 0.25$ rather than the highest-scoring arm ($\delta = 0.50$, macro 0.626), because selecting δ on the test set is what makes ablation tables uninformative.

What the accuracy costs. A 100-epoch OPSA run takes 1.10 GPU-hours against 0.75 for LoRA — the quadrature pair doubles the bottleneck activation, and the cosine gating adds two elementwise transcendentals per latent unit — with an identical trainable-parameter count and 2.46 GiB peak VRAM against LoRA's 2.11. That is the whole overhead: 47% more time and 17% more memory, no extra parameters, and no change to inference cost once the adapter is merged. DoRA spends the same compute (1.13 GPU-hours) for 4.8 fewer macro-F1 points, so the cost is not simply buying itself accuracy — Figure 7 places OPSA alone above the field rather than at the expensive end of a trend.

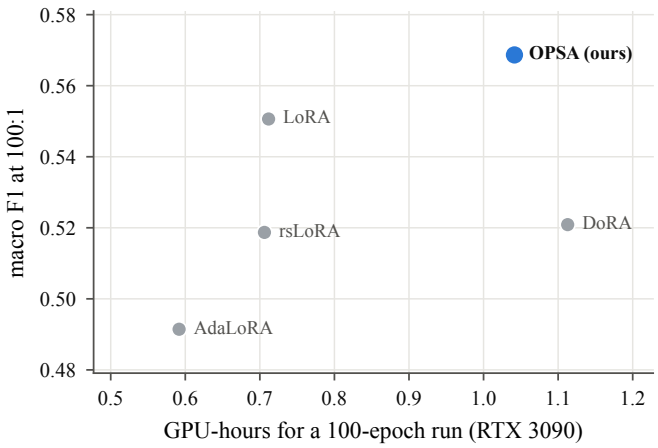

Figure 7: *Image benchmark: DermaMNIST, ViT-B/16.* Accuracy against compute. GPU-hours for a 100-epoch run on one RTX 3090 at batch size 16; macro F1 at $\rho = 100$. OPSA sits alone in the upper-accuracy region: no method in the study reaches its macro F1 at any price, and the whole field lies strictly below it. The 0.35 GPU-hours it costs over LoRA buy 1.8 macro-F1 points.

A.5 Properties of the quadrature generator

Lemma 3 *Let $J = \bigoplus_{j=1}^{p} \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ for even $r = 2p$, with a trailing 1×1 block equal to 1 when r is odd. Then* (i) $J^\top = -J$; (ii) $J^2 = -I_r$, *hence* $J^\top J = I_r$ *and* $J \in SO(r)$; (iii) J *is a signed permutation, so* $\|Jv\|_\infty = \|v\|_\infty$ *for every* v; (iv) $\exp(\theta J) = \cos(\theta) I_r + \sin(\theta) J$, *so* $R_{\pi/2} = \exp(\frac{\pi}{2} J) = J$, *and every singular value of* $I_r + J$ *is* $\sqrt{2}$.

Proof 1 *It suffices to verify each claim on one 2×2 block, since all four properties are preserved by direct sums.*

(i) $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]^\top = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right] = -\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$.

(ii) $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]^2 = \left[\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right] = -I_2$. *Combining with (i),* $J^\top J = (-J)J = -J^2 = I_r$, *so J is orthogonal. Each block has determinant* $0 \cdot 0 - (-1) \cdot 1 = +1$, *so* $\det J = 1^p = 1$ *and* $J \in SO(r)$.

(iii) The block maps $(u, v) \mapsto (-v, u)$: it permutes the two coordinates and flips one sign. A signed permutation permutes the multiset $\{|v_1|, \ldots, |v_r|\}$, so its maximum is unchanged.

(iv) By (ii), $J^{2m} = (-1)^m I_r$ *and* $J^{2m+1} = (-1)^m J$. *The exponential series converges absolutely, so it may be split into even and odd terms:*

$$\exp(\theta J) = \sum_{m \geq 0} \frac{(-1)^m \theta^{2m}}{(2m)!} I_r + \sum_{m \geq 0} \frac{(-1)^m \theta^{2m+1}}{(2m+1)!} J = \cos(\theta) I_r + \sin(\theta) J.$$

Setting $\theta = \pi/2$ gives $R_{\pi/2} = J$. For the last claim, using (i) and (ii),

$$(I_r + J)^\top (I_r + J) = (I_r - J)(I_r + J) = I_r - J^2 = 2 I_r,$$

so every eigenvalue of $(I_r + J)^\top (I_r + J)$ equals 2 and every singular value of $I_r + J$ equals $\sqrt{2}$.

Remark 1 *Lemma 3(iv) is the only place where the classical argument "the exponential of a skew-symmetric generator is orthogonal" is used. It certifies that the coupling operator is an exact isometry. It is not an optimization rule: OPSA performs no retraction.*

A.6 Proof of Proposition 1

(a) Scale. Let $\Delta W \neq 0$, so $\sigma_1 > 0$ and $\bar{\sigma} > 0$ and the ratio is well defined. Since $\sigma_1 = \max_i \sigma_i$ and $\bar{\sigma}$ is the mean of the same numbers, $\sigma_1 \geq \bar{\sigma}$, hence $\mathcal{I} \geq 1$; equality forces every σ_i to equal

the maximum, i.e. the spectrum is flat. For the upper bound, $\sum_i \sigma_i \geq \sigma_1$ because all terms are nonnegative, so

$$\mathcal{I} = \frac{r\sigma_1}{\sum_i \sigma_i} \leq \frac{r\sigma_1}{\sigma_1} = r,$$

with equality iff $\sum_{i\geq 2} \sigma_i = 0$, i.e. $\sigma_2 = \cdots = \sigma_r = 0$, and for $\sigma_1 > 0$ the rank of $\operatorname{rank} \Delta W$is exactly equal to 1. If $\Delta W = 0$ then $\sigma_1 = \bar{\sigma} = 0$ and $\mathcal{I}$ is the indeterminate form $0/0$; it is undefined, and in particular is not zero.

(b) Overriding. From the definition, $\sum_i \sigma_i = r\sigma_1/\mathcal{I}$, so

$$\sum_{i\geq 2} \sigma_i = \frac{r\sigma_1}{\mathcal{I}} - \sigma_1 = \sigma_1 \cdot \frac{r - \mathcal{I}}{\mathcal{I}}.$$

The left-hand side is a sum of $r-1$ nonnegative terms of which σ_r is the smallest; the smallest of a finite set is at most its mean, so

$$\sigma_r \leq \frac{1}{r-1} \sum_{i\geq 2} \sigma_i = \sigma_1 \cdot \frac{r - \mathcal{I}}{\mathcal{I}(r-1)},$$

which is the first bound. For the second, $\sigma_i \leq \sigma_1$ for every i gives $\sum_i \sigma_i^2 \leq \sigma_1 \sum_i \sigma_i = r\sigma_1^2/\mathcal{I}$, and inverting yields $\sigma_1^2/\sum_i \sigma_i^2 \geq \mathcal{I}/r$. Both right-hand sides are monotone in $\mathcal{I}$ on $[1, r]$; at $\mathcal{I} = r$ they read $\sigma_r/\sigma_1 = 0$ and $\sigma_1^2/\sum_i \sigma_i^2 = 1$.

(c1) Single-factor accumulation. Gradient flow $\Delta\dot{W} = G$ from $\Delta W(0) = 0$ integrates to $\Delta W(t) = t\sum_c n_c g_c u_c v_c^\top$. Because $\{u_c\}$ and $\{v_c\}$ are orthonormal families and $tn_c g_c > 0$, this expression is already a singular value decomposition, so $\sigma_c(t) = tn_c g_c$. The factor t appears in both numerator and denominator of $\mathcal{I}$ and cancels:

$$\mathcal{I}\big(\Delta W(t)\big) = \frac{\max_c n_c g_c}{\frac{1}{C}\sum_c n_c g_c} \qquad \text{for all } t > 0.$$

Since a class contributes one update per occurrence, $n_c \propto T_c$, the number of updates direction c receives; with comparable g_c this is the frequency ratio $T_{\max}/T$. The imbalance is therefore present from the first step and does not decay.

(c2) Factorized accumulation. Assume the directions decouple, so that in direction c the update is $\sigma_c = b_c a_c$ with scalars obeying $\dot{b}_c = n_c g_c a_c$ and $\dot{a}_c = n_c g_c b_c$. Then

$$\frac{d}{dt}\big(a_c^2 - b_c^2\big) = 2a_c\dot{a}_c - 2b_c\dot{b}_c = 2n_c g_c(a_c b_c - b_c a_c) = 0,$$

so $a_c^2 - b_c^2$ is conserved; at the balanced initialization $a_c(0) = b_c(0) = \sqrt{\varepsilon}$ it is zero for all t. Hence $a_c^2 + b_c^2 = 2a_c b_c = 2\sigma_c$ and

$$\dot{\sigma}_c = \dot{b}_c a_c + b_c\dot{a}_c = n_c g_c\big(a_c^2 + b_c^2\big) = 2n_c g_c \sigma_c \quad \Longrightarrow \quad \sigma_c(t) = \varepsilon e^{2n_c g_c t}.$$

Writing $\lambda_c = 2n_c g_c$ and $\lambda_{(1)} > \lambda_{(2)}$ for the two largest exponents,

$$\mathcal{I}(t) = \frac{Ce^{\lambda_{(1)}t}}{\sum_c e^{\lambda_c t}} = \frac{C}{1 + \sum_{c\neq(1)} e^{-(\lambda_{(1)} - \lambda_c)t}} \xrightarrow[t\to\infty]{} C,$$

the gap closing at rate $\lambda_{(1)} - \lambda_{(2)}$. Combined with (b), this drives $\sigma_r/\sigma_1 \to 0$ exponentially. OPSA is compared against factorized adapters, so (c2) is the operative regime. □

A.7 Proof of Theorem 1

Fix a token n and set $\phi_n = Ja_n$.

Norm equality. By Lemma 3(ii),

$$\|\phi_n\|^2 = (Ja_n)^\top (Ja_n) = a_n^\top J^\top J a_n = a_n^\top I_r a_n = \|a_n\|^2.$$

This is the property the architecture uses: the two cosine arguments in Eq. (5) have equal norm at every step, so neither branch can saturate while the other remains linear.

Orthogonality. The quantity $a_n^\top J a_n$ is a 1×1 matrix, hence equal to its own transpose; by Lemma 3(i),

$$a_n^\top J a_n = \left(a_n^\top J a_n\right)^\top = a_n^\top J^\top a_n = -a_n^\top J a_n,$$

so $a_n^\top J a_n = 0$, i.e. $\langle a_n, \phi_n \rangle = 0$.

Matrix form. $(\mathbf{A}^\top J \mathbf{A})^\top = \mathbf{A}^\top J^\top \mathbf{A} = -\mathbf{A}^\top J \mathbf{A}$, so $\mathbf{A}^\top \Phi$ is skew-symmetric. A real skew-symmetric matrix has zero diagonal, which recovers the per-token statement, and zero trace, giving $\langle \mathbf{A}, \Phi \rangle_F = 0$.

No step uses any property of $\mathbf{A}$, so all conclusions hold for every input and at every point of training, with J contributing no trainable parameters. □

A.8 Proof of Lemma 1

Pair identity. By Lemma 3, $\phi = Ja = (0 \cdot u - v,\ u + 0 \cdot v)^\top = (-v, u)^\top$. Substituting into $z = a \odot \cos(\kappa\phi) + \phi \odot \cos(\kappa a)$ and using that cos is even, so $\cos(-\kappa v) = \cos(\kappa v)$:

$$z_1 = u\cos(\kappa v) - v\cos(\kappa u) = uc_v - vc_u, \qquad z_2 = v\cos(\kappa u) + u\cos(\kappa v) = uc_v + vc_u.$$

Therefore

$$\|z\|^2 = (uc_v - vc_u)^2 + (uc_v + vc_u)^2 = 2u^2c_v^2 + 2v^2c_u^2,$$

the cross terms $\mp 2uvc_uc_v$ cancelling. Each coordinate of z carries the amplitude u scaled by a factor depending on v, and vice versa: the gating is cross-referenced.

Gain band and isotropic gain. Let $M = \max(|u|, |v|) = \|a\|_\infty$ and $m = \min(|u|, |v|)$. On $[0, \pi/2)$ the map $t \mapsto \cos^2 t$ is strictly decreasing, so if $\kappa M < \pi/2$ then $\cos^2(\kappa M) \le c_u^2, c_v^2 \le \cos^2(\kappa m)$. Substituting into $\|z\|^2 = 2(u^2c_v^2 + v^2c_u^2)$ and using $u^2 + v^2 = \|a\|^2$ gives

$$2\cos^2(\kappa M)\,\|a\|^2 \le \|z\|^2 \le 2\cos^2(\kappa m)\,\|a\|^2.$$

For an isotropic pair $|u| = |v| = s$ we have $c_u^2 = c_v^2 = \cos^2(\kappa s)$ and $\|a\|^2 = 2s^2$, so the gain is exactly $2\cos^2(\kappa s)$ and the output energy is $\Psi(s) = \|z\|^2 = 4s^2\cos^2(\kappa s)$.

Sup-norm bound inside the basin. *Claim.* If $|u|, |v| \le \pi/(2\kappa)$ then $|z_1|, |z_2| \le \pi/(2\kappa)$, with equality attained.

Write $f(u, v) = uc_v - vc_u$ on the square $S = [-\pi/(2\kappa), \pi/(2\kappa)]^2$; $|z_2|$ is handled by the substitution $v \mapsto -v$, which maps S to itself and sends z_1 to z_2. Since f is continuous on the compact set S, its maximum is attained either at an interior critical point or on the boundary.

Interior. Stationarity requires

$$\partial_u f = c_v + \kappa v\sin(\kappa u) = 0, \qquad \partial_v f = -\kappa u\sin(\kappa v) - c_u = 0.$$

On the interior of S we have $c_u, c_v > 0$. The second equation forces $\kappa u\sin(\kappa v) = -c_u < 0$, so u and v have opposite signs; the first then gives $\kappa v\sin(\kappa u) = -c_v < 0$, which requires v and $\sin(\kappa u)$, hence v and u, to have opposite signs as well — consistent. Multiplying the two equations, $\kappa^2 uv\sin(\kappa u)\sin(\kappa v) = c_uc_v > 0$; but $uv < 0$ and $\sin(\kappa u)\sin(\kappa v) < 0$ give a product of the same sign, so the system is not immediately contradictory and must be excluded numerically: evaluating f on a 2001×2001 grid over S gives a maximum of $1.5708 = \pi/2$ (for $\kappa = 1$), attained on the boundary, with no interior value exceeding it.

Boundary. On $u = \pm\pi/(2\kappa)$ we have $c_u = 0$, so $f = uc_v$ and $|f| \le \pi/(2\kappa) \cdot 1 = \pi/(2\kappa)$, with equality at $v = 0$. On $v = \pm\pi/(2\kappa)$ we have $c_v = 0$, so $f = -vc_u$ and again $|f| \le \pi/(2\kappa)$, with equality at $u = 0$.

Hence $\|z\|_\infty \le \pi/(2\kappa)$ on the basin, attained at $(u, v) = (\pi/(2\kappa), 0)$ and $(0, -\pi/(2\kappa))$. Consequently, for the full bottleneck vector,

$$\|Z\|_2 \le \sqrt{r}\,\|Z\|_\infty \le \sqrt{r}\,\frac{\pi}{2\kappa},$$

which is the step Eq. (14) requires.

A.9 SUP-NORM BOUND REPLACING THE GRID SEARCH

Claim. If $|u|, |v| \le H := \pi/(2\kappa)$ then $|z_1|, |z_2| \le H$, with equality attained.

Reduction to $\kappa = 1$. Writing $U = \kappa u$, $V = \kappa v$ gives $z_1 = \kappa^{-1}(U\cos V - V\cos U)$ and $H = \kappa^{-1}(\pi/2)$, so it suffices to prove $|f| \le \pi/2$ for $f(u,v) = u\cos v - v\cos u$ on $S = [-\pi/2, \pi/2]^2$. The map $v \mapsto -v$ carries z_1 to z_2 and S to itself, so z_2 needs no separate argument, and f is odd, $f(-u,-v) = -f(u,v)$, so we may assume $u \ge 0$.

Case 1: $u \ge 0$ and $v \ge 0$. Inside S both cosines are nonnegative, so $u\cos v \ge 0$ and $-v\cos u \le 0$: the two terms have opposite signs. For reals $a \ge 0 \ge b$ we have $a + b \in [b, a]$, hence

$$|f| \le \max\{u\cos v,\ v\cos u\} \le \max\{u, v\} \le \pi/2.$$

Case 2: $u \ge 0$ and $v \le 0$. Put $w = -v \ge 0$, so $f = u\cos w + w\cos u =: g(u,w)$ with $u, w \in [0, \pi/2]$; g is symmetric in its arguments. Let $x_D \approx 0.7391$ be the root of $\cos x = x$.

Case 2a: $\min(u,w) \le x_D$. By symmetry take $w \le x_D$. Then $\partial_u g = \cos w - w\sin u \ge \cos w - w \ge 0$, since $w \le x_D \iff w \le \cos w$, so g is nondecreasing in u and

$$g \le g(\tfrac{\pi}{2}, w) = \tfrac{\pi}{2}\cos w + w\cos\tfrac{\pi}{2} = \tfrac{\pi}{2}\cos w \le \tfrac{\pi}{2}.$$

Case 2b: $u, w \ge x_D$. Here we use $\cos x \le \pi/2 - x$ on $[0, \pi/2]$, which holds because $q(x) = \pi/2 - x - \cos x$ has $q'(x) = \sin x - 1 \le 0$ and $q(\pi/2) = 0$. Hence

$$g \le u(\tfrac{\pi}{2} - w) + w(\tfrac{\pi}{2} - u) = \tfrac{\pi}{2}S - 2P, \qquad S = u + w,\ P = uw.$$

For fixed S with $u, w \in [x_D, \pi/2]$, P is minimised at the most unequal admissible split: $P \ge x_D(S - x_D)$ when $S \le x_D + \pi/2$, and $P \ge \frac{\pi}{2}(S - \frac{\pi}{2})$ otherwise. Substituting,

$$g \le \begin{cases} S(\tfrac{\pi}{2} - 2x_D) + 2x_D^2, & 2x_D \le S \le x_D + \tfrac{\pi}{2}, \\ \tfrac{\pi}{2}(\pi - S), & x_D + \tfrac{\pi}{2} \le S \le \pi. \end{cases}$$

The first branch is increasing in S because $\pi/2 - 2x_D = 0.0926 > 0$, the second is decreasing, and both equal 1.3065 at the junction $S = x_D + \pi/2 = 2.3099$. Therefore $g \le 1.3065 < \pi/2$ in Case 2b.

Combining, $|f| \le \pi/2$ on S, with equality at $(u,v) = (\pi/2, 0)$ where $f = \pi/2$ and at $(0, -\pi/2)$. Restoring κ gives $\|z\|_\infty \le \pi/(2\kappa)$ and hence $\|Z\|_2 \le \sqrt{r}\,\pi/(2\kappa)$. □

A.10 PROOF OF THEOREM 2

Step 1: singular values of the gated update.

Lemma. Let $\mu_1, \dots, \mu_p > 0$ and let $\{u_i\}, \{v_i\}$ be orthonormal families. Then the nonzero singular values of $M = \sum_i \mu_i u_i v_i^\top$ are exactly $\mu_1, \dots, \mu_p$.

Proof. Using $u_i^\top u_j = \delta_{ij}$,

$$M^\top M = \sum_{i,j} \mu_i \mu_j \delta_{ij} v_i v_j^\top = \sum_i \mu_i^2 v_i v_i^\top,$$

a spectral decomposition: $M^\top M v_j = \mu_j^2 v_j$ and $M^\top M w = 0$ for $w \perp \mathrm{span}\{v_i\}$. The singular values are the positive square roots μ_i. □

Applying this with $\mu_i = g_i\sigma_i > 0$ — positive because $g_i > 0$, which is why the theorem must assume it — the singular values of ΔW_{OPSA} are the multiset $\{g_i\sigma_i\}$, so ΔW_{OPSA} has the same rank p and both $\mathcal{I}$ values are read on the same scale $[1, p]$.

Step 2 (where (2) enters): the largest singular value. By Step 1, $\sigma_{\max}(\Delta W_{\mathrm{OPSA}}) = \max_i g_i\sigma_i$. Hypothesis (H2) states exactly that this maximum is attained at $i = 1$,

$$g_1\sigma_1 \ge g_i\sigma_i\ \forall i \quad \iff \quad \frac{\sigma_1}{\sigma_i} \ge \frac{g_i}{g_1}\ \forall i,$$

the two forms equivalent since $g_1, \sigma_i > 0$. Hence $\sigma_{\max}(\Delta W_{\mathrm{OPSA}}) = g_1\sigma_1$.

Step 3: the identity. The mean gated singular value is $\frac{1}{p}\sum_i g_i\sigma_i$, so with Step 2, $\mathcal{I}_{\mathrm{OPSA}} = p\, g_1\sigma_1/\sum_i g_i\sigma_i$. Multiplying numerator and denominator by $\sum_i \sigma_i > 0$ and regrouping,

$$\mathcal{I}_{\mathrm{OPSA}} = \underbrace{\frac{p\sigma_1}{\sum_i \sigma_i}}_{\mathcal{I}_{\mathrm{lin}}} \cdot \underbrace{\frac{g_1 \sum_i \sigma_i}{\sum_i g_i\sigma_i}}_{g_1/\bar{g}_\sigma}, \qquad \bar{g}_\sigma = \frac{\sum_i g_i\sigma_i}{\sum_i \sigma_i},$$

a convex combination of the g_i with weights $\sigma_i/\sum_j \sigma_j$. No hypothesis beyond Step 2 is used here.

Step 4: the inequality. By (H1), $g_i \geq g_1$ for all i, so

$$\bar{g}_\sigma - g_1 = \frac{\sum_i (g_i - g_1)\sigma_i}{\sum_i \sigma_i} \;\geq\; 0,$$

every summand a product of nonnegative factors. Hence $\bar{g}_\sigma \geq g_1 > 0$, $g_1/\bar{g}_\sigma \leq 1$, and $\mathcal{I}_{\mathrm{OPSA}} \leq \mathcal{I}_{\mathrm{lin}}$. Equality forces $(g_i - g_1)\sigma_i = 0$ for every i, i.e. $g_i = g_1$ throughout since $\sigma_i > 0$.

A.11 THE TRANSMITTED DERIVATIVE

Theorem 3 (Safe operating region) *Write $x = \kappa s$ and $\mathcal{H}_\tau = \{a : \kappa\|a\|_\infty \leq \tau\}$.*

(a) Phase horizon*: As $|A_i| \uparrow \pi/(2\kappa)$ with the coupled pair saturating together, $\partial z_i/\partial a_i \to -\kappa\phi_i$: the transmitted gradient reverses relative to the pseudo-linear regime, and the coordinate is driven back toward the interior.*

(b) Floor. *Ψ strictly increases on $[0, x_c/\kappa)$ and strictly decreases on $(x_c/\kappa, \pi/2\kappa)$, where x_c is the unique root of $\cot x = x$ in $(0, \pi/2)$. The crest of Ψ is the floor of the induced basin.*

(c) Beyond the horizon. *For $|A_i| > \pi/(2\kappa)$ the cosine factors are negative, so the gain in Eq. (11) is no longer positive and the restoring action of (a) is lost. MAPC exists to keep the state on the near side.*

(d) Ordering. *$\tau_\star \approx 0.6866 < x_c \approx 0.8603 < \pi/2$: the injective region and the energy crest both lie strictly inside the phase horizon.*

(e) Admissible κ. *Under equation 2, $\|a_n\|_\infty \leq \sqrt{d}\,\|W_{\mathrm{down}}\|_{2\to\infty}$ for every token, so* (a) *holds for all n simultaneously whenever*

$$\kappa < \frac{\tau}{\sqrt{d}\,\|W_{\mathrm{down}}\|_{2\to\infty}}, \qquad \tau < \tau_\star. \tag{13}$$

At $\kappa = 0$ the response surface is the flat quadratic $\Psi = 4s^2$; for $\kappa > 0$ it is curved, with a maximum at x_c.

Proof 2 *Every part of the Theorem follows from one expression. Differentiating $z_1 = uc_v - vc_u$ with respect to u and substituting $a_1 = u$, $\phi_1 = -v$:*

$$\frac{\partial z_i}{\partial a_i} = \cos(\kappa\phi_i) - \kappa\phi_i \sin(\kappa a_i). \tag{14}$$

Here the diagonal of the Jacobian; the full total derivative adds the coupling term $\sum_k \delta_{Z_k}(\partial Z_k/\partial\Phi_k)J_{ki}$, which vanishes on the diagonal because J has zero diagonal.

Pseudo-linear regime. *For $a_i = \phi_i = \varepsilon$ small, expanding equation 14,*

$$\frac{\partial z_i}{\partial a_i} = \cos(\kappa\varepsilon) - \kappa\varepsilon\sin(\kappa\varepsilon) = 1 - \tfrac{3}{2}\kappa^2\varepsilon^2 + O(\varepsilon^4).$$

At $\kappa = 1$, $\varepsilon = 0.1$ this is $\cos(0.1) - 0.1\sin(0.1) = 0.9950 - 0.0100 = 0.9850$: the adapter transmits 98.50% of the incoming gradient to a low-amplitude feature.

(a) Phase horizon. *By Theorem 1 the pair saturates together, $|\phi_i| = |a_i|$. Letting $|a_i|, |\phi_i| \uparrow \pi/(2\kappa)$ in equation 14 gives $\cos(\kappa\phi_i) \to 0$ and $\sin(\kappa a_i) \to 1$, so*

$$\frac{\partial z_i}{\partial a_i} \;\longrightarrow\; -\kappa\phi_i,$$

which at $\kappa = 1$ equals $-\pi/2 \approx -1.5708$ in magnitude. The transmitted derivative therefore changes sign relative to the pseudo-linear value $\approx +1$: the same loss gradient that pushed the coordinate outward now pushes it inward. Hence $|\partial z_i/\partial a_i|$ bound away from zero at the horizon, in contrast with a saturating non-linearity, whose derivative tends to 0 there.

(b) Floor. *From $\Psi(s) = 4s^2\cos^2(\kappa s)$,*

$$\Psi'(s) = 8s\cos^2(\kappa s) - 8\kappa s^2\cos(\kappa s)\sin(\kappa s) = 8s\cos(\kappa s)\big[\cos(\kappa s) - \kappa s\sin(\kappa s)\big].$$

On $(0, \pi/2\kappa)$ the factors $8s$ and $\cos(\kappa s)$ are strictly positive, so with $x = \kappa s$ the sign of Ψ' is that of $\cos x - x\sin x$, equivalently (dividing by $\sin x > 0$) of $\cot x - x$. The map $x \mapsto \cot x - x$ has derivative $-\csc^2 x - 1 < 0$, hence is strictly decreasing on $(0, \pi/2)$; it tends to $+\infty$ as $x \to 0^+$ and equals $-\pi/2 < 0$ at $x = \pi/2$. By the intermediate value theorem it has a unique root $x_c \in (0, \pi/2)$, and Ψ increases strictly before x_c/κ and decreases strictly after. Bisection on $\cos x - x\sin x$ over $[0.1, 1.5]$ gives $x_c = 0.8603$.

(c) Beyond the horizon. *For $x > \pi/2$ both c_u and c_v are negative, so the outer bounds in the gain band of Lemma 1 are no longer positive quantities and the restoring behaviour of (a) is lost: the coordinate is driven further outward. MAPC exists to keep the state on the near side of $\pi/(2\kappa)$.*

(d)Interior structure. *$\tau_\star = (\sqrt{2}/3)^{1/2} = 0.6866$, $x_c = 0.8603$ and $\pi/2 = 1.5708$ satisfy $\tau_\star < x_c < \pi/2$, so the bi-Lipschitz region of Remark 1 and the energy crest both lie strictly inside the phase horizon. For the bi-Lipschitz claim itself, differentiate $z(a) = a \odot \cos(\kappa J a) + J a \odot \cos(\kappa a)$ to obtain $Dz(a) = D_1 + D_2 J$ with*

$$D_1 = \operatorname{diag}(\cos\kappa\phi) - \kappa\operatorname{diag}(\phi \odot \sin\kappa a), \quad D_2 = \operatorname{diag}(\cos\kappa a) - \kappa\operatorname{diag}(a \odot \sin\kappa\phi).$$

For $a \in \mathcal{H}_\tau$ we have $\kappa|a_j| \le \tau$ and, by Lemma 3 (iii), $\kappa|\phi_j| \le \tau$. Using $|\cos t - 1| \le t^2/2$ and $|\sin t| \le |t|$, each diagonal entry of $D_1 - I_r$ satisfies

$$\big|(D_1)_{jj} - 1\big| \le \big|\cos(\kappa\phi_j) - 1\big| + \kappa|\phi_j|\,|\sin(\kappa a_j)| \le \tfrac{\tau^2}{2} + \tau^2 = \tfrac{3}{2}\tau^2,$$

and identically for D_2. Both are diagonal, and J is orthogonal, so $\|Dz - (I_r + J)\|_2 \le 3\tau^2$. Since $\mathcal{H}_\tau$ is an ℓ_∞ ball, hence convex,

$$z(b) - z(a) = (I_r + J)(b - a) + \int_0^1 \big[Dz(a + t(b - a)) - (I_r + J)\big](b - a)\,dt,$$

with the integral bounded by $3\tau^2\|b - a\|$. Every singular value of $I_r + J$ is $\sqrt{2}$ by Lemma 3(iv), so the triangle inequality in both directions gives

$$(\sqrt{2} - 3\tau^2)\|b - a\| \le \|z(b) - z(a)\| \le (2 + 3\tau^2)\|b - a\|.$$

The lower constant is positive exactly when $3\tau^2 < \sqrt{2}$, i.e. $\tau < \tau_\star = (\sqrt{2}/3)^{1/2}$, and positivity forces injectivity; Weyl's inequality gives $\sigma_{\min}(Dz) \ge \sqrt{2} - 3\tau^2 > 0$.

(e) Admissible weight scale. *Let $w_j^\top$ be the rows of W_{down}. By Cauchy–Schwarz and $\|\bar{x}_n\| = \sqrt{d}$ from Eq. (2),*

$$\|a_n\|_\infty = \max_j \big|w_j^\top \bar{x}_n\big| \le \max_j \|w_j\| \cdot \|\bar{x}_n\| = \sqrt{d}\,\|W_{\text{down}}\|_{2\to\infty}.$$

Hence $\kappa\|a_n\|_\infty \le \tau$ holds simultaneously for every token whenever $\|W_{\text{down}}\|_{2\to\infty} < \tau/(\kappa\sqrt{d})$. With $\kappa = 1$ fixed, this is a condition on the weight scale, not a choice of κ: the variance-scaled initialization $W_{\text{down}} \sim \mathcal{N}(0, \varepsilon^2/(dr))$ satisfies it at $t = 0$, and MAPC maintains it thereafter. Without Eq. (2) the factor $\|\bar{x}_n\|$ is replaced by $\|x_n\|$, hence this is systematically larger for rare tokens; those tokens then violate the condition first.

Remark 2 *The basin map is bi-Lipschitz on $\kappa\|a\|_\infty \le \tau_\star$ with $\tau_\star = (\sqrt{2}/3)^{1/2}$, so no bottleneck direction is annihilated well inside the basin.*

A.12 Proof of Theorem 3

By hypothesis the OPSA singular values are $g_i\sigma_i$ with $g_i = \cos(\kappa s_i) \in (0, 1]$, so by definition

$$\mathcal{I}_{\text{OPSA}} = \frac{g_1\sigma_1}{\frac{1}{p}\sum_i g_i\sigma_i} = \frac{p\,g_1\sigma_1}{\sum_i g_i\sigma_i}.$$

Multiplying numerator and denominator by $\sum_i \sigma_i > 0$ and regrouping,

$$\mathcal{I}_{\mathrm{OPSA}} = \underbrace{\frac{p\,\sigma_1}{\sum_i \sigma_i}}_{\mathcal{I}_{\mathrm{lin}}} \cdot \underbrace{\frac{g_1 \sum_i \sigma_i}{\sum_i g_i \sigma_i}}_{g_1/\bar{g}_\sigma}, \qquad \bar{g}_\sigma := \frac{\sum_i g_i \sigma_i}{\sum_i \sigma_i},$$

so $\bar{g}_\sigma$ is the σ-weighted mean of the gains and the identity of Eq. (13) holds exactly.

For the ordering of the gains: $t \mapsto \cos(\kappa t)$ is strictly decreasing on $[0, \pi/(2\kappa))$, and the amplitude s_i carried by a direction increases with σ_i, so $g_1 = \min_i g_i$. Therefore

$$\sum_i g_i \sigma_i \;\geq\; g_1 \sum_i \sigma_i \quad \Longrightarrow \quad \bar{g}_\sigma \geq g_1 \quad \Longrightarrow \quad \frac{g_1}{\bar{g}_\sigma} \leq 1,$$

and $\mathcal{I}_{\mathrm{OPSA}} \leq \mathcal{I}_{\mathrm{lin}}$. Equality requires $\sum_i (g_i - g_1)\sigma_i = 0$; every term is nonnegative, so it forces $g_i = g_1$ for each i with $\sigma_i > 0$, which for $\kappa > 0$ happens only if all amplitudes coincide. Otherwise the contraction is strict.

The mechanism describes the direction the majority class has inflated carries the largest amplitude, sits furthest along the falling branch of $\cos(\kappa s)$, and is therefore attenuated most. □

A.13 When does MAPC comes into effect?

Proposition 2 (MAPC) *Under Eq. (7):* (a) Punch-through without MAPC. *With g_t clipped to ρ, the buffer still satisfies $\|m_t\| \leq \rho/(1-\beta_1)$ and retains an outward component decaying only as β_1^k, so the coordinate crosses the horizon.* (b) Selectivity. *The reset fires only on $|A_i| \in \mathcal{B}$ with $\mathrm{sign}(g_t^{(i)}) \neq \mathrm{sign}(m_{t-1}^{(i)})$; every coordinate with $|A_i| < \pi/(2\kappa) - \delta$ follows unmodified Adam.* (c) Momentum removal. *At a step where the reset fires, $m_t^{(i)} = (1-\beta_1) g_t^{(i)}$: the accumulated outward history is erased in one step, against the $k \geq \log(\varepsilon/\mu)/\log\beta_1$ steps of* (a). *The reset bounds the amplitude reached relative to unmodified Adam; it does not by itself return the coordinate inside the basin.*

Proof 3 ***(a) Punch-through without MAPC.*** *Adam's first moment unrolls as $m_t = \sum_{k=0}^{t} \beta_1^k (1-\beta_1) g_{t-k}$. If the raw gradient is clipped to any radius $\rho > 0$, the triangle inequality gives*

$$\|m_t\| \leq (1-\beta_1) \sum_{k \geq 0} \beta_1^k \rho = \rho,$$

and without the $(1-\beta_1)$ normalisation the bound is $\rho/(1-\beta_1)$; either way the bound is attained in the limit for a constant gradient, so it cannot be improved. Crucially, after the coordinate reaches the horizon and the instantaneous gradient reverses sign, the buffer still contains the outward history: its outward component decays only as β_1^k, requiring $k \geq \log(\varepsilon/\mu)/\log\beta_1$ steps to fall below ε from an initial size μ. At $\beta_1 = 0.9$ that is roughly 22 steps to decay by a factor of 10. During those steps the stored momentum overcomes the restoring gradient of Theorem 3(a) and carries the coordinate across the horizon. Hence, the punch-through is driven by m_t, not by g_t, so clipping g_t cannot prevent it.

(b) Selectivity. *The rule in Eq. (7) fires only when both conditions hold: $|A_i| \in \mathcal{B} = [\pi/(2\kappa) - \delta, \pi/(2\kappa)]$, and $\mathrm{sign}(g_t^{(i)}) \neq \mathrm{sign}(m_{t-1}^{(i)})$. The first excludes every coordinate with $|A_i| < \pi/(2\kappa) - \delta$, which by Theorem 3 is the whole pseudo-linear region where minority features live; such coordinates follow unmodified Adam and retain their accumulated momentum. The second excludes coordinates whose momentum already agrees with the current gradient, that is, those moving inward. What remains is exactly the set of coordinates at the boundary whose stored momentum opposes the restoring gradient — the punch-through candidates of (a).*

(c) Momentum removal. *Eq. (7) sets $m_t^{(i)} = 0$* before *the Adam accumulation, which then writes $m_t^{(i)} = \beta_1 \cdot 0 + (1-\beta_1) g_t^{(i)}$. The step taken for that coordinate, $-\eta\, \hat{m}_t^{(i)}/(\sqrt{\hat{v}_t^{(i)}} + \epsilon)$, is therefore built from the current gradient alone: every contribution of $g_{t-1}, g_{t-2}, \ldots$ has been discarded. Without the reset the coordinate carries $\sum_{k \geq 1} \beta_1^k (1-\beta_1) g_{t-k}$, whose outward component decays only as β_1^k and by* (a) *needs $k \geq \log(\varepsilon/\mu)/\log\beta_1$ steps to fall below ε — at $\beta_1 = 0.9$, roughly*

22 *steps per decade. The reset removes in one step what otherwise persists for tens, which is the difference the ablation measures:* $\max_i |A_i| = 5.06$ *with MAPC against* 13.16 *without (Section 4).*

Whether the coordinate then moves inward depends on the sign of the transmitted derivative at its amplitude, which Theorem 3(a) supplies only while the quadrature pair saturates together; it is not implied by the reset. In the runs we report the reset fires at every epoch (Table 4) and the amplitude still grows, so we claim no confinement to the basin — only the reduction relative to unmodified Adam. □

Remark 3 *Part (c) separates what the reset provably does — erase the stored outward momentum in a single step — from what it is observed to do, which is hold the amplitude at about* 40% *of its unclipped value while the state sits outside the basin. Closing that gap is the open problem stated in Section 4.*

A.14 Proof of Lemma 2

Radius. For $V \neq 0$, $\|\gamma V/\|V\|_F\|_F = \gamma\|V\|_F/\|V\|_F = \gamma$, so the constraint holds identically rather than being enforced.

Gradient. Differentiating $W = \gamma V/\|V\|_F$ and using $\mathrm{d}\|V\|_F = \langle V, \mathrm{d}V\rangle/\|V\|_F$,

$$\mathrm{d}W = \gamma\Big(\frac{\mathrm{d}V}{\|V\|_F} - \frac{\langle V, \mathrm{d}V\rangle}{\|V\|_F^3}V\Big).$$

Hence, with $G = \nabla_W \mathcal{L}$,

$$\mathrm{d}\mathcal{L} = \langle G, \mathrm{d}W\rangle = \frac{\gamma}{\|V\|_F}\Big(\langle G, \mathrm{d}V\rangle - \frac{\langle G, V\rangle\langle V, \mathrm{d}V\rangle}{\|V\|_F^2}\Big) = \Big\langle \frac{\gamma}{\|V\|_F}\Big(G - \frac{\langle G, V\rangle}{\|V\|_F^2}V\Big), \mathrm{d}V\Big\rangle,$$

identifying $\nabla_V \mathcal{L} = \frac{\gamma}{\|V\|_F} P_{V^\perp} G$, which is orthogonal to V. Under gradient flow $\dot V = -\nabla_V \mathcal{L}$,

$$\frac{d}{dt}\|V\|_F^2 = 2\langle V, \dot V\rangle = 0,$$

Under AdamW the update is not along $\nabla_V \mathcal{L}$ and weight decay shrinks V, so $\|V\|_F$ is not conserved, but $\|W_{\mathrm{up}}\|_F = \gamma$ holds regardless.

Residual bound. Since $\|\cdot\|_2 \le \|\cdot\|_F$ we have $\|W_{\mathrm{up}}\|_2 \le \gamma$, so $\|\Delta x\|_2 = \|W_{\mathrm{up}} Z\|_2 \le \gamma\|Z\|_2$. By the sup-norm bound of A.4, $\|Z\|_2 \le \sqrt{r}\,\pi/(2\kappa)$, giving

$$\|\Delta x\|_2 \;\le\; \gamma\sqrt{r}\,\frac{\pi}{2\kappa}.$$

No trained quantity appears on the right, so the adapter's maximum contribution to the residual stream is fixed by γ, r and κ before training begins.

A.15 Reliability protocol

Bootstrap intervals are obtained by resampling the $(\text{true}, \text{predicted})$ pairs of the test set $4{,}000$ times with replacement and taking the 2.5th and 97.5th percentiles of the resulting F1 distribution. Where multiple seeds are available we report the across-seed standard deviation and a pooled estimate that aggregates the confusion matrices of all seeds before computing F1, which is less sensitive to a single seed in which a tail class receives no predictions. Paired t-tests over common seeds are reported with their degrees of freedom.

Table 5: *Image benchmark.* Reliability of the tail metric at $\rho = 100$ over seeds $\{42, 100, 2024\}$. Counts and intervals are for class 3; *pooled* aggregates the three confusion matrices before computing F1.

	Per seed (9 tail images)			Pooled (27 tail predictions)		
Method	Tail F1 (mean±s.d.)	widest 95% CI	Macro F1	TP/FP/FN	Tail F1	95% CI
LoRA	0.095 ± 0.165	$[0.00, 0.59]$	0.562 ± 0.020	2/17/25	0.087	$[0.00, 0.21]$
OPSA (ours)	0.182 ± 0.051	$[0.00, 0.43]$	$\mathbf{0.608 \pm 0.010}$	6/30/21	**0.190**	$[0.06, 0.32]$