CERTIFICATE TIGHTNESS IS A TRAINING PROPERTY

Anonymous authors
Paper under double-blind review

ABSTRACT

A deterministic robustness certificate rests on a global Lipschitz bound, usually the product of the layers' spectral norms. On a standard network under standard training, that bound exceeds the sensitivity the network exhibits by orders of magnitude. We study the ratio of the two, the gap. The bound certifies the true constant from above, and any exhibited Jacobian certifies it from below. So the gap is the factor between the two ends of a certified enclosure. No verifier returning a global constant can improve the certified radius by more than that factor. This paper characterizes what sets the gap. It does not build a certified system. Its one intervention leaves the architecture and the loss fixed. We track the gap along training over 380 runs on four datasets. At a fixed convolutional architecture, training sets it. At learning rates fixed in advance, the gap grows $3\times$ under SGD, $60\times$ under AdamW and $2{,}300\times$ under Muon. Tuned rates compress that spread to one order, SGD still mildest and Muon below AdamW. At those fixed rates on ImageNet-64, the spread widens to ten orders. Every quantitative prediction was registered before its test run, with two marked exceptions, and every miss is printed. The gap factors exactly into misalignment between layers and slack from units the activations switch off. Projecting each layer onto a spectral-norm cap after every step removes the slack term. Under Muon it moves the gap from 3.2 million to 7.4, at 0.55 clean accuracy against 0.88. The product bound is then within a factor of about seven of the best radius any global-constant verifier could grant. Repair after training recovers a factor of two to six, against five orders. Under the cap the optimizers' roles invert. At the operating budget and the rates fixed in advance, SGD and AdamW collapse toward rank one and stop learning. Muon, which inflates the gap most when free, keeps training. We name this the constraint inversion. The cap moves only the bound, and verified accuracy also needs a trained margin. On four published 1-Lipschitz constructions at one fixed bound, Muon verifies more than SGD, with disjoint seed ranges. Training decides how much of a tight bound is usable.

1 INTRODUCTION

A deterministic robustness certificate is a promise built from one number. Suppose no input change is amplified by more than a factor L. An input whose top logit leads the runner-up by m then keeps its label within a ball of radius $m/(\sqrt{2}\,L)$. The $\sqrt{2}$ converts a logit difference to the pairwise convention (Tsuzuku et al., 2018). Most such certificates build L as the product of the layers' spectral norms (Virmaux & Scaman, 2018; Sedghi et al., 2019). It is inexpensive to compute, sound, and available for any network. It is also loose. On an unconstrained six-layer CIFAR-10 network, it certifies about 3×10^{-9} in pixel units. That is six orders of magnitude below the $1/255$ step of an eight-bit image. A verifier measuring this network's sensitivity perfectly could still reach about 10^{-2} (Fig. 1a, Proposition 1). Reported looseness reaches tens of orders of magnitude on ImageNet-scale networks (Khromov & Singh, 2024). Across the unconstrained networks in this paper, it spans a few hundred to beyond 10^{18}.

Two lines of work answer the looseness. One builds networks whose layers are orthogonal or norm-controlled by construction (Cissé et al., 2017; Anil et al., 2019; Trockman & Kolter, 2021; Singla & Feizi, 2021; Prach & Lampert, 2022; Araujo et al., 2023). The other folds the bound into the loss and trains against it (Tsuzuku et al., 2018; Leino et al., 2021; Hu et al., 2023). Both accept that a standard network under standard training carries a loose certificate. Its size has been measured and

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.

its growth with depth described (Khromov & Singh, 2024; Ma et al., 2023). Yet it has not been treated as a quantity to move at a fixed architecture. What decides the looseness of the product bound on a standard network? Can the training that sets it be made to keep it small?

Asking it needs a measurable quantity. We take the ratio G of the product bound to the largest Jacobian norm the network exhibits on data. The numerator is a certified ceiling on the true Lipschitz constant. The denominator is a certified floor, since an exhibited Jacobian is a proof of sensitivity. The true constant therefore lies between them. G is the factor between the ends of that certified enclosure (Section 2). Consider any sound verifier that returns a global constant. Small G means the simplest bound in the literature is already within a factor G of its best radius. The proof needs no verifier (Proposition 1). That reading makes the ratio, not the bound, the object here.

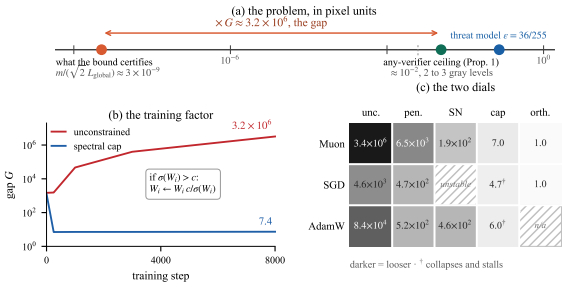

Figure 1: **The gap, its origin, and its dials. (a)** The product bound certifies far less than one gray level, so the looseness is not a verification problem. **(b)** On one network, one projection per training step separates a gap of 3.2 million from one of 7.4. So the looseness is a training property. **(c)** Endpoint gap over three optimizers and five routes of spectral control, darker looser. The optimizer sets the gap and the route hardens it. A † marks capped runs that collapse toward rank one and stop learning. Their low gap certifies nothing. Settings, seeds and exact values are in Section 3.1 and Appendix B.

We tracked G over a grid of 380 training runs. It spans four datasets, three optimizers, four routes of spectral control and a by-construction anchor. Every quantitative prediction was registered before its test run, with two marked exceptions. Every miss is printed (Section 2.4). Training sets the gap, and Fig. 1b isolates it. Two runs of one network share an initialization, a data order and 8,000 steps. They differ only in a rescaling that shrinks any layer whose spectral norm exceeds a fixed budget. It is applied at initialization and after each step. Both begin with the bound overstating the witnessed sensitivity, by 1,478 and 1,279. The unconstrained run ends at 3.2 million and the capped run at 7.4. The cap costs clean accuracy, which Section 3.7 quantifies. The panel shows seed 42, and the other seeds lie within 14 percent. That an optimizer moves a Lipschitz bound has been observed (Khromov & Singh, 2024; Ma et al., 2023), but not what it does to the ratio. Three choices make the question answerable. First, the ratio rather than the bound. A loose certificate can be the fault of the network or of the verifier, and only the ratio separates them. By Proposition 1, it bounds what any sound verifier returning a global constant could recover. So it decides whether to retrain or to buy a better verifier. Second, the optimizer rather than the architecture. It shapes the spectra the bound multiplies and can be swapped without redesigning the model. Third, controlled enforcement rather than observation. Moving one lever, the cap alone, turns a correlation into an intervention with one cause. That design costs verified accuracy, the share of inputs a certificate covers, and pays for it in attribution. The paper characterizes training. It does not build a certified system. Concurrently, Kuang et al. (2026) take the ratio itself as their object, on three-layer MNIST networks at one endpoint. They compare no optimizers and apply no hard projection. No prior work takes that combination as its object (Section 4).

- **At a fixed architecture, training sets the gap.** Capping every weight matrix's spectral norm after each step moves G by over five orders on convolutional networks. At rates fixed in advance, the optimizer grades G across three orders. SGD gives the smallest gap, AdamW a larger one and Muon (Jordan et al., 2024) the largest. We call this ordering, and the distances between the three, the grading. It was registered before the AdamW runs existed and reproduces on three datasets and a residual convolutional family (Section 3.1). Tuned rates compress this to one order, SGD still mildest and Muon below AdamW (Section 3.5).
- **Where the gap lives is measured.** $G = M \times A$ separates cross-layer misalignment from activation gating, exactly. The cap removes the gating term under all three optimizers, though only Muon still trains there. Without it the looseness is in gating under Muon and

AdamW, and in misalignment under SGD. Under Muon, skip connections relocate it into misalignment, as registered before the residual runs (Section 3.2).

- **Repair after training recovers a factor of two to six, and the levers compose.** By proof, function-preserving rescaling cannot move G. Applying the cap's projection after training shifts it by a factor of six, against five orders. Swapping optimizers mid-run shows G follows the one acting now (Section 3.4). The budget traces an accuracy-tightness frontier. A margin term in the loss raises verified accuracy at unchanged G (Section 3.7).
- **Under a hard spectral cap SGD and AdamW stop learning and Muon does not, the constraint inversion.** Projection after every step drives SGD to stable rank one at every budget tried. Stable rank counts a layer's effective directions. AdamW collapses the same way at our operating cap of 2. Muon, which inflates G most when free, keeps training. At a thousand classes it needs a logit temperature and a lower rate. AdamW trains at looser budgets and on a dense network. Its death is specific to convolutional stacks at the tight budget. SGD's is not. Neither a temperature nor a smaller rate alone rescues the two that die. Muon's tolerance was found by Newhouse et al. (2025). The deaths by rank collapse are new. Two registered predictions about the update explain half of them (Section 3.3).

The result is a convolutional network whose product bound is within a factor of seven of the true constant. It keeps 0.55 clean accuracy against 0.88. Prior routes to the same property change what is trained, through orthogonal layers or a bound-aware loss. This one changes only how.

2 Preliminaries and protocol

2.1 Networks and notation

We study feed-forward networks $f_\theta = \phi_n \circ \sigma \circ \phi_{n-1} \circ \cdots \circ \sigma \circ \phi_1$. Each $\phi_\ell(\boldsymbol{x}) = \boldsymbol{W}_\ell \boldsymbol{x} + \boldsymbol{b}_\ell$ is an affine map, dense or convolutional. The activation σ is the ReLU, positively 1-homogeneous and applied elementwise. Every model carries biases and no normalization layers. We write $\|\boldsymbol{W}\|_2$ for the spectral norm of the linear part. For an input $\boldsymbol{x}$ at which f_θ is differentiable, the Jacobian factorizes as

$$\boldsymbol{J}(\boldsymbol{x}) \;=\; \boldsymbol{W}_n \, \boldsymbol{D}_{n-1}(\boldsymbol{x}) \, \boldsymbol{W}_{n-1} \cdots \boldsymbol{D}_1(\boldsymbol{x}) \, \boldsymbol{W}_1, \tag{1}$$

where $\boldsymbol{D}_\ell(\boldsymbol{x})$ is the diagonal $0/1$ matrix of active units at layer ℓ. Biases thus move the gates and then drop out of the derivative. The residual family is treated in Section 2.3. Per-layer operator norms are exact for linear layers, and every convolution here is zero-padded. We compute the exact largest singular value of its circular relaxation (Sedghi et al., 2019). This certifiably upper-bounds the deployed operator. A zero-padded convolution is a compression of the circular one, and compressions never increase an operator norm. Average pooling enters the product with its exact norm, $1/2$ for a $2{\times}2$ pool and $1/4$ for the global pool. On the plain stack, L_{global} is the product of the seven weight-layer norms over 32 (Appendix Q). Depth counts convolutions, so the head is the seventh weight layer at depth 6.

2.2 The certified enclosure, the gap, and verification headroom

The true Lipschitz constant $L_\star = \sup_{\boldsymbol{x}\neq\boldsymbol{y}} \|f_\theta(\boldsymbol{x}) - f_\theta(\boldsymbol{y})\|_2 / \|\boldsymbol{x} - \boldsymbol{y}\|_2$ is what a certificate needs and is intractable to compute. Exact evaluation is NP-hard even for two-layer networks (Virmaux & Scaman, 2018). We therefore work with one certified bound on each side of it,

$$L_{\mathrm{global}} \;=\; \prod_{\ell=1}^{n} \|\boldsymbol{W}_\ell\|_2, \qquad L_{\mathrm{local}} \;=\; \max_{\boldsymbol{x}\in\mathbb{D}_{\mathrm{wit}}} \|\boldsymbol{J}(\boldsymbol{x})\|_2, \tag{2}$$

the product bound and the witnessed sensitivity over a witness set $\mathbb{D}_{\mathrm{wit}}$ of data points. The first is what a certificate may promise. Multiply the layers' worst-case stretches, and no input change is amplified by more. The second is what the network is caught doing: the largest stretch any data point actually exhibits. Submultiplicativity, the ReLU's unit Lipschitz constant, and the construction of Section 2.1 make L_{global} a certified ceiling. Biases translate a layer's output without entering its constant. Every witness is a certificate in the other direction. If the network stretches some direction by a factor at one input, the true constant is at least that factor. An exhibited Jacobian thus proves

$L_\star \geq L_{\text{local}}$. Hence

$$L_{\text{local}} \;\leq\; L_\star \;\leq\; L_{\text{global}}, \qquad \text{and we define} \qquad G \;=\; L_{\text{global}}/L_{\text{local}} \;\geq\; 1. \tag{3}$$

At $G = 1$ the two instruments meet and no verifier can do better. Above one, the true constant lies in a band whose ends differ by that factor. The certificate prices its worst end. So a factor of a million in G is up to a factor of a million of certified radius forgone. That factor, as $\log G$, is the object of this paper. G is defined where $L_{\text{local}} > 0$. A dead network whose witnessed Jacobian vanishes is reported as dead, not as a gap. Every gap is quoted under the 64-witness convention of Section 2.4. Each is an upper estimate of that factor, since a better witness set can only raise L_{local} (Appendix E). Every quantity is rechecked by exhaustion on a two-dimensional problem (Appendix D).

The gap prices a guarantee. Write $m(\boldsymbol{x}) = f_\theta(\boldsymbol{x})_y - \max_{j \neq y} f_\theta(\boldsymbol{x})_j$ for the logit margin. Take any input with $m(\boldsymbol{x}) > 0$ and any sound bound $L_V \geq L_\star$. The prediction is then constant on the ℓ_2 ball of radius $r_{L_V}(\boldsymbol{x}) = m(\boldsymbol{x})/(\sqrt{2}\, L_V)$ (Tsuzuku et al., 2018). Projected gradient descent (PGD) at $0.9\times$ these radii flips no prediction (Appendix C). Verified accuracy at a radius is the share of test inputs classified correctly and certified there. Published baselines test each pair of logits instead, which never certifies less. Table 9 prints our routes under both.

Proposition 1 (Verification headroom)**.** *Let V be any sound verifier, i.e. any procedure returning L_V with $L_\star \leq L_V$. Then the certified radius computed from L_V exceeds the one computed from L_{global} by a factor of at most G.*

Proof. Soundness and equation 3 give $L_V \geq L_\star \geq L_{\text{local}} = L_{\text{global}}/G$, so $r_V(\boldsymbol{x})/r_{L_{\text{global}}}(\boldsymbol{x}) = L_{\text{global}}/L_V \leq G$. □

The proposition applies in both directions. At $G \approx 3{\times}10^6$ (Fig. 1b), up to six orders of certified radius may be recoverable. No verifier alone can tell how much. At $G \approx 7$, the trivial bound is within a factor 7 of what any sound verifier could grant (Fig. 1a). It covers only verifiers returning a global constant. A local verifier can lie below $L_\star$ on a ball and grant more.

2.3 Two places the gap can live

The gap has exactly two places to live. Between layers, the product of norms ignores that consecutive layers need not align their principal directions. Inside the data path, activations gate units off. Both causes were noted by Zou et al. (2020). The accounting below makes the split exact.

Let L_{linear} denote the operator norm of the affine skeleton, the network with every activation deleted and biases removed. It is computed exactly, so $L_{\text{linear}} \leq L_{\text{global}}$ holds by construction (Appendix B). Define

$$M \;=\; \frac{L_{\text{global}}}{L_{\text{linear}}}, \qquad A \;=\; \frac{L_{\text{linear}}}{L_{\text{local}}}, \qquad \text{so that} \qquad G \;=\; M \times A \tag{4}$$

exactly, by construction. M, the misalignment factor, measures the looseness of multiplying layer ceilings that the composed map never attains. A, the gating factor, is the skeleton norm over the sensitivity the data path exhibits. Each $\boldsymbol{D}_\ell$ in equation 1 deletes the rows of units that are off. That lowers each factor's norm but not necessarily the product's. A deleted row can be one that cancelled another. Hence A is an exact identity rather than an inequality. M also carries the slack each certified factor holds over its exact operator norm. At the capped operating point this slack compounds to 1.05, two percent of $\log M$. There M is 7.25 with the slack and 6.9 without (Appendix B).

Submultiplicativity gives $L_{\text{linear}} \leq L_{\text{global}}$, so $M \geq 1$ always. A carries no such guarantee, because gating is not monotone on the product norm. Take two units with rows $(1, 0)$ and $(-1, 0)$ feeding a head $(1, 1)$. They have a zero skeleton, and a unit Jacobian wherever one unit is off. So a witness can exhibit a Jacobian larger than the skeleton norm.

In the residual family, a block $\boldsymbol{x} \mapsto s_\ell \boldsymbol{x} + \boldsymbol{W}_{\ell,2}\,\sigma(\boldsymbol{W}_{\ell,1}\sigma(\boldsymbol{x}))$ contributes the certified factor $s_\ell + \|\boldsymbol{W}_{\ell,2}\|_2\|\boldsymbol{W}_{\ell,1}\|_2$ to L_{global}. Stem, head and pooling are separate factors. Its skeleton $s_\ell \boldsymbol{I} + \boldsymbol{W}_{\ell,2}\boldsymbol{W}_{\ell,1}$, both branch activations deleted, enters L_{linear}. The decomposition yields one registered prediction about architecture, tested in Section 3.2. A residual block carries its skip past every activation. So we expected gating to weaken under skips, and the looseness to relocate into composition. This is intuition, not a theorem, since the branch can cancel the skip. One scoped floor holds, and it covers only the dead $c = 0.5$ network (Appendix N).

2.4 PROTOCOL

The gap is a ratio of a product of layer norms to a witnessed sensitivity. Only one handle acts on the numerator without touching the architecture or the loss: each layer's spectral norm. Holding every layer at a ceiling caps the product, and leaving training otherwise untouched lets the denominator answer. A cap profile is a vector $\boldsymbol{c} \in \mathbb{R}^n_{>0}$ whose entries c_ℓ are those per-layer ceilings. We write c when every layer shares one budget, as throughout unless a sweep says otherwise. The capped constraint set is $\{\theta : \|\boldsymbol{W}_\ell\|_2 \le c_\ell \ \forall \ell\}$, and the hard cap enforces it by scalar rescaling after each optimizer step,

$$\boldsymbol{W}_\ell \ \leftarrow \ \boldsymbol{W}_\ell \cdot \min\bigl(1,\, c_\ell / \|\boldsymbol{W}_\ell\|_2\bigr). \tag{5}$$

That rescaling is the whole intervention (Algorithm 1).

The grid holds 380 training runs across 358 unique configuration-seed cells. Later registered blocks add runs outside it. The core configuration is a plain six-convolution network of base width 64 on CIFAR-10. Its channels double from 64 to 256 over three stages. It is extended along depth, width, a residual family and three further datasets. Each configuration has two to four seeds, and the reference decomposition reruns two of them. Numbers quoted in the text for a route are geometric means over those two seeds unless a table says otherwise. Each is trained with SGD, AdamW, and Muon along four routes. The routes are unconstrained, a spectral penalty, spectral normalization, and the hard cap equation 5. Each such combination is an arm. A fifth route fixes the bound by construction. Its reference network is made orthogonal by the Cayley transform (Trockman & Kolter, 2021). We call it the Cayley anchor, since it fixes the tight end at $G \approx 1$. Three further published constructions run beside it. Each optimizer runs at one rate, fixed before any run and held across every arm: its declared rate. They are 0.05 for SGD, 10^{-3} for AdamW and 0.02 for Muon, at batch size 256 with no schedule. Section 3.5 reports what tuning changes. The witness set for L_{local} holds 64 held-out test inputs with exact Jacobians. Three stronger instruments raise L_{local} by at most $1.62\times$ and move no ordering (Appendix E). They are the full test set, a secant search over input pairs, and input-space gradient ascent on the Jacobian norm. Every grid run trains for 8,000 steps under weight decay 5×10^{-4}, decoupled for AdamW and Muon and coupled for SGD. We use six terms throughout. A route is one way of enforcing spectral control, and an arm is one optimizer on one route. The grading is the ordering of the three optimizers' gaps and the distances between them. The operating budget is $c = 2$, and the witnesses are the 64 inputs behind L_{local}. A gauge is a rescaling of the weights that leaves the function unchanged. Every quantitative prediction was registered before its test run, as a direction or band committed to the code history. Each was scored as it returned, misses included. The ledger marks two exceptions to the registration rule (Appendix R), and Table 1 lists every setting.

3 RESULTS

This section covers the grid of Fig. 1c in the order the claims were made. Its cells are geometric means over seeds. It asks what the optimizer does to the gap and where the gap lives. It then asks what enforcement does and what repair recovers. Stress tests and the two levers follow (Table 25).

3.1 THE OPTIMIZER SETS THE GAP

The optimizer is the one part of a training recipe that can be swapped by itself. Architecture, data, loss and step count stay fixed. If the gap moves with it, looseness is not a property of the model alone. The shared scale is CIFAR-10 at width 64 and depth 6 over 8,000 steps. There, unconstrained training multiplies G by $3\times$, $60\times$ and $2{,}300\times$ under SGD, AdamW and Muon (Table 6). These are geometric means over seeds. The same ordering holds on CIFAR-100 and on the residual family (Table 6, Table 7). With skips, the spread between SGD and Muon widens to nine orders. Section 3.5 reports where each result stops.

The grading, the ordering of the three optimizers and the distance between them, also scales with depth. G grows roughly exponentially. Unconstrained Muon moves from about 7×10^3 at four convolutions to above 10^{10} at eight. Capped runs stay below 20 at every depth and gain clean accuracy as they deepen (Fig. 4). Network width, 16 to 256, is harder to read, since its effects on both bounds are non-monotone here (Khromov & Singh, 2024).

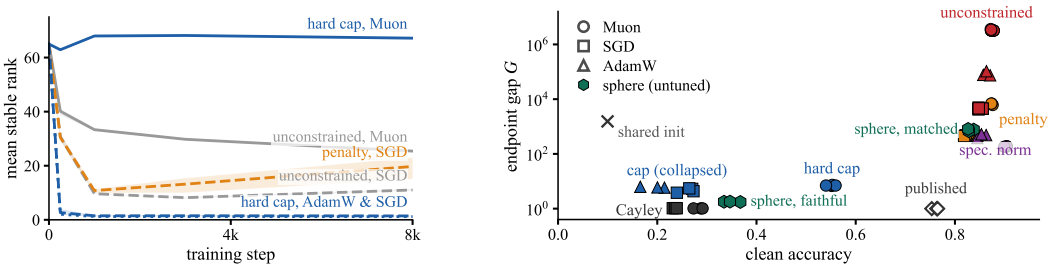

Figure 2: **The response to enforcement. (a)** Mean stable rank along training. Under the cap, SGD and AdamW collapse toward rank one and stall, while Muon holds near 67. **(b)** Clean accuracy against G for the 47 runs of the family, color the route and marker the optimizer. Tightness is bought with accuracy, and only Muon keeps both under the cap. Hollow diamonds are published by-construction results at $G \approx 1$ by definition (Trockman & Kolter, 2021; Singla & Feizi, 2021; Xu et al., 2022). Seed bands, the shared initialization and four undrawn spectral-norm SGD seeds are in Section 3.3, Appendix G and Appendix S.

The registration extends beyond the trained grid (Fig. 7, Table 25). At depth 10, Muon is within a factor of 1.3 of its registered center. AdamW's seeds at $10^{11.80}$ and $10^{12.12}$ straddle the band's edge. We record them as a partial miss rather than pool them. SGD diverged at the declared rate.

3.2 Two causes, and the looseness moves between them

A gap says how loose a certificate is, not why. The exact factorization $G = M \times A$ answers the second question. M is what the bound loses by multiplying layers as though their principal directions were aligned. A is what it loses by counting directions the activations switch off. Which factor carries the looseness says which one an intervention must move. The split needs endpoint weights. They were saved on the Muon runs, and on plain-stack corners we reran under the other two optimizers. On the Muon arms, unconstrained networks carry the gap in gating and capped networks carry it in composition. Skips relocate it as registered (Fig. 5, Table 11). Under the cap the witnessed Jacobian matches the skeleton norm within two percent in plain stacks. Under skips it exceeds that norm. So gating contributes no net looseness. A stronger witness strengthens the reading (Appendix E). The gating has not gone (Section 3.6). What remains is composition slack, which the gating partly offsets when skips are present. In the capped residual arms M overstates the gap threefold. There it measures the size of the composition term, not its location. The other two optimizers are decomposed at the plain-stack corners. Which factor is larger without the cap depends on the optimizer: gating under AdamW, misalignment under SGD. Under the cap A is near one on all three, 0.57 to 1.79 over six endpoints. Without the cap it runs 21 to 6,500. Those two capped arms are the collapsed ones of Section 3.3, so only Muon is both tight and trained.

The gap is not an artifact of the instruments. Under the tighter LipSDP bound (Fazlyab et al., 2019) on a dense network, the unconstrained Muon gap still grows about $40\times$. Capped arms stay flat (Table 21). The grading is claimed for convolutional families only (Appendix T).

3.3 The response to enforcement

Within every optimizer the routes harden in the registered order (Fig. 2b, Table 9). Unconstrained Muon ends at 3.2×10^6 and the penalty stratifies by weight. Spectral normalization ends near 190 at clean accuracy 0.90. The hard cap ends near 7, or 5.8 under the strongest witness (Appendix E).

Under the hard cap the roles invert (Fig. 2a). Unconstrained references drift from stable rank 65 at initialization to 11 for SGD and 25 for Muon. This is training's known low-rank bias (Huh et al., 2023; Galanti et al., 2025). Under the cap, SGD and AdamW collapse toward stable rank one. Stable rank is squared Frobenius over squared spectral norm, averaged over the seven weight layers. Both stall at every declared rate on both datasets, while capped Muon holds stable rank near 67 and keeps learning. The difference is in the update, not the constraint. An orthogonalized step spreads over every direction at once. Repeated projection then has no single direction to squeeze the weights onto. The two deaths differ in reach. At budgets four and eight times looser capped SGD still dies to

chance. Capped AdamW there trains to 0.55 and 0.82, against capped Muon's 0.81 and 0.84. On a dense network at the operating budget it trains level with capped Muon (Appendix G). SGD's death is budget-independent. AdamW's is confined to convolutional stacks at the operating budget. Gouk et al. (2021) trained SGD under the same projector with batch normalization. So the candidates narrow to normalization, architecture, the rate, or a logit temperature τ. Constrained networks are known to need one (Béthune et al., 2022). It multiplies the logits before the loss. A capped network with a bound near 4 compresses its logits. So we ran a registered factorial over temperature and rate (Appendix J). Rescue was defined in advance as accuracy within 15 points of capped Muon's. Neither axis alone rescues (Fig. 6). Only both together lift them, to 0.49 and 0.52 at 1,000 steps and to 0.62 and 0.57 by 8,000. The death is therefore reported at the declared rate and temperature. A registered extension to 8,000 steps settles the rate axis: rate alone buys delay, not survival. The tempered cells that do clear it certify nothing at any radius. Nor does the untempered cell there.

Half a mechanism was registered before the runs that test it (Appendix I). The cap acts after the update, so the next weight is proportional to $\boldsymbol{W} + \Delta\boldsymbol{W}$. A step with one dominant direction and a size comparable to the budget drags the weights onto it. The three update rules differ exactly in the step's rank and in how its size scales with the budget. SGD's step is the gradient, which grows with the product of the other layers' norms. A looser budget should kill it sooner. AdamW's step is sign-like and fixed in size, so a looser budget should let it live longer. Muon's step is full-rank by construction and has nothing to collapse onto. Table 12 scores the three predictions. The collapse step is the step at which mean stable rank first falls below 2. SGD's falls from about 290 at $c = 2$ to a single step at $c = 8$. AdamW's rises from about 260 to beyond the 1,000 logged steps. Muon's step keeps stable rank above 80 at every budget. The size prediction is partial. At $c = 2$, SGD's step is a fifth of the budget, so that death accumulates along a persistent direction. At $c = 8$, the step-0 logits are already saturated. That is a scale pathology a loss-scale account predicts alike. Why the direction persists at $c = 2$ stays open. The collapse is not an artifact of early training. We registered that annealing the budget from 8 to 2 would rescue capped AdamW. It does not (Appendix I).

The spectral sphere optimizer of Xie et al. (2026) was added after the grid was fixed. It reaches $G = 751$ at clean accuracy 0.831, where our cap at the matching budget reaches 0.551. The accuracy price thus depends on how the constraint is enforced (Appendix G).

3.4 After-training repair recovers a factor of two to six

No function-preserving per-layer rescaling changes the gap. With 1-homogeneous activations, inversely rescaling two layers is invisible to the function once biases rescale in step. The symmetry is classical (Neyshabur et al., 2015; Dinh et al., 2017; Du et al., 2018; Stock et al., 2019; Gonon et al., 2025) and reappears as the rescaling lemma of Kuang et al. (2026). Proposition 2, proved in Appendix H, says two cap profiles with the same product are related by a covariant rescaling. That rescaling leaves the function, both instruments, G and every certified radius identical. Only the total budget is physical (Appendix L). By Proposition 3, covariantly projecting a trained network onto any profile rescales the function by a scalar. Nothing we measure changes. The projector of Algorithm 1 leaves biases unchanged, so we measure it instead. It moves the gap by a factor of six, five orders short of the capped regime (Appendix O). Clipping singular values changes the function (Appendix P). Neither proposition covers the neuron-wise gauge, which also preserves the function (Neyshabur et al., 2015; Kuang et al., 2026). Optimized on the unconstrained endpoints of Fig. 1b, it recovers at least 2.1 to 2.2. G stays above 1.4×10^6 on both seeds (Appendix M).

The gap also follows the optimizer acting now. We swapped optimizers at the midpoint of an unconstrained run, with self-swap controls. Every swapped run moves more than 1.6 decades, factors of ten, over its second half. Both controls stay inside their bands. The Muon-to-SGD arm falls 1.7 decades while its clean accuracy rises. The registration had capped it at 0.2 decades (Appendix K). Training undoes what training did. Covariant projection provably cannot (Proposition 3), and the weight-only projector recovers only a factor of six.

3.5 Stress and scale

Two sweeps stress the grading. A horizon of 40,000 steps leaves SGD's gap flat while AdamW and Muon keep rising (Table 20). Tuning each optimizer to its rate compresses the spread to about one order (Table 8). Tuned, the optimizer still moves the gap, but Muon's falls below AdamW's.

We repeated the unconstrained and capped arms on ImageNet-64, a thousand classes. The grading persists and widens. The gap is smallest under SGD, then AdamW, then Muon, on two seeds each. Neighbours are four and six orders apart, and the ends ten, against about three at the shared scale. That is six orders beyond the registered band, in the direction of the effect. The cap alone does not train there. Capped Muon reaches 0.03 against 0.23 unconstrained, a registered miss. A logit temperature of 40 at a tenth of the rate rescues it. This grid does not separate the two. At budget 4 the capped network reaches 0.39 and 0.38 on two seeds. That passes every unconstrained arm, the best being SGD at 0.32, at a gap twelve orders below Muon's. At budget 2 one seed matches Muon's 0.23 under a bound of 1.7×10^3 against 8.0×10^{21}. The untempered control is at that bound and reaches 0.03. Nothing verifies at any radius, and live gaps run 1.4×10^3 to 1.7×10^6 (Appendix Q).

3.6 TIGHTNESS IS NOT LINEARITY

Capped tightness could be trivial if the cap simply linearized the network. A linear network has $A = 1$ by construction, and under the cap A is close to one. Gating entropy settles it. It is the entropy of a unit's on-off pattern, zero for a unit that never switches. At the operating budgets it is 0.42 to 0.51 per unit, above the unconstrained 0.37. So the cap removes the bound's overcount of gated-off directions, not the gating. At $c \leq 1$ the entropy instead collapses to 0.06–0.09. At $c = 8$ the cap prunes, killing 41 to 44 percent of units (Appendix Q). The budgets where tightness is linearity reach at most 0.19 clean accuracy. The verified-accuracy peak at $c = 1$ is a warning.

3.7 THE TWO LEVERS

Gap and clean accuracy rise together with the budget, a frontier with no interior optimum. Verified accuracy does not follow them. It peaks at $c = 1$, 0.140 without a margin term against 0.0015 at $c = 2$. A verified-accuracy number means nothing without its clean accuracy beside it (Appendix Q). Certified accuracy is a product of two levers, a tight bound and a trained margin. The cap moves only the bound. G falls from 3.2×10^6 to 7, while verified accuracy at $\varepsilon = 36/255$ stays at 0.15% (Table 9). Of the trained routes we certified, only two verify anything without a margin term. They are the hard cap and the spectral sphere at its faithful radius (Appendix G). The by-construction routes verify more. The Cayley anchor reaches 0.217, at a bound fixed by design up to the cap on stem and head. Even there training decides how much of that bound is usable. We retrained four published constructions, three seeds per optimizer. Muon verifies more than SGD on every one, by four to five points. The ranges are disjoint under either certificate (Appendix S). One of the four never orthogonalizes, so the effect is not orthogonalization's. Certified training works the margin lever (Tsuzuku et al., 2018; Leino et al., 2021), and this paper isolates the other. The capped run of Fig. 1b keeps 0.55 clean accuracy against 0.88, or 0.62 when tuned. A logit temperature of 40 lifts capped Muon to 0.77 at the same certificate. The gap pays, rising from 7 to 58 through misalignment alone. The recovered accuracy is uncertified, and capped SGD replicates it (Appendix J).

The levers compose, and the composition was itself a registered prediction. Adding a constant competing-logit margin term (Tsuzuku et al., 2018) to capped Muon raises verified accuracy at $\varepsilon = 36/255$ about $35\times$. It rises from 0.0015 to 0.054 at unchanged G, for about two points of clean accuracy. GloRo (Leino et al., 2021) reaches 0.58 at the same radius by training both levers together. That is a certified system, a different question from ours. We reproduce that number in our code (Appendix B). Moving its hidden layers to Muon, at Muon's own rate, costs verified accuracy, 0.41 against 0.59. It inflates the gap from 2.8 to 10. Under a capped cross-entropy the same optimizer is the one that still trains. So the optimizer moves tightness under both objectives, and the sign depends on the objective. Adding the cap to that Muon arm recovers 0.54 and takes the gap to 2.4. That is tighter than the published recipe, at five points less verified accuracy. Every clause was registered before the runs. A fourth arm, reported and not scored, caps the recipe under its own Adam. It reaches 0.56 verified, and its gap falls from 2.8 to 1.6. So the cap tightens under a certified loss as under cross-entropy. It cuts the gap $4.4\times$ with Muon and $1.8\times$ with Adam, so the registered halving holds only with Muon. Our own grid trains nothing for verified accuracy. A gap can be attributed to a cause only while the other levers are held fixed. The budget alone at $c = 1$ exceeds what the margin term buys at $c = 2$, at a third of the clean accuracy, so neither lever dominates.

4 Related work

Bounding the constant, and tracking it. The product of layer spectral norms is the bound everything else refines (Virmaux & Scaman, 2018). Refinements use semidefinite, polynomial, and local relaxations (Fazlyab et al., 2019; Latorre et al., 2020; Jordan & Dimakis, 2020; Xu & Sivaranjani, 2026). The convolutional norm is exact under circular padding (Sedghi et al., 2019). Along training, Khromov & Singh (2024) track both bounds. They find adaptive optimizers driving the upper bound higher than SGD. Ma et al. (2023) make that comparison their subject. The closest prior observation to our capped result is in Leino et al. (2021). Under their certified loss, an optimization-based lower bound reaches 83 percent of the product bound on MNIST. It reaches about 70 percent on CIFAR-10, where a standard model is orders below. Downward pressure on the bound thus tightens it. The ratio is implicit in those paired curves, but never the object. Nor is it graded by update geometry or moved under enforcement. Kuang et al. (2026), concurrently, take the same ratio as their object on three-layer MNIST networks at one endpoint. They tighten it by penalizing the product of the norms. After training, they lower the bound by the function-preserving per-neuron rescaling that Section 3.4 measures on our endpoints. They measure nothing by optimizer or under hard projection. Newhouse et al. (2025) train transformers under enforced bounds and find Muon reaching a lower bound than AdamW at equal performance. They give a guarantee for Muon and none for AdamW, the Muon half of the inversion we report. They run no SGD and no stable rank along training. Under their per-layer rescaling, AdamW trains on CIFAR-10 MLPs. Ours dies on convolutional stacks and trains on a dense one (Section 3.3). So the difference is architectural. Hyperball (Wen et al., 2026) wraps Adam and Muon and trains to 1.2B parameters on a Frobenius sphere, not a spectral one. Both Gouk et al. (2021), with batch normalization, and Khromov & Singh (2024), with a logit temperature, trained SGD under per-layer budgets. Section 3.3 scopes our collapse.

Tightness by construction, spectral control, and optimizer geometry. The by-construction line (Cissé et al., 2017; Anil et al., 2019; Trockman & Kolter, 2021; Singla & Feizi, 2021; Prach & Lampert, 2022; Araujo et al., 2023) and its certified pipelines (Leino et al., 2021; Hu et al., 2023) accept a constrained function class (Huster et al., 2018; Neumayer et al., 2023). We retrain four of them under our protocol (Appendix S). Our routes descend from the spectral penalty (Yoshida & Miyato, 2017), spectral normalization (Miyato et al., 2018), and the projector of Gouk et al. (2021), here an instrument. Optimizers interpreted as steepest descent under matrix norms (Bernstein & Newhouse, 2024; Large et al., 2024) are realized in Muon (Jordan et al., 2024; Liu et al., 2025). Muon optimizes under an implicit spectral-norm constraint (Chen et al., 2026). Training's low-rank bias is documented (Arora et al., 2019; Huh et al., 2023), and we find no prior measurement of stable rank under hard spectral projection. The full survey is Appendix A.

5 Conclusion

Three rules follow for convolutional networks. Choose the optimizer with the care given to the architecture. At fixed architecture it alone moves the gap three orders at declared rates, one when tuned. Never pay for post-hoc repair. By symmetry, Propositions 2 and 3 close profile search and covariant projection. By measurement, the weight-only projector returns a factor of six and the neuron-wise gauge a factor of two. Measure G before upgrading the verifier. By Proposition 1, a small gap means no global-constant verifier can help, so retrain. A collapsed network also reports a small gap. A large gap is indeterminacy, not guaranteed headroom.

Limitations. The boundary is the grid of Section 3.1. At ImageNet-64 the meter carries over and the grading widens. The cap trains there only with a temperature and a lower rate, not separated, and then certifies nothing. At the declared rate $\tau = 1$ trains and $\tau = 10$ dies, on one seed each. The verified accuracies are isolation measurements, not a certified system. We quote LiResNet and LOT and did not retrain them (Appendix S). Dot-product self-attention has no finite Lipschitz constant on an unbounded domain (Kim et al., 2021). A transformer needs the ℓ_2 variant before a gap exists. Token mixing, the nearest family we ran, keeps the cap effect but not the ordering (Appendix T).

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks in the required-disclosure category. We used them to draft text for propositions we had formulated. We used them to draft the pre-registration documents from hypotheses we had set. We also used them to check those drafts adversarially before we approved them. We used them to write some scripts and measurement code to our specification. We did not use them to develop the theoretical framework or to formulate mathematical claims. They supplied no ingredient of any proof. Synthetic data, translation, dataset cleaning and qualitative analysis are not applicable to this work. In the recommended-disclosure category, we used them to create and edit figure scripts and software code. We also used them to draft and edit prose, to summarize literature, to search for information, and to format references.

We have reviewed all AI-assisted work. We set every hypothesis and approved every prediction, band and scoring rule before its runs fired. The ledger marks two exceptions. We produced and interpreted every measurement. We checked and corrected every proof. We ran all code and checked it against the experimental outputs, and every reported number traces to a released aggregate. We read every cited work and verified its claims and its reference entry against the source. We take responsibility for the final content of this work, including every text, claim and artifact produced with generative AI.

ETHICS STATEMENT

This work studies the tightness of Lipschitz-based robustness certificates on standard image-classification benchmarks. It involves no human subjects, collects no data, and releases no dataset. The datasets used, CIFAR-10, CIFAR-100, TinyImageNet, and a downsampled ImageNet, are public and used under their existing terms. The findings concern when a robustness guarantee is loose and how training affects it. We see no path by which they enable harm beyond what the underlying certification methods already enable. One practical implication is methodological rather than ethical. We state it in Section 3.7 because it could mislead if omitted. A network that has collapsed to chance can report a higher verified-accuracy figure than a trained one. Verified accuracy should never be read without clean accuracy beside it. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The paper is self-contained. The instruments are defined in Section 2 and Algorithm 1. Every training setting is in Appendix B, and every proof is given in full. The predictions the text relies on are printed in Table 26 with their outcomes, and the release carries every registered row. The supplementary material carries the code, the script and data behind every figure and table, and the registration files. A reproduction guide gives the commands for each experiment. Registrations are dated and labeled by code state rather than by commit hash, since a hash would identify the repository. Some figure-data files were placed by hand from run logs, and the guide marks each. Per-step training logs and endpoint weights are omitted for size. The guide gives the commands that regenerate the logs. Every number drawn from the weights is in the released data.

REFERENCES

Cem Anil, James Lucas, and Roger B. Grosse. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, pp. 291–301. PMLR, 2019.

Alexandre Araujo, Aaron J. Havens, Blaise Delattre, Alexandre Allauzen, and Bin Hu. A unified algebraic perspective on Lipschitz neural networks. In *International Conference on Learning Representations*, 2023.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pp. 573–582. PMLR, 2019.

Pierfrancesco Beneventano, Mahmoud Abdelmoneum, and Tomaso Poggio. The spectral dynamics and noise geometry of Muon. *arXiv preprint arXiv:2606.08388*, 2026.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Louis Béthune, Thibaut Boissin, Mathieu Serrurier, Franck Mamalet, Corentin Friedrich, and Alberto González-Sanz. Pay attention to your loss: understanding misconceptions about Lipschitz neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20077–20091. Curran Associates, Inc., 2022.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026.

Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pp. 854–863. PMLR, 2017.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Blaise Delattre, Quentin Barthélemy, Alexandre Araujo, and Alexandre Allauzen. Efficient bound of Lipschitz constant for convolutional layers by Gram iteration. In *International Conference on Machine Learning*, pp. 7513–7532. PMLR, 2023.

Blaise Delattre, Quentin Barthélemy, and Alexandre Allauzen. Spectral norm of convolutional layers with circular and zero paddings. *arXiv preprint arXiv:2402.00240*, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Tomer Galanti, Zachary S. Siegel, Aparna Gupte, and Tomaso A. Poggio. SGD with weight decay secretly minimizes the ranks of your neural networks. In *Conference on Parsimony and Learning*, pp. 1388–1412. PMLR, 2025.

Matteo Gamba, Hossein Azizpour, and Mårten Björkman. On the Lipschitz constant of deep networks and double descent. In *British Machine Vision Conference*, 2023.

Paul Geuchen, Dominik Stöger, Thomas Telaar, and Felix Voigtlaender. Upper and lower bounds for the Lipschitz constant of random neural networks. *Information and Inference: A Journal of the IMA*, 14(2):iaaf009, 2025. doi: 10.1093/imaiai/iaaf009.

Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A rescaling-invariant Lipschitz bound based on path-metrics for modern ReLU network parameterizations. In *International Conference on Machine Learning*, pp. 20047–20074. PMLR, 2025.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021. doi: 10.1007/s10994-020-05929-w.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Kai Hu, Andy Zou, Zifan Wang, Klas Leino, and Matt Fredrikson. Unlocking deterministic robustness certification on ImageNet. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42993–43011. Curran Associates, Inc., 2023.

Kai Hu, Haoqi Hu, and Matt Fredrikson. LipNeXt: Scaling up Lipschitz-based certified robustness to billion-parameter models. In *International Conference on Learning Representations*, 2026.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J. Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local Lipschitz bounds. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22745–22757. Curran Associates, Inc., 2021.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023.

Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the Lipschitz constant as a defense against adversarial examples. In *ECML PKDD 2018 Workshops*, pp. 16–29. Springer International Publishing, 2018. doi: 10.1007/978-3-030-13453-2_2.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Matt Jordan and Alexandros G. Dimakis. Exactly computing the local Lipschitz constant of ReLU networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7344–7353. Curran Associates, Inc., 2020.

Tyler Kastner, Nimrod De La Vega, and Amir massoud Farahmand. Singularclip: Preventing spectral collapse to maintain plasticity in continual and reinforcement learning. *arXiv preprint arXiv:2608.18319*, 2026.

Grigory Khromov and Sidak Pal Singh. Some fundamental aspects about Lipschitz continuity of neural networks. In *International Conference on Learning Representations*, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Simon Kuang, Yuezhu Xu, S. Sivaranjani, and Xinfan Lin. Demystifying Lipschitz verification: Positive matrices, negative results. *arXiv preprint arXiv:2603.28113*, 2026.

Bo-Han Lai, Pin-Han Huang, Bo-Han Kung, and Shang-Tse Chen. Enhancing certified robustness via block reflector orthogonal layers and logit annealing loss. In *International Conference on Machine Learning*, pp. 32246–32277. PMLR, 2025.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. In *Advances in Neural Information Processing Systems*, volume 37, pp. 73501–73548. Curran Associates, Inc., 2024.

Fabian Latorre, Paul Rolland, and Volkan Cevher. Lipschitz constant estimation of neural networks via sparse polynomial optimization. In *International Conference on Learning Representations*, 2020.

Sungyoon Lee, Jaewook Lee, and Saerom Park. Lipschitz-certifiable training with a tight outer bound. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16891–16902. Curran Associates, Inc., 2020.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021.

Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B. Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in Lipschitz constrained convolutional networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Róisín Luo, James McDermott, Christian Gagné, Qiang Sun, and Colm O'Riordan. Optimization-induced dynamics of Lipschitz continuity in neural networks. *arXiv preprint arXiv:2506.18588*, 2025.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Avery Ma, Yangchen Pan, and Amir massoud Farahmand. Understanding the robustness difference between stochastic gradient descent and adaptive gradient methods. *Transactions on Machine Learning Research*, 2023.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Julian McGinnis, Suprosanna Shit, Florian A. Hölzl, Paul Friedrich, Paul Büschl, Vasiliki Sideri-Lampretsa, Mark Mühlau, Philippe C. Cattin, Bjoern Menze, Daniel Rueckert, and Benedikt Wiestler. Beyond uniformity: Regularizing implicit neural representations through a Lipschitz lens. In *International Conference on Learning Representations*, 2026.

William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. In *Empirical Methods in Natural Language Processing*, pp. 1766–1781. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.133.

Laurent Meunier, Blaise Delattre, Alexandre Araujo, and Alexandre Allauzen. A dynamical system perspective for Lipschitz neural networks. In *International Conference on Machine Learning*, pp. 15484–15500. PMLR, 2022.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Toby Murray. Lipschitz-based robustness certification under floating-point execution. *arXiv preprint arXiv:2603.13334*, 2026.

Sebastian Neumayer, Alexis Goujon, Pakshal Bohra, and Michael Unser. Approximation of Lipschitz functions using deep spline neural networks. *SIAM Journal on Mathematics of Data Science*, 5(2):306–322, 2023. doi: 10.1137/22M1504573.

Laker Newhouse, R. Preston Hess, Franz Cesista, Andrii Zahorodnii, Jeremy Bernstein, and Phillip Isola. Training transformers with enforced Lipschitz constants. *arXiv preprint arXiv:2507.13338*, 2025.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning*, volume 267, pp. 49069–49104. PMLR, 2025.

Bernd Prach and Christoph H. Lampert. Almost-orthogonal layers for efficient general-purpose Lipschitz networks. In *European Conference on Computer Vision*, pp. 350–365. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-19803-8_21.

Bernd Prach, Fabio Brau, Giorgio C. Buttazzo, and Christoph H. Lampert. 1-Lipschitz layers compared: Memory, speed, and certifiable robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24574–24583. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02320.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon and Scion great again! (bridging theory and practice of LMO-based optimizers for LLMs). *arXiv preprint arXiv:2505.13416*, 2025.

David Rolnick and Konrad P. Kording. Reverse-engineering deep ReLU networks. In *International Conference on Machine Learning*, pp. 8178–8187. PMLR, 2020.

Amartya Sanyal, Philip H. S. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and GANs. In *International Conference on Learning Representations*, 2020.

Hanie Sedghi, Vineet Gupta, and Philip M. Long. The singular values of convolutional layers. In *International Conference on Learning Representations*, 2019.

Sabari Shanmugam, Nick Barnes, and Kerry Taylor. Spectral gradient orthogonalization improves differentially private training at scale. In *European Conference on Computer Vision*, pp. 636–652. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37252-9_35.

Zhouxing Shi, Yihan Wang, Huan Zhang, J. Zico Kolter, and Cho-Jui Hsieh. Efficiently computing local Lipschitz constants of neural networks via bound propagation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2350–2364. Curran Associates, Inc., 2022.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In *International Conference on Machine Learning*, pp. 9756–9766. PMLR, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Pierre Stock, Benjamin Graham, Rémi Gribonval, and Hervé Jégou. Equi-normalization of neural networks. In *International Conference on Learning Representations*, 2019.

Teodor-Mihai Stupariu and Andrei Manolache. How the optimizer shapes learned solutions in equivariant neural networks. *arXiv preprint arXiv:2605.27662*, 2026.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the Cayley transform. In *International Conference on Learning Representations*, 2021.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

Ruigang Wang and Ian R. Manchester. Direct parameterization of Lipschitz-bounded deep networks. In *International Conference on Machine Learning*, pp. 36093–36110. PMLR, 2023.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29909–29921. Curran Associates, Inc., 2021.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them II: Hyperball optimization. *arXiv preprint arXiv:2606.16899*, 2026.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018.

Tian Xie, Haoming Luo, Haoyu Tang, Yiwen Hu, Jason Klein Liu, Qingnan Ren, Yang Wang, Wayne Xin Zhao, Rui Yan, Bing Su, Chong Luo, and Baining Guo. Controlled LLM training on spectral sphere. *arXiv preprint arXiv:2601.08393*, 2026.

Xiaojun Xu, Linyi Li, and Bo Li. LOT: Layer-wise orthogonal training on improving ℓ_2 certified robustness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18904–18915. Curran Associates, Inc., 2022.

Yuezhu Xu and S. Sivaranjani. ECLipsE-Gen-Local: Efficient compositional local Lipschitz estimates for deep neural networks. *Transactions on Machine Learning Research*, 2026.

Jun Yan, Weiquan Huang, Jiankai Zuo, Yujian Mo, Xi Fang, Chengliang Wu, and Zeming Wei. When Muon optimizer meets adversarial training: A theoretical and empirical study. *arXiv preprint arXiv:2605.26929*, 2026.

Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Dongmian Zou, Radu V. Balan, and Maneesh Singh. On Lipschitz bounds of general convolutional neural networks. *IEEE Transactions on Information Theory*, 66(3):1738–1759, 2020. doi: 10.1109/TIT.2019.2961812.

A Related work in full

Bounding the constant, and tracking it. Exact Lipschitz constants are NP-hard to compute. The product of layer spectral norms is the bound everything else refines (Virmaux & Scaman, 2018). Refinements include semidefinite relaxations (Fazlyab et al., 2019), polynomial optimization (Latorre et al., 2020), and exact local computation (Jordan & Dimakis, 2020). Scalable local bounds and estimates refine it too (Weng et al., 2018; Shi et al., 2022; Xu & Sivaranjani, 2026). For convolutions, the spectral norm is computable exactly under circular padding (Sedghi et al., 2019). Fast estimators exist (Delattre et al., 2023; 2024). SDP bounds can inherit the product bound's failures (Kuang et al., 2026). We measure the simplest certified bound because it is the one deterministic certificates actually pay. Along training, Khromov & Singh (2024) track upper and lower bounds and find double descent in both. They report that adaptive optimizers drive the upper bound higher than SGD does. Ma et al. (2023) make that comparison their subject. They attribute the robustness of SGD-trained networks to smaller Lipschitz constants. Gamba et al. (2023) connect the constant to double descent at interpolation. Geuchen et al. (2025) bound it at initialization, and Luo et al. (2025) model its evolution under gradient noise. McGinnis et al. (2026) spread a product-bound Lipschitz budget non-uniformly across the layers of implicit neural representations. In their SIREN experiments, the bound grows during training when the initial weight scale is small. Certified training uses bounds as a loss signal (Lee et al., 2020; Huang et al., 2021). That the optimizer moves the bound is therefore established. The ratio is implicit in the paired curves of Khromov & Singh (2024), as the distance between them. It is never made the object there, graded by update geometry, or moved under enforcement. A bound-only measurement cannot see the roles invert once the weights are constrained. The quantity that inverts is not the bound. Kuang et al. (2026), concurrently, take the same ratio as their object on three-layer MNIST networks at one endpoint. They tighten it by penalizing the product of the norms. Separately, they lower the trivial bound after training by optimizing the function-preserving per-neuron rescaling. They track the enclosure along training only on a tiny Iris classifier. They measure nothing by optimizer or under hard projection. Section 3.4 measures their rescaling on our endpoints. Our penalty arm uses the sum the literature practices, not their product.

Tightness by construction, and spectral control. A rich line builds networks whose bound is tight by design (Cissé et al., 2017; Anil et al., 2019; Li et al., 2019; Trockman & Kolter, 2021; Singla & Feizi, 2021; Xu et al., 2022; Prach & Lampert, 2022; Meunier et al., 2022; Araujo et al., 2023; Wang & Manchester, 2023). Certified pipelines scale them (Leino et al., 2021; Hu et al., 2023; Lai et al., 2025; Hu et al., 2026), and a benchmark compares them (Prach et al., 2024). The price is a constrained function class whose limits are known (Huster et al., 2018; Anil et al., 2019; Neumayer et al., 2023). Our object is different. We characterize what sets the gap for standard networks under standard training. One Cayley anchor (Trockman & Kolter, 2021), run under our protocol, places both regimes on one frontier (Table 9). Our enforcement routes have named ancestors: the spectral penalty (Yoshida & Miyato, 2017) and spectral normalization (Miyato et al., 2018). The third is the projection operator of Gouk et al. (2021), applied after every step at a fixed budget. Singular-value clipping reappears concurrently as a plasticity device (Kastner et al., 2026). Residual branches have been constrained for invertibility (Behrmann et al., 2019). Two protocol differences matter for what follows. Gouk et al. (2021) trained SGD and Adam under this projection with batch normalization present. Their per-layer budgets were chosen by search, and they report failure to converge at budgets too tight. Khromov & Singh (2024) renormalized SGD to a per-layer budget with a logit temperature of 10 to 40, and it trained. The double-descent trend was reduced there. The collapse we report is thus a property of our protocol: no normalization layers and a budget of two. Section 3.3 scopes it so. The contribution is not the operator but its use as an instrument.

Optimizer geometry. That the optimizer's geometry selects the solution is classical (Gunasekar et al., 2018). Margin-maximization results imply that gradient descent grows norms slowly but without bound (Soudry et al., 2018; Ji & Telgarsky, 2019; Lyu & Li, 2020). Transformers show that growth in practice (Merrill et al., 2021). That is why every mildness claim here carries its horizon. A recent line reads optimizers as steepest descent under matrix norms (Bernstein & Newhouse,

2024; Large et al., 2024). It reinterprets Shampoo (Gupta et al., 2018) and builds on the spectral scaling condition of Yang et al. (2023). Muon and its relatives realize it (Jordan et al., 2024; Liu et al., 2025; Pethick et al., 2025; Kovalev, 2025; Riabinin et al., 2025). SOAP instead runs Adam in Shampoo's eigenbasis (Vyas et al., 2025). The mechanism of Section 3.3 measures the spectral shape of the update. That shape also matters in Shanmugam et al. (2026), who study differentially private training at scale. They show that orthogonalizing the gradient spectrum is what lets Muon-style updates survive its clipping and noise. There the spectral shape of the update, rather than its magnitude, decides whether training proceeds. Chen et al. (2026) prove that Muon with decoupled weight decay implicitly optimizes under a spectral-norm constraint. How special the orthogonalized spectrum is remains debated (Shumaylov et al., 2026). Concurrently, Xie et al. (2026) constrain weights and updates together, on their diagnosis that controlling updates alone lets weights drift. They report gains at billion-parameter scale. They measure no Lipschitz quantity. Even so, the pairing is the one our capped arms require, and their result is evidence it persists at scale. Yan et al. (2026) study Muon's per-update spectral ceiling for adversarial training, and Muon's spectral dynamics are being mapped (Beneventano et al., 2026; Stupariu & Manolache, 2026). Newhouse et al. (2025) train transformers under enforced bounds and ask which bounds are achievable at a given performance. Their sweep over optimizer and constraint method found Muon reaching a lower bound than AdamW at equal performance. They give a guarantee that capping controls the weight norm under Muon, and none for AdamW. That is half of the inversion we report. They did not run SGD or track stable rank along training. Their object is the bound rather than the gap. Their spectral normalization under AdamW rescales each MLP layer to its budget every step, across a learning-rate sweep. It trained where our capped AdamW on a convolutional stack at the declared rate dies. Hyperball (Wen et al., 2026) rescales Adam's and Muon's weights and updates onto a Frobenius sphere after every step. It trains transformers to 1.2B parameters, under a Frobenius rather than a spectral constraint.

Rank, and other certificate families. Training's low-rank bias, and the spectral structure around it, is documented (Arora et al., 2019; Huh et al., 2023; Galanti et al., 2025; Martin & Mahoney, 2021; Papyan et al., 2020). Stable rank has served as a control (Sanyal et al., 2020). We find no prior measurement of stable rank under hard spectral projection, where that bias becomes functional death. Randomized smoothing certifies a smoothed classifier probabilistically, a different object (Cohen et al., 2019). Per-input verification certifies a fixed network around individual inputs (Zhang et al., 2018; Wang et al., 2021; Gowal et al., 2018). Our subject is instead a global property of the trained weights. Training against the IBP bound makes that bound tight (Gowal et al., 2018). That is the bound-aware loss route, which changes what is trained, where we change only how.

B PROTOCOL DETAILS

Algorithm 1 Capped training and the gap meter. Line 4 is the whole intervention. Lines 7 to 10 are the measurement and change no weights.

Require: layers $\boldsymbol{W}_1 \dots \boldsymbol{W}_n$, cap profile $\boldsymbol{c}$, any optimizer step OPT, witness set $\mathbb{D}_{\text{wit}}$, checkpoints $\mathbb{T}$

1: **for** $t = 0, 1, 2, \dots$ **do**
2: $\quad \boldsymbol{W} \leftarrow \text{OPT}\big(\boldsymbol{W}, \nabla\mathcal{L}(\boldsymbol{W})\big)$ ▷ any optimizer, unchanged
3: $\quad$ **for** $\ell = 1 \dots L$ **do**
4: $\quad\quad \boldsymbol{W}_\ell \leftarrow \boldsymbol{W}_\ell \cdot \min\big(1,\, c_\ell / \|\boldsymbol{W}_\ell\|_2\big)$ ▷ the intervention
5: $\quad$ **end for**
6: $\quad$ **if** $t \in \mathbb{T}$ **then**
7: $\quad\quad L_{\text{global}} \leftarrow \prod_\ell \|\boldsymbol{W}_\ell\|_2$ ▷ certified upper bound
8: $\quad\quad L_{\text{linear}} \leftarrow \|\boldsymbol{W}_n \cdots \boldsymbol{W}_1\|_2$ ▷ bias-free affine skeleton
9: $\quad\quad L_{\text{local}} \leftarrow \max_{\boldsymbol{x} \in \mathbb{D}_{\text{wit}}} \|\boldsymbol{J}(\boldsymbol{x})\|_2$ ▷ witnessed lower bound
10: $\quad\quad G \leftarrow L_{\text{global}}/L_{\text{local}}, \quad M \leftarrow L_{\text{global}}/L_{\text{linear}}, \quad A \leftarrow L_{\text{linear}}/L_{\text{local}}$ ▷ the gap and its two causes
11: $\quad$ **end if**
12: **end for**

The depth extension runs 4, 6, and 8 convolutions, with a registered extension to 10. The width extension runs 16, 32, 64, 128, and 256. The further datasets are CIFAR-100, TinyImageNet, and ImageNet at 64 by 64 resolution. The penalty runs at weight $\beta \in \{0.001, 0.01, 0.1\}$, written β to keep λ for the gauge factor. Checkpoints fall at fixed steps, and a registered horizon extension takes the core arms to 40,000 steps (Section 3.5).

Muon uses heavy-ball momentum at 0.95 with Nesterov correction. It orthogonalizes the momentum buffer, viewed as a matrix with convolutions flattened. It uses five quintic Newton-Schulz iterations followed by three cubic polish steps. The orthogonalization is therefore approximate rather than exact. It leaves singular values within a few percent of one at modest condition numbers. No claim in this paper depends on exact orthogonalization. Weight decay is decoupled and applied before the update. The classifier head and all one-dimensional parameters take plain momentum steps without orthogonalization. Capped arms are rescaled onto their budgets at initialization, which is why their step-0 gap differs from the unconstrained arm's. Power iteration appears only inside spectral normalization and the penalty, because that is how those methods are practiced (Miyato et al., 2018; Yoshida & Miyato, 2017). All measurement runs with reduced-precision matrix multiplication disabled.

At each logged checkpoint, L_{linear} is computed by power iteration through the composed map. In the exact split, it comes from a singular value decomposition of the materialized skeleton. Decoupled weight decay at 5×10^{-4} is the setting in which Muon provably optimizes under an implicit spectral-norm constraint (Chen et al., 2026). Unconstrained Muon still inflates G by more than $2{,}000\times$ there. That implicit radius is thus far above certificate-relevant scales.

Table 1: Every training and measurement setting of the grid. Learning rates are the declared rates.

Setting	Value
Learning rate, Muon	0.02
Learning rate, SGD	0.05
Learning rate, AdamW	10^{-3}
Momentum, Muon	0.95
Momentum, SGD	0.9
AdamW (β_1, β_2)	$(0.9, 0.999)$
Weight decay	5×10^{-4}
Learning-rate schedule	constant
Batch size	256
Training steps, main grid	8,000
Training steps, mechanism and temperature-and-rate runs	1,000
Augmentation	reflection padding by 4 pixels, random crop to the input size, horizontal flip with probability 0.5
Convolutions	3×3, stride 1, padding 1, with bias
Channels at width 64, depth 6	64, 64, 128, 128, 256, 256, in three stages of two convolutions
Normalization layers	none
Skip connections	none in the plain stack, pre-activation skips in the residual family
Pooling	2×2 average pooling after each of the three stages, then global average pooling
Pooling attenuation	$1/32 = (1/2)^3 \cdot 1/4$, from three 2×2 average pools and one 4×4 global pool
Head	linear, 256 features to one logit per class, at width 64
Initialization of capped arms	each layer scaled to spectral norm c, then projected to $c/(1 + 10^{-3})$
Projection margin	10^{-3}, so weights are at $c/1.001 \approx 0.999c$
Penalty	$\sum_i \max(0, \sigma_i - c)^2$ over layers i
Spectral normalization	one power-iteration step per forward pass (Miyato et al., 2018)
Spectral sphere radius	$\sqrt{n_{\text{out}}/n_{\text{in}}}$, scale factor 1
Spectral sphere, power-iteration steps	5
Spectral sphere, matrix-sign iterations	8
Input bound	10^6, so effectively unbounded; the input projection is 1-Lipschitz
LipSDP slice	three-layer MLP of width 128 on flattened CIFAR-10 inputs of dimension 3,072

Table 9 quotes GloRo from its main results table and SOC from its Table 2 at LipConvnet-20. LiResNet comes from its Table 1, with generated data. We also ran GloRo's recipe unchanged in our own code, on three seeds. That is 6C2F with MinMax activations, 800 epochs, and Adam at the published 10^{-3}. It reaches clean accuracy 0.773 and kept-epoch verified accuracy 0.588 at $36/255$ under their convention. They publish 0.770 and 0.584. The registered gate was one-sided and two points below their number, and the arm passes. Their converged power-iteration certificate and our sound certificate agree to 0.001 at the endpoint. Our meter reads $G = 2.79$ on their network, between the by-construction anchor and our capped runs. Moving the six hidden convolutions and the first dense layer to Muon changes the optimizer and its rate. Muon runs at 0.01, against Adam's published 0.001. Verified accuracy falls to 0.414, over the seed range 0.407 to 0.427, and clean accuracy to 0.651. These are endpoint values under our sound certificate, against 0.586 for the published recipe. The certified bound rises from about 22 to between 1.4×10^5 and 3.0×10^5, and the gap from 2.79 to 10.4. Both directions were written into the registration before the runs. A third arm adds the hard cap to that Muon arm and changes nothing else (Table 2). The registration asked for two things. The cap should at least halve the Muon arm's gap, and cost no more than a point of verified accuracy. It does both. The gap falls to 2.35, a ratio of 0.23 against the required 0.5. Verified accuracy rises to 0.537, twelve points above the Muon arm. The bound falls to between 12.65 and 12.67, below the published recipe's 21 to 23. The comparison with the Muon arm is clean, since both share the optimizer and its rate. Mean stable rank follows the same path, 48 under Adam, 16 under Muon and 31 with the cap. The capped arm stays below the published recipe on both accuracies, by 0.049 verified and 0.038 clean. Its only advantage is tightness, and both halves

of that trade belong together. A fourth arm caps the published recipe under its own Adam, on three seeds. The registration declared no prediction on it, so it is reported and scores nothing. It reads 0.555 verified at a gap of 1.56, under a bound of 11.19 to 11.45. That is tighter than the capped Muon arm, at 0.018 more verified accuracy and 0.015 less clean accuracy. It is 0.031 verified below the published recipe, where the capped Muon arm is 0.049 below. It is the only arm that changes the cap with the optimizer held fixed. The registration asked that its ratio be printed beside the scored one. Under Adam the cap takes the gap from 2.80 to 1.56, a ratio of 0.56. That misses the 0.5 bar the Muon arms clear at 0.23. So the cap tightens the bound under both optimizers, and the halving belongs to the cap and Muon together.

Table 2: GloRo's recipe in our code, 800 epochs, three seeds per arm, as mean $\pm$ standard deviation. Verified accuracy is at $\varepsilon = 36/255$ under our sound certificate. The second arm moves the hidden layers to Muon at rate 0.01, against Adam's 0.001. The third adds the hard cap to the second. The fourth caps the published recipe under its own Adam. * reported and not scored.

Arm	Clean	Verified	L_{global}	G
Published recipe, Adam	0.773 ± 0.001	0.586 ± 0.001	21 to 23	2.80 ± 0.04
Hidden layers on Muon	0.651 ± 0.007	0.414 ± 0.011	1.4 to 3.0×10^5	10.4 ± 1.1
Muon with the hard cap	0.735 ± 0.002	0.537 ± 0.001	12.65 to 12.67	2.35 ± 0.06
Adam with the hard cap*	0.720 ± 0.003	0.555 ± 0.001	11.19 to 11.45	1.56 ± 0.03

An earlier attempt at this comparison left every arm below its registered liveness floor and scored nothing. That is the purpose of the floors. GloRo's published network uses MinMax activations, and its released snapshot script hardcodes ReLU. In our port of that script the activation is the only difference. ReLU costs about eight points of clean accuracy and ten of verified accuracy, under their own loss. We report that as a diagnostic, and it scores no prediction of ours. Every upper-block row there is at the declared rate. Its G is the main-grid geometric mean over every seed of the route. So the unconstrained row reads $10^{6.53}$ over four seeds. The text's 3.2 million is the two corner seeds that carry the decomposition and the certificates. Tuned for accuracy, the hard cap reaches 0.6215 at its own tuned rate on the six-layer stack. At the declared rate it reaches 0.55. The Cayley anchor runs under the shared protocol rather than a recipe developed for orthogonal architectures.

The cap is not free. At ImageNet-64 it costs about four times an unconstrained run's wall-clock time, 4.9 hours per seed against 72 minutes. On the small network it costs twenty-five to seventy times. The training step is inexpensive there. The per-step projection is a fixed cost that does not shrink with the model. These durations come from run timestamps rather than instrumented timing, so they are accurate to about a minute. Each ratio compares arms run on the same GPU type. Capped SGD is worse again and more variable. Its spectra degenerate, and the projector falls back to a slower path. The projection is an instrument in this paper rather than a proposal. Anyone adopting it should price it at the scale they intend to use.

C CERTIFIED-RADIUS VALIDATION

Table 3 gives the numbers behind this appendix. A certified radius is only meaningful if the prediction is constant inside it. We attack at $0.9\times$ the certified radius with PGD on a margin objective, with three restarts, on the first thousand test inputs. We discard points whose margin falls below 10^{-3}, where the radius is numerically meaningless. Across six arms the attack flips no prediction. The floor drops between zero and three points per arm.

Table 3: Certified-radius validation. Points certified, points dropped by the margin floor, and predictions flipped by PGD at $0.9\times$ the certified radius. Values are mean $\pm$ standard deviation over two seeds.

Arm	Certified	Dropped	Flips at $0.9r$
Hard cap	556.5 ± 9.2	0.0 ± 0.0	0.0 ± 0.0
Cayley anchor	270.5 ± 33.2	1.5 ± 2.1	0.0 ± 0.0
Cap with margin term	541.0 ± 17.0	0.5 ± 0.7	0.0 ± 0.0

D A TWO-DIMENSIONAL CHECK OF EVERY INSTRUMENT

Every quantity in this paper can be recomputed on a laptop on one small problem. No CIFAR gap offers that. The device is that of Khromov & Singh (2024), a two-dimensional input, so that the true constant can be found by exhaustion. A four-layer ReLU network, $2 \to 32 \to 32 \to 32 \to 2$, trains full-batch on two thousand two-moons points for 2,000 steps. It runs under SGD, AdamW, and Muon at declared rates, and under capped Muon at the paper's budget $c = 2$. Muon here is the exact polar factor of the momentum buffer rather than the Newton–Schulz iteration of the large runs. L_{global} is the product of exact spectral norms, and L_{linear} the exact norm of the bias-free skeleton. L_{local} is the maximum Jacobian norm over 64 held-out points from a second sample, the paper's witness convention. $L_\star$ is the maximum over a $1{,}500 \times 1{,}500$ grid on the data box, refined twenty-fold around its top fifty cells. It is thus the constant on the box, not the plane, and the enclosure is stated on the box. A ReLU network is piecewise linear. So the grid maximum is the constant itself, up to regions thinner than the refined spacing. In every arm the refinement found nothing above the coarse grid. The whole check runs in about a minute and a half (Fig. 3, Table 4).

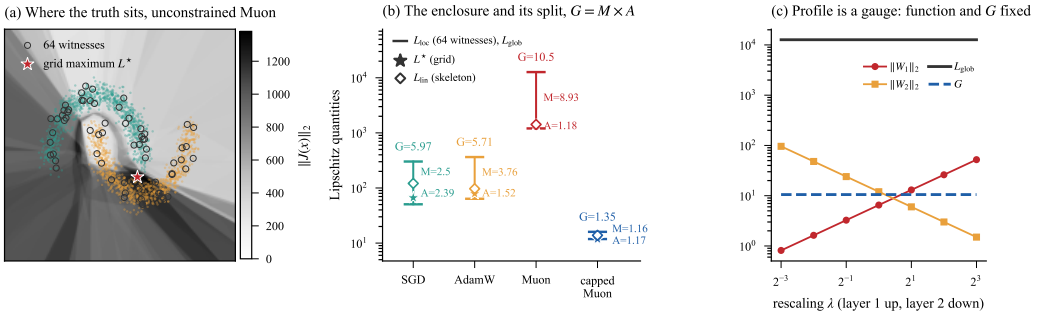

Figure 3: **Every instrument on a problem small enough to check.** **(a)** Witnessed Jacobian norm over the data box for the unconstrained Muon network. The 64 witnesses are circled and the grid maximum $L_\star$ starred. **(b)** The enclosure $L_{\text{local}} \le L_\star \le L_{\text{global}}$ for four arms. The skeleton norm L_{linear} lies on the ladder, so M and A are the two bracket lengths. **(c)** Rescaling layer one by λ and layer two by $1/\lambda$ moves the two norms by 64 each way. The function, L_{global}, and G stay unchanged to every printed digit.

Table 4: The toy in numbers, seed 0. Every arm reaches 1.000 clean accuracy, the capped arm 0.999. G is over the 64 witnesses of the figure. Its range over twenty further random draws of 64 witnesses is in parentheses. The last column is the true enclosure $L_{\text{global}}/L_\star$, which no witness draw moves.

Arm	L_{local}	$L_\star$	L_{linear}	L_{global}	G (draws)	M	A	$L_{\text{global}}/L_\star$
SGD	50.6	66.5	121.1	302.1	5.97 (4.61–6.60)	2.50	2.39	4.54
AdamW	63.7	78.8	96.9	364.2	5.71 (5.13–6.13)	3.76	1.52	4.62
Muon	1,204.4	1,381.4	1,420.7	12,682.7	10.53 (9.31–11.12)	8.93	1.18	9.18
Capped Muon, $c = 2$	11.8	11.8	13.8	16.0	1.35 (1.35–1.47)	1.16	1.17	1.35

The enclosure holds in every arm, and the true constant is near its floor. The 64 witnesses understate it by between 1.00 and 1.31. That is the order of the 1.18 to 1.34 the wider witness sets recover at the shared scale (Appendix E). The split is exact, and so is the gauge. Across $\lambda \in [1/8, 8]$, the two norms move from 0.8 to 52 and from 96 to 1.5. Over that range the function changes by exactly zero, and L_{global} and G hold to every printed digit. Under the cap every layer is at its budget, so $L_{\text{global}} = 2^4$ exactly. The certified bound is within 1.35 of the true constant, at 0.999 accuracy against 1.000 unconstrained.

What the toy does not reproduce is the size of the effects. Its gaps are below ten where the paper's run to millions. Its looseness is composition rather than gating, with M of 2.5 to 8.9 against A of 1.2 to 2.4. The plain CIFAR stacks put gating first (Section 3.2). A two-dimensional input leaves few directions for gating to remove. Muon still ends with the largest bound, 35 to 42 times the other two, and the largest enclosure, twice theirs. SGD and AdamW are within witness noise of each other. Their G ranges over the twenty draws overlap, while Muon's does not. The true enclosure has no such sensitivity. We also draw 200 further random bias-free networks of the same shape. They use He-scaled Gaussian weights and standard normal inputs, with 64 random witnesses each. A falls below one in 43 of them, the non-monotonicity of Section 2.3 as a rate. The toy is a check of the instruments, not evidence for the findings.

E WITNESS SET AND THE LOWER INSTRUMENT

Table 5 gives the numbers behind this appendix. L_{local} is a maximum over a witness set, so it can only understate the true constant. The reported G can therefore only overstate the enclosure width. The question is by how much. We recompute L_{local} for six arms over the 64-image convention and wider sets. These are the full 10,000-image test set, the 10,000-image training set, and PGD-perturbed points at $\varepsilon \in \{8, 36\}/255$. At the shared scale, the widest set exceeds the 64-image convention by between $1.18\times$ and $1.34\times$. The factor does not vary systematically by arm. Adversarially perturbed points never exceed the plain data. The correction is scale-dependent. On ImageNet-64 the same sweep gives $1.58\times$ and $2.04\times$ on the two unconstrained seeds. There the adversarial points do exceed the 64-image maximum, though they stay below the full test set. The gating statistics of Table 23 show the same limit from the other side. In the unconstrained network, 14 percent of units are silent on the 64 witnesses. Each is active somewhere in the full test set. Every such unit is a Jacobian direction the 64-witness set never exercises. That is the case for searching the Jacobian norm directly rather than sampling it. What does not change with scale is the direction. A better witness set can only raise L_{local} and so can only lower G. Every gap reported in this paper is thus an upper estimate of the enclosure width, not a possibly optimistic one. No ordering here is decided by a factor of two. At ImageNet-64, the correction is more than three orders of magnitude below the smallest gap between optimizers it could confound. So no verdict at that scale turns on it. The capped gap of about 7 at the shared scale is at least 5.8 after its own correction. The certificate itself never depends on witnesses.

A two-point secant search, after Kuang et al. (2026), seeks the input pair with the largest ratio of output to input distance. It reaches 348 and 379 on the two unconstrained endpoints. That is $1.44\times$ and $1.62\times$ the 64-image convention, and $1.22\times$ and $1.29\times$ the full test set. On the capped arms it finds nothing beyond the witnesses: $1.02\times$ and $1.01\times$ the convention, and below the test set. So the capped correction stays with the test-set scan. On the penalty arms it reaches $1.20\times$ to $1.24\times$. The search exceeds the convention on the unconstrained arms and falls below plain enumeration on

the capped ones. That is a finding about the instruments, not a defect. A local ascent stalls where the landscape is flat. The capped networks are flat by construction, so enumeration is stronger there. A direct gradient ascent on the Jacobian norm starts from the witnesses, both free and inside the $36/255$ ball. It finds nothing beyond the test-set scan on any of the six arms. On ImageNet-64 the secant search reaches 2,306 and 2,831 on the two unconstrained endpoints. That is $2.60\times$ and $1.85\times$ the 64-image convention. The ball-constrained search exceeds the free one on both seeds, a finite-step basin effect. Every point is a valid lower bound. The corrected gap is L_{global} over the largest instrument, computed once. On the unconstrained CIFAR-10 corners it moves from 3.2×10^6 to 2.2×10^6 and from 3.2×10^6 to 2.0×10^6. So the 3.2 million printed under the 64-witness convention is 2.1×10^6 under the strongest measured witness. On the capped corners the test-set scan remains the largest instrument. The corrected gap there is the floor of 5.8 above. The ImageNet-64 correction is measured at the step-8,000 endpoints, whose weights were saved. It is quoted as representative of the 20,000-step endpoints the text reports, which were not saved. The gauge optimization of Appendix M is a separate axis on the numerator and is never composed with these.

Table 5: Witnessed sensitivity L_{local} under the witness sets, for Muon runs at the shared scale and on ImageNet-64. Values are mean $\pm$ standard deviation over two seeds. The last column divides the largest instrument by the 64-image convention, per seed before averaging. The instruments include the secant search and the Jacobian ascent of the text. The correction grows with scale and never reverses direction.

Arm	64 images	test set	training set	ratio
Hard cap	0.552 ± 0.018	0.672 ± 0.009	0.664 ± 0.001	1.22 ± 0.06
Penalty	29.97 ± 0.28	35.28 ± 1.62	37.87 ± 3.83	1.29 ± 0.08
Unconstrained	237.5 ± 4.8	290.0 ± 6.0	283.3 ± 22.9	1.53 ± 0.13
ImageNet-64				
Unconstrained	$1{,}209 \pm 453$	$2{,}111 \pm 426$	$1{,}716 \pm 20$	2.22 ± 0.53

F THE OPTIMIZER GRADING IN FULL

Section 3.1 quotes the optimizer grading by its headline factors. This appendix (Appendix F) carries every arm. Table 6 gives every unconstrained arm at the shared protocol as a mean and standard deviation over seeds. Table 7 shows the collapse under the cap replicating on each dataset. Figure 4 carries the depth and width sweeps behind the scaling claims.

Table 6: Unconstrained endpoint gap G by family and optimizer, backing the grading sequence of Section 3.1. Values are mean $\pm$ standard deviation over seeds. CIFAR-10 and CIFAR-100 rows are the plain width-64 depth-6 network. The residual row is the implemented skip architecture. All run the shared 8,000-step protocol. Every cell has four seeds except AdamW on the plain network, which has three on each dataset.

Family	SGD	AdamW	Muon
CIFAR-10 plain	$(4.60 \pm 0.26) \times 10^3$	$(8.55 \pm 1.70) \times 10^4$	$(3.38 \pm 0.20) \times 10^6$
CIFAR-100 plain	$(2.35 \pm 0.20) \times 10^4$	$(1.18 \pm 0.17) \times 10^5$	$(3.08 \pm 0.34) \times 10^6$
Residual	$(3.59 \pm 0.80) \times 10^7$	$(4.08 \pm 0.55) \times 10^{12}$	$(4.64 \pm 1.66) \times 10^{16}$

Table 7: The gap collapse replicates across datasets. Endpoint gap G as mean $\pm$ standard deviation over seeds, for unconstrained and hard-capped Muon runs. TinyImageNet capped runs are untuned at this scale and reach accuracy 0.08.

Dataset	Unconstrained	Hard cap	Seeds
CIFAR-10	$(3.38 \pm 0.20) \times 10^6$	7.02 ± 0.24	4
CIFAR-100	$(3.08 \pm 0.34) \times 10^6$	3.87 ± 0.13	4
TinyImageNet	$(4.33 \pm 0.37) \times 10^6$	2.73 ± 0.14	3

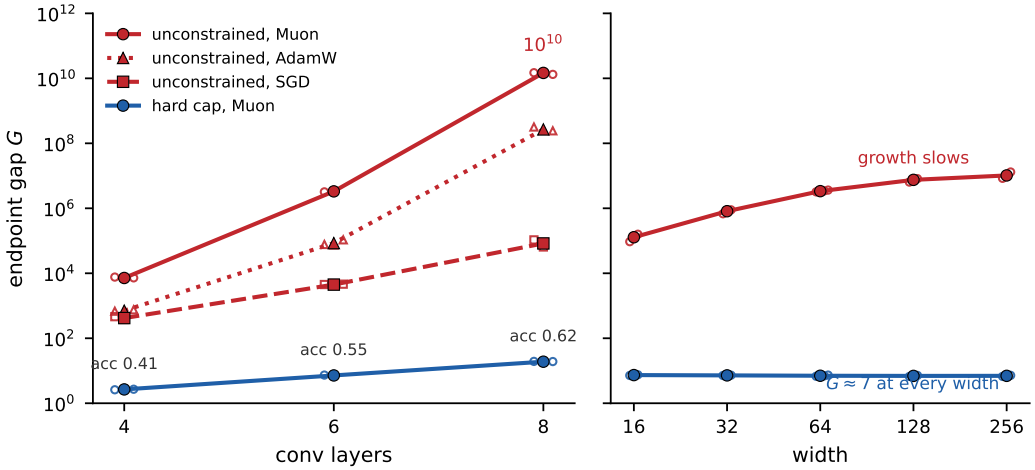

Figure 4: **The scale panels.** *(Left)* Endpoint gap G versus depth on CIFAR-10, with 4, 6, and 8 convolutional layers. Unconstrained Muon rises roughly exponentially in depth, from about 7×10^3 at depth 4 to above 10^{10} at depth 8. Unconstrained AdamW rises between them, from about 7×10^2 to above 10^8. Unconstrained SGD rises far more shallowly over the same 8,000-step runs, from about 4×10^2 to about 10^5. Capped runs lie flat, G below 20 at every depth, while their inherited-lr accuracy rises (0.41, 0.55, 0.62). The depth trend is consistent with the exponential-in-depth looseness reported qualitatively for product bounds (Virmaux & Scaman, 2018). *(Right)* Endpoint gap G versus width at depth 6, five widths from 16 to 256. Unconstrained Muon rises steeply through width 64, from about 10^5 at width 16 to 3.2–3.6×10^6 at width 64. The growth then slows, to 0.8–1.3×10^7 at width 256. Capped runs hold $G \approx 7$ at every width. Color encodes route and marker the optimizer, per Fig. 1. Open marks are individual seeds, lines join seed means, 2 to 4 seeds per point.

The grading at tuned rates. Table 8 gives the unconstrained endpoint gap and clean accuracy at every rate of the learning-rate grid. Each optimizer's accuracy-optimal rate is in bold. SGD's two best rates tie at 0.852, and the bold one has the smaller gap. The bold rates give the comparison Section 3.5 reports beside the declared-rate grading. At those rates the spread is about one order, SGD still mildest and Muon and AdamW exchanging places.

Table 8: Unconstrained endpoint gap and clean accuracy over the learning-rate grid, CIFAR-10 width 64 depth 6. Values are means over the seeds run at each rate, geometric for the gap. Bold marks each optimizer's accuracy-optimal rate.

Optimizer	rate	G	clean	seeds
SGD	0.01	1.3×10^3	0.737	3
	0.05	4.4×10^3	0.852	2
	0.1	1.0×10^4	0.852	3
AdamW	10^{-4}	6.2×10^2	0.657	3
	3×10^{-4}	2.9×10^3	0.789	3
	$\mathbf{10^{-3}}$	7.6×10^4	0.864	2
	3×10^{-3}	1.7×10^6	0.817	3
Muon	**0.005**	1.3×10^4	0.898	3
	0.01	1.3×10^5	0.897	3
	0.02	3.2×10^6	0.875	2
	0.05	1.7×10^9	0.629	3

G THE ENFORCEMENT ROUTES AND THE INVERSION IN FULL

Section 3.3 and Section 3.2 report the response to enforcement by its endpoints. Figure 2b places every arm of the shared-scale family on the accuracy-versus-gap plane. Figure 5 shows where the looseness lives under the exact factorization. Figure 2a traces the stable rank along training that defines the constraint inversion. Four runs of Fig. 2b are not drawn, all of them SGD under spectral normalization. Two seeds diverged to chance accuracy, at gaps of 10^{11} and 7×10^{11}. The other two trained unstably to 0.49 clean accuracy, at gaps of 338 and 450.

Table 9: **Routes to a tight certificate at the shared scale** (CIFAR-10, width 64, depth 6), a frontier of one rate setting. Upper block, our runs at the declared rates, geometric means over 2 to 4 seeds. No margin term is used except in the marked row, so the verified columns show the bound lever alone. Verified accuracy is at $\varepsilon = 36/255$. The product column is the certificate of Section 2.2. The per-pair column is the test the published baselines report, which contains the product test and so is never smaller. Over the 70 endpoints we certified both ways it adds 2.1 points on average. Lower block, published results under their own protocols, margin-trained, quoted as reported and not comparable. The GloRo row also reproduces in our code under their recipe (Appendix B). SOC appears in both blocks. What separates its two rows is the protocol, not the construction (Appendix S).

			Verified acc.	
Route	$\log_{10} G$	Clean acc.	product	per-pair
Trained here, shared protocol				
By construction, Cayley	0.01	0.28	0.217	0.263
By construction, SOC[¶]	0.02	0.25	0.182	0.234
By construction, Björck[¶]	0.02	0.26	0.187	0.241
By construction, SLL[¶]	0.02	0.27	0.199	0.250
Spectral sphere, faithful radius[‡]	0.25	0.35	0.194	0.236
Hard cap[†]	0.85	0.55	0.0015	0.054
Hard cap with a margin term[§]	0.85	0.53	0.054	0.138
Spectral normalization	2.27	0.90	0	0
Spectral sphere, matched budget[‡]	2.88	0.83	0	0
Unconstrained, weight decay 0.1	3.43	0.88	n/r	n/r
Unconstrained	6.53	0.87	0	0
Published, native protocol, margin-trained				
GloRo (Leino et al., 2021)	n/r	0.77	0.58	
SOC (Singla & Feizi, 2021)	n/r	0.76	0.62	
LiResNet (Hu et al., 2023)	n/r	0.82	0.70	

[†] at the declared rate. Tuned for accuracy the same route reaches 0.62 clean (Appendix B). [§] capped Muon with a constant competing-logit margin term, G unchanged (Section 3.7). [‡] run at Muon's declared rate, never tuned for this optimizer. [¶] retrained here, Muon at the screened rate. Both optimizers and all three constructions are in Table 27. n/r, not reported, or for our weight-decay row not evaluated, because its endpoint weights were not saved. The published SOC row's certified bound is 1 by construction, and its witnessed sensitivity, hence its gap, is not reported.

Capped AdamW at looser budgets and on a dense network. Two registered arms scope the AdamW half of the inversion. At $c = 4$ and $c = 8$, the budgets at which capped SGD dies, capped AdamW trains (Table 10, top). Its stable rank is 4 at $c = 4$ and about 20 at $c = 8$, with no collapse toward one. That is consistent with the mechanism's prediction that AdamW's collapse step rises with the budget. Its verified accuracy at those budgets is zero, the vacuous loose-budget certificate. Take a dense network on the same input, with three hidden layers of width 128. Under the same shrink-only projector at $c = 2$ and the declared rate, capped AdamW trains with no rank collapse. It reaches the accuracy of capped Muon on the same network (Table 10, bottom). So the death is specific to convolutional stacks at the tight budget.

Table 10: Top, capped AdamW at loose budgets on the convolutional stack for 8,000 steps. Columns give clean accuracy, stable rank, gap, and verified accuracy at $36/255$. Bottom, capped AdamW and capped Muon on the dense network at $c = 2$, 1,000 steps: clean accuracy and collapse step. Both panels give the mean $\pm$ standard deviation over two seeds.

c	Clean	srank	G	VRA
4	0.545 ± 0.015	4.0 ± 0.1	149 ± 30	0.000 ± 0.000
8	0.819 ± 0.001	19.7 ± 0.9	$9{,}622 \pm 679$	0.000 ± 0.000

Optimizer	Clean	T_c
AdamW	0.473 ± 0.017	none
Muon	0.473 ± 0.007	none

The spectral sphere in full. The arm behind the Section 3.3 paragraph, three seeds, untuned at Muon's declared rate. At the budget matching our cap it ends at $G = 751$ and clean accuracy 0.831. At its faithful radius it ends at $G = 1.8$, clean accuracy 0.35, and verified accuracy 0.194 at $\varepsilon = 36/255$. That is the highest of any freely trained route here, and below the by-construction anchor's 0.217 (Table 9). Its reference implementation holds its sphere only to about $1.5\times$, with per-layer ratios from 1.07 to 1.49. Our per-layer measurements surfaced the discrepancy.

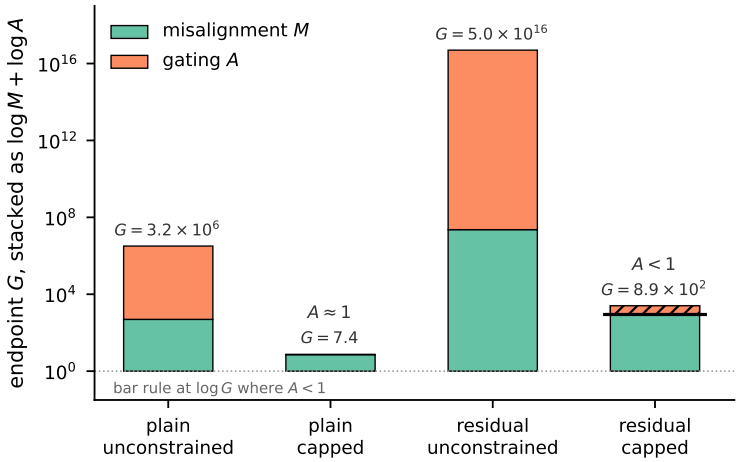

Figure 5: **Where the looseness lives.** The exact split $G = M \times A$, stacked in log space. The total height is $\log G$, and the segments are literally the two factors. Muon runs, seed 42, computed from the bias-free skeleton. Unconstrained, the gating factor A carries the gap in both families. Under the cap, A falls to about one. The composition factor M carries what remains, the relocation of Section 2.3. The capped residual bar's A segment draws downward, because a witnessed Jacobian can exceed the skeleton norm. That is legitimate, since A is not a certified ordering (Section 2.3).

Table 11: The exact split $G = M \times A$ by family, route and optimizer at the shared scale, from the bias-free skeleton. Values are mean $\pm$ standard deviation over the two replication seeds. The split is exact on each seed, so the product of the mean factors need not equal the mean gap. The penalty route and the residual family were decomposed under Muon only. † the arms that collapse under the cap (Section 3.3). Their gaps are those of a network that stopped learning.

Family	Route	Optimizer	M	A	G
Plain	Unconstrained	Muon	581 ± 125	$5{,}667 \pm 1{,}189$	$(3.22 \pm 0.02) \times 10^6$
Plain	Unconstrained	AdamW	133 ± 21	612 ± 31	$(8.19 \pm 1.68) \times 10^4$
Plain	Unconstrained	SGD	190 ± 29	23.3 ± 3.1	$4{,}381 \pm 87$
Plain	Penalty	Muon	51 ± 11	129 ± 16	$6{,}526 \pm 552$
Plain	Hard cap	Muon	7.20 ± 0.07	1.00 ± 0.02	7.20 ± 0.23
Plain	Hard cap	AdamW†	8.11 ± 3.38	0.80 ± 0.33	5.96 ± 0.01
Plain	Hard cap	SGD†	2.85 ± 0.82	1.55 ± 0.33	4.27 ± 0.32
Residual	Unconstrained	Muon	$(1.92 \pm 0.53) \times 10^7$	$(2.38 \pm 0.31) \times 10^9$	$(4.49 \pm 0.67) \times 10^{16}$
Residual	Hard cap	Muon	$2{,}248 \pm 390$	0.36 ± 0.02	811 ± 106

H THE SYMMETRY PROPOSITIONS, STATED AND PROVED

Proposition 2 (Cap profiles are a gauge choice)**.** *Let $\mathbf{c}, \mathbf{c}'$ satisfy $\prod_\ell c_\ell = \prod_\ell c'_\ell$, and set $\lambda_\ell = c'_\ell / c_\ell$. For any θ feasible for $\mathbf{c}$, the covariant rescaling*

$$\boldsymbol{W}_\ell \;\mapsto\; \lambda_\ell \boldsymbol{W}_\ell, \qquad \boldsymbol{b}_\ell \;\mapsto\; \Big(\prod\nolimits_{j \le \ell} \lambda_j\Big)\, \boldsymbol{b}_\ell \tag{6}$$

is feasible for $\mathbf{c}'$ and defines the same function exactly, biases included. Consequently L_{global}, L_{local}, G, every margin $m(\boldsymbol{x})$, and every certified radius $r_{L_{\mathrm{global}}}(\boldsymbol{x})$ are identical under $\mathbf{c}$ and $\mathbf{c}'$.

Proof of Proposition 2. Feasibility follows from $\|\lambda_\ell \boldsymbol{W}_\ell\|_2 = \lambda_\ell \|\boldsymbol{W}_\ell\|_2 \le c'_\ell$. For function equality, induct on layers. Suppose the input to layer ℓ is scaled by $\prod_{j<\ell} \lambda_j$. Then the rescaled affine map outputs $\prod_{j \le \ell} \lambda_j$ times the original pre-activation, biases included by equation 6. The 1-homogeneity of σ propagates the scalar. At the output the accumulated factor is $\prod_\ell \lambda_\ell = 1$. Invariance of L_{global} follows from the same product, and all remaining quantities depend only on the function and L_{global}. □

Proposition 3 (Post-hoc projection is inert)**.** *Let θ be any trained network, and let θ' be its covariant projection onto an arbitrary profile $\mathbf{c}$. The projection maps $\boldsymbol{W}_\ell \mapsto t_\ell \boldsymbol{W}_\ell$ and $\boldsymbol{b}_\ell \mapsto \big(\prod_{j \le \ell} t_j\big)\boldsymbol{b}_\ell$ with $t_\ell = \min(1, c_\ell / \|\boldsymbol{W}_\ell\|_2)$ and $t = \prod_\ell t_\ell$. Then $f_\theta[\theta'] = t \cdot f_\theta[\theta]$ exactly, biases included, and therefore predictions are unchanged, $G(\theta') = G(\theta)$, and $r_{L_{\mathrm{global}}}(\boldsymbol{x})$ is unchanged for every $\boldsymbol{x}$.*

Proof of Proposition 3. Run the induction of Proposition 2 with $\lambda_\ell = t_\ell$, whose product is now t rather than one. It gives $f_\theta[\theta'] = t\, f_\theta[\theta]$, with biases carried by the covariant rule. Both L_{global} and L_{local} scale by t, so G is fixed. The margin $m(\boldsymbol{x})$ scales by t and L_{global} scales by t, so $r_{L_{\mathrm{global}}}(\boldsymbol{x}) = m(\boldsymbol{x}) / (\sqrt{2} L_{\mathrm{global}})$ is fixed. □

I THE MECHANISM RUN

The mechanism of Section 3.3 was registered with three predictions and their failure conditions before the runs that test it. It came after the capped-SGD budget runs and before the AdamW ones. So SGD's half of the first sign was known, and AdamW's was not. Under projection the next weight is proportional to $\boldsymbol{W} + \Delta\boldsymbol{W}$. A step of stable rank near one and size comparable to the budget thus drags the weights onto its direction. Muon's step is orthogonalized, so its operator norm is set by the learning rate and its stable rank is full. SGD's step is the gradient, which in a capped network scales with the product of the other layers' norms. Its size relative to the budget therefore grows with the budget. AdamW's step is sign-like, with operator norm set by the learning rate and the layer's shape. Its size relative to the budget therefore shrinks as the budget loosens. The sign

pattern of a near-rank-one gradient is itself near rank one. The predictions were written with pass and fail conditions first. The collapse step T_c is the first step at which the flattened kernel's stable rank falls below 2. That stable rank is averaged over the seven weight layers. At initialization the per-layer values run from 7.8 to 145, each below its layer's smaller dimension. The update's stable rank is averaged over the six orthogonalized convolution layers. First, T_c decreases with the budget for SGD and increases for AdamW. Second, consider the per-step update in operator norm, the same instrument the projector uses. For SGD it exceeds half the budget on some layer in the 50 steps before collapse, at every budget. For Muon it never exceeds a tenth of the budget. Third, in that window the step's stable rank is below 2 for SGD and below 5 for AdamW. For Muon it is above 20. Every capped arm at every budget was re-run for 1,000 steps with per-step logging. Every collapse step is thus read at the same cadence.

Table 12: **Three registered predictions about the update mechanism.** Collapse step T_c is the first step at which mean stable rank falls below 2. It is read at per-step cadence on every arm. Two seeds per cell, collapsed to their means, with standard deviations in Table 13.

Registered sign	SGD	AdamW	Muon	Verdict
Step stable rank at T_c	1.0 to 1.5	1.3	88 to 134	confirmed
T_c at $c = 2/4/8$	290 / 5 / 1	258 / none / none	none / none / none	confirmed, both ways
Peak $\|\Delta \boldsymbol{W}\|_2/c$ before T_c, $c = 2/4/8$	0.18 / 5.5 / 71.6	never crosses	≤ 0.05	partial at $c = 2$
Step-0 largest logit, $c = 2/8$	0.12 / 1,329	0.12 / –	0.12 / –	diagnostic

Table 13: The mechanism run, per cell, as mean $\pm$ standard deviation over two seeds. Columns give the collapse step T_c and the largest per-step operator-norm update relative to the budget. That update is taken over the 50 steps before collapse. Muon never collapses, so it is measured over its first 50 steps and its orthogonalized group. The last column is the mean stable rank of the update in that window.

Optimizer	c	T_c	$\max\|\Delta \boldsymbol{W}\|_2/c$	srank$(\Delta \boldsymbol{W})$
SGD	2	290.5 ± 2.1	0.184 ± 0.034	1.45 ± 0.11
SGD	4	5.0 ± 1.4	5.5 ± 1.8	1.14 ± 0.12
SGD	8	1.0 ± 0.0	71.6 ± 2.8	1.05 ± 0.02
AdamW	2	258.0 ± 14.1	0.250 ± 0.012	1.30 ± 0.01
AdamW	4	beyond 1,000	–	–
AdamW	8	beyond 1,000	–	–
Muon	2	beyond 1,000	0.046 ± 0.000	111.9 ± 2.1
Muon	4	beyond 1,000	0.023 ± 0.000	130.0 ± 5.6
Muon	8	beyond 1,000	0.012 ± 0.000	89.7 ± 2.6

Two of the three were confirmed in full. The sign is confirmed both ways. SGD's collapse step falls from 292 and 289 at $c = 2$ to 6 and 4 at $c = 4$. At $c = 8$ it is a single step. AdamW's rises from 268 and 248 at $c = 2$ to beyond the window at both looser budgets. The rank is confirmed for all three, 1.0 to 1.5 for SGD, 1.3 for AdamW, 88 to 134 for Muon. The size is partial in the way the registration named as its risk. SGD's step crosses half the budget at $c = 4$ and $c = 8$, but reaches only 0.16 and 0.21 of it at $c = 2$. So the operating-budget collapse is an accumulation along a persistent direction. Muon's largest relative step is 0.046, and AdamW's never crosses. The step-0 diagnostic explains why the loose budgets die differently. At $c = 8$ the initial network, every layer scaled to its cap, has logits of about 1,300. There 99 percent of examples are saturated, under a certified bound of 65,536. At $c = 2$ the logits are 0.12, with no saturation. So $c = 8$ is a scale pathology that the update account and a loss-scale account predict alike. The budget at which a logit temperature can discriminate them is $c = 2$.

Annealing the budget. A registered test asks whether the operating budget or the opening phase does the killing. The budget is held at $c = 8$ to step 2,000, lowered linearly to $c = 2$ by step 4,000, and held there. Capped AdamW keeps training during the loose phase at stable rank 65. It falls below stable rank 2 at steps 5,418 and 5,599. It ends at stable rank 1.91, with clean accuracy 0.297 and 0.296. The registration asked for stable rank above 2 and clean accuracy at least 0.40, so that prediction is refuted. Annealed SGD is already collapsed at the first logged step and ends at chance, as registered. The operating budget kills whenever it arrives.

J THE TEMPERATURE-AND-RATE FACTORIAL

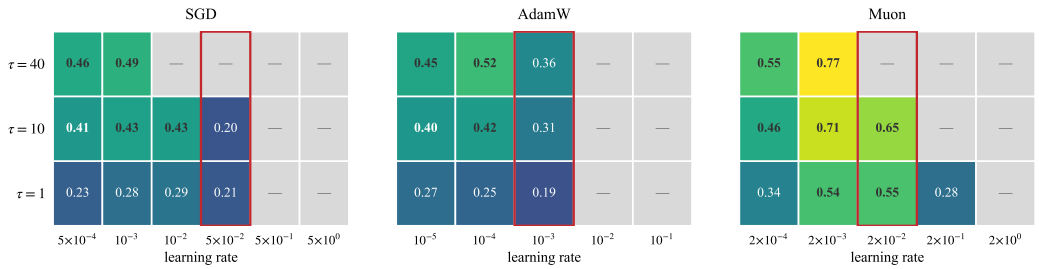

Figure 6: **The registered factorial, every cell.** Clean accuracy at step 1,000, the mean over two seeds, for each capped optimizer at $c = 2$. Rows are the logit temperature, columns the learning rates surrounding each declared rate. The declared rate is boxed in red, and it is where the paper reports the death. Bold values clear the registered rescue bar of 0.40. A dash marks a cell at chance. It is drawn grey, so a dead cell is not mistaken for a poor one. The bar is cleared only up and to the left, where a lowered rate meets a raised temperature. Neither axis clears it alone. Table 14 adds the collapse step for every cell.

The factorial behind Section 3.3 runs every capped optimizer at $c = 2$ across logit temperatures $\tau \in \{1, 10, 40\}$. The logits are multiplied by τ before the loss, and nothing else changes. A learning-rate grid surrounds each declared rate, with two seeds and 1,000 steps of per-step logging. It was registered with the two mechanisms' opposite predictions. If the loss's scale is the cause, a temperature rescues SGD. If the update's is, it brings the collapse forward. Rescue was defined as clean accuracy at step 1,000 within 15 points of capped Muon's untempered cell. That bar is 0.40, against dead controls of 0.21 for SGD and 0.19 for AdamW. Against the tempered Muon cell of 0.77 the bar would be 0.62, which no SGD or AdamW cell reaches. No cell rescues on one axis alone. The cells that clear the bar are at $\tau \geq 10$. They are also at a fifth of the declared rate or lower (Fig. 6). A registered extension of the same cells to 8,000 steps settles the horizon. Capped Muon's control gains almost nothing from eight times the steps, 0.554 against 0.548. So the matched bar is 0.404, and the two bars agree. Every rate-only cell finishes below it. SGD's two lower rates settle at 0.324 and 0.328. Its highest rate collapses after the shorter window, at steps 1,598 and 1,329. AdamW's does too, at 2,848 and 3,049. Both were censored at 1,000 steps. Rate alone buys delay, not survival. All five tempered SGD cells hold. The two at $\tau = 40$ rise further, reaching 0.620 to 0.641 clean accuracy, above the capped Muon control. A registered certificate pass on the $\tau = 40$ endpoints at rate 10^{-3} says what that accuracy certifies. The tempered pair verifies nothing at $\varepsilon \in \{8, 16, 36\}/255$. At the same rate, the untempered control verifies 48 and 71 images of 10,000 at the smallest radius. It verifies nothing at larger radii. Both counts fall below the bar the registration fixed in advance. So at this rate capped SGD certifies almost nothing, tempered or not. The zero cannot be charged to the temperature. What the temperature does is move the gap, from 5.4 and 5.5 to 53 and 54. That is about tenfold, more than the eightfold of the Muon cell, and entirely through misalignment. The certified bound stays fixed at the $c = 2$ ceiling on all four endpoints. At that fixed bound M rises from about 10 to 112 and 120. The gating does not freeze. Its entropy rises from 0.20 to 0.44, so the untempered network is the more nearly linear of the two. The reading of Section 3.7 thus replicates on a second optimizer. A temperature loosens the certificate through misalignment rather than linearization. The result is scoped, as registered, to $\tau = 40$, rate 10^{-3},

$c = 2$ and this architecture. The temperature is confounded with a fortyfold larger effective step, since the loss is taken on τ times the logits. AdamW's branch reads the same at 8,000 steps. Its rate-only cells settle below the bar or collapse after the shorter window. All three of its tempered cells hold, reaching 0.43 to 0.57. A replica of the setting of Khromov & Singh (2024) runs SGD without momentum at rate 10^{-3} and $\tau = 40$. It reaches 0.38, slightly under the bar. At the declared rate, a temperature of 10 delays SGD's collapse and 40 brings it forward. The temperature axis is thus non-monotone, and its registered sign is scored as mixed. Step-0 saturation is absent at $c = 2$ at every temperature, so the tempered training is not a saturation artifact.

Table 14: The factorial at step 1,000, as mean ± standard deviation over two seeds. Each temperature has two rows, clean accuracy above and the collapse step below, where – means no collapse within 1,000 steps. The declared rates are 0.05, 10^{-3}, and 0.02.

SGD, learning rate						
τ	0.0005	0.001	0.01	0.05	0.5	5
1	0.23 ± 0.01	0.28 ± 0.00	0.29 ± 0.01	0.21 ± 0.01	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	290.5 ± 2.1	26.5 ± 2.1	3.5 ± 0.7
10	0.41 ± 0.00	0.43 ± 0.01	0.43 ± 0.03	0.20 ± 0.02	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	583.5 ± 153.4	2.0 ± 0.0	2.0 ± 0.0
40	0.46 ± 0.01	0.49 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	3.0 ± 0.0	2.0 ± 0.0	1.0 ± 0.0

AdamW, learning rate					
τ	1e-05	0.0001	0.001	0.01	0.1
1	0.27 ± 0.00	0.25 ± 0.00	0.19 ± 0.01	0.10 ± 0.00	0.10 ± 0.00
	–	–	258.0 ± 14.1	–	3.0 ± 0.0
10	0.40 ± 0.01	0.42 ± 0.00	0.31 ± 0.01	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	–	2.0 ± 1.4
40	0.45 ± 0.00	0.52 ± 0.00	0.36 ± 0.07	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	–	1.0 ± 0.0

Muon, learning rate					
τ	0.0002	0.002	0.02	0.2	2
1	0.34 ± 0.01	0.54 ± 0.01	0.55 ± 0.02	0.28 ± 0.02	0.10 ± 0.00
	–	–	–	–	548.5 ± 4.9
10	0.46 ± 0.00	0.71 ± 0.00	0.65 ± 0.00	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	658.5 ± 20.5	603.0 ± 28.3
40	0.55 ± 0.01	0.77 ± 0.02	0.10 ± 0.00	0.10 ± 0.00	0.10 ± 0.00
	–	–	–	668.5 ± 36.1	643.0 ± 24.0

Capped Muon is the second reading. The registration named the branch in which a temperature lifts its accuracy. The cap's accuracy price would then be a loss-scaling artifact. It does, and the price moves rather than disappears. Table 16 gives the exact instruments on the endpoints. At $\tau = 40$ and a tenth of the declared rate, clean accuracy rises by more than twenty points. The certified bound stays unchanged at the cap's ceiling. The witnessed sensitivity falls eightfold, so G rises from 7 to 58. The linearity instrument of Section 3.7 and the certificate pass ran on the same endpoints and their controls. They say what the fall is (Table 15). The pattern entropy is unchanged, 0.49 against 0.48. The gating stays as input-dependent as before, and the network has not linearized. A falls from 0.96 to 0.38, and the loosening is entirely in M, from 7 to 152. The composed operator norm collapses from 0.58 to 0.026. The certified product stays fixed at the cap's ceiling. The tempered network verifies nothing at $\varepsilon \in \{8, 16, 36\}/255$ on either seed, against 0.26 at $8/255$ for the control. The temperature divides the raw-logit margins by 40.

Table 15: Tempered against control capped Muon at $c = 2$, step 1,000. Rows give the split, the gating entropy over the test set, and verified accuracy at three radii. Values are mean $\pm$ standard deviation over two seeds.

	Control	Tempered
Clean accuracy	0.548 ± 0.016	0.770 ± 0.022
M	7.5 ± 0.9	152.3 ± 0.9
A	0.96 ± 0.12	0.38 ± 0.01
Gating entropy	0.48 ± 0.01	0.49 ± 0.00
VRA at $8/255$	0.258 ± 0.016	0.000 ± 0.000
VRA at $16/255$	0.100 ± 0.010	0.000 ± 0.000
VRA at $36/255$	0.001 ± 0.001	0.000 ± 0.000

Table 16: Capped Muon at $c = 2$ under a logit temperature, with exact instruments on the step-1,000 endpoints. Values are mean $\pm$ standard deviation over two seeds. The cap sets L_{global}, which is identical on both seeds.

τ	rate	Clean	L_{global}	L_{local}	G
1	0.02	0.548 ± 0.016	3.9721 ± 0.0000	0.557 ± 0.002	7.13 ± 0.02
40	0.002	0.770 ± 0.022	3.9721 ± 0.0000	0.068 ± 0.001	58.38 ± 1.25

K THE OPTIMIZER SWAP

Section 3.4 reports the mid-run swap. Each unconstrained run trains 4,000 steps under one optimizer. The optimizer state is then reset, and training continues to 8,000 under the other. Self-swap controls reset the state and continue under the same optimizer. Table 17 gives the gap at the swap and at the end. It gives the movement between them in decades, factors of ten, and the endpoint clean accuracy. It uses two seeds per swapped direction and one per control.

Table 17: The mid-run optimizer swap at the shared scale. Gap at the swap step and at the end, and its movement in decades over the second half. The last column is endpoint clean accuracy. Each swap is the mean $\pm$ standard deviation over two seeds, and each control is a single seed. Both controls stay within their bands while every swapped run moves more than 1.6 decades.

Swap	G at 4,000	G at 8,000	decades	clean
SGD to Muon	$(3.64 \pm 0.16) \times 10^3$	$(1.17 \pm 0.13) \times 10^6$	$+2.50 \pm 0.07$	0.876 ± 0.004
Muon to SGD	$(8.33 \pm 0.14) \times 10^5$	$(1.71 \pm 0.05) \times 10^4$	-1.69 ± 0.02	0.906 ± 0.004
Muon to Muon (control)	8.23×10^5	3.03×10^6	$+0.57$	0.878
SGD to SGD (control)	3.52×10^3	4.18×10^3	$+0.07$	0.857

L GAUGE INVARIANCE CHECK

The covariant rescaling equation 6 moves a trained capped network across six cap profiles that share one total budget. The profiles are uniform, front-heavy, back-heavy, alternating, single-spike, and mild-tilt, on two seeds. Five of the six reproduce G, L_{global}, and L_{local} bit for bit, at $G = 7.3641356$ and 7.0346887 on the two seeds. The mild-tilt profile, the one whose rescaling factors are not powers of two, agrees to a relative 2×10^{-6}. An earlier check rescaled weights alone, which is exact only for the bias-free idealization. It agreed to three significant figures on the same biased networks. Certificates evaluated in floating point can degrade in ways real arithmetic does not predict (Murray, 2026). So we report the agreement at this precision rather than assert it from the proof.

M NEURON-WISE GAUGE OPTIMIZATION

The scalar rescaling of Propositions 2 and 3 is one function-preserving family. The neuron-wise diagonal gauge $\boldsymbol{W}_\ell \mapsto \boldsymbol{\Lambda}_\ell \boldsymbol{W}_\ell \boldsymbol{\Lambda}_{\ell-1}^{-1}$, with $\boldsymbol{\Lambda}_\ell$ positive and diagonal per channel, is the larger one. It changes $\|\boldsymbol{W}_\ell\|_2$ and so the certified bound (Kuang et al., 2026). We optimized it on the two unconstrained Muon corner endpoints at the shared scale. We ran 2,000 Adam steps on the log-scales, minimizing a power-iteration surrogate of the product of flattened-matrix norms. The surrogate never enters the reported number. The gauged copy is re-measured with the exact per-frequency instrument of Section 2.1. Function preservation is checked directly, as the largest logit deviation over the test set. The optimization is one-sided: a better search could tighten further, so the recovery is a floor (Table 18). A rerun at three times the optimization budget recovers 2.06 to 2.14, within four percent of the shorter run. So the factor of two is a plateau rather than an under-optimized search.

Table 18: Optimized neuron-wise gauge on the unconstrained Muon corner endpoints, CIFAR-10, width 64, depth 6, step 8,000. Values are mean $\pm$ standard deviation over the two endpoints. Rows give the certified bound and the gap over the same 64 witnesses, before and after. They also give the factor recovered and the largest logit drift the gauge introduced.

	before	after
Certified bound L_{global}	$(7.64 \pm 0.12) \times 10^8$	$(3.55 \pm 0.05) \times 10^8$
Gap G	$(3.22 \pm 0.02) \times 10^6$	$(1.49 \pm 0.05) \times 10^6$
Factor recovered		2.15 ± 0.06
Largest logit drift		$(6.1 \pm 2.2) \times 10^{-5}$

The gauge accounts for a factor of about two. The gap of six and a half orders is not a parameterization artifact. Six orders survive the largest function-preserving family of a generic ReLU network (Rolnick & Kording, 2020). Those six orders are what the paper calls a training property.

N A GATED-GAIN FLOOR FOR NARROW BUDGETS

The relocation hypothesis of Section 2.3 has a provable core, and the network's shape sets how far it reaches. Take a residual block with identity skip and branch Jacobian $\boldsymbol{B} = \boldsymbol{W}_2\boldsymbol{D}_2\boldsymbol{W}_1\boldsymbol{D}_1$, one gate per branch activation. Under caps $\|\boldsymbol{W}_i\|_2 \le c$, the block Jacobian $\boldsymbol{I} + \boldsymbol{B}$ is square, with largest singular value at most $1 + c^2$. By Weyl's inequality, its smallest singular value is at least $1 - \|\boldsymbol{B}\|_2 \ge 1 - c^2$. For any $\boldsymbol{X}$ and any square $\boldsymbol{Y}$ with smallest singular value s, $\|\boldsymbol{X}\boldsymbol{Y}\|_2 \ge \|\boldsymbol{X}\|_2\, s$. Indeed $\boldsymbol{Y}$ carries the unit ball onto an ellipsoid that contains the ball of radius s. Apply this from the head downward, across the run of identity-skip blocks that follows the last projection. Collect everything above that run into one affine map $\boldsymbol{X}$. In the implemented family, $\boldsymbol{X}$ is average pooling and a linear head, with no activation after the run. The cancellation below requires this, since an activation there would gate $\boldsymbol{X}$ in the Jacobian alone. At every input $\boldsymbol{h}$ to the run, this gives $\|\boldsymbol{X}\boldsymbol{J}_{\text{run}}(\boldsymbol{h})\|_2 \ge \|\boldsymbol{X}\|_2\,(1 - c^2)^{n_{\text{run}}}$. The same sub-network's bias-free skeleton has norm at most $\|\boldsymbol{X}\|_2\,(1 + c^2)^{n_{\text{run}}}$. The gating factor of the sub-network from the last projection block's output to the logits therefore obeys

$$A_{\text{run}} \;\le\; \left(\frac{1 + c^2}{1 - c^2}\right)^{n_{\text{run}}} \qquad \text{for } c < 1, \tag{7}$$

with no hypothesis on the head, whose norm cancels.

The stem and the three projection blocks lie outside this argument, and for a reason rather than a technicality. Each is a tall map. The stem takes three channels to the first width. Every channel change in the family is an increase at stride one. Their ranges are therefore proper subspaces of the spaces they map into. A lower bound on the norm of a product cannot pass through a tall factor. The layers above it may act as zero on its range. The caps constrain norms, not that alignment. The floor is therefore proven for the top run, not for the map we measure. That map's A includes the stem and every projection. In the implemented family, each group is a projection block followed by one identity block. The top run is thus a single block. The provable core of the relocation hypothesis is

one block deep. The floor is zero at $c = 1$ and negative beyond. The statement is thus vacuous at the $c = 2$ operating point and at every budget the paper's live networks use. It constrains only the collapsed $c = 0.5$ network, and there only a quantity we did not measure. What it does say about the operating point is that $A < 1$ there is consistent with the argument, not anomalous. At $c = 2$ the branch can cancel the skip entirely, and the skeleton does not floor the witnessed Jacobian.

O WARM-START PROJECTION CHECK

Table 19 gives the numbers behind this appendix.

The projector of Algorithm 1 leaves biases unchanged, so it is not the covariant projection of Proposition 3. This check is its empirical cover (Section 3.4).

Projection alone does not repair a trained biased network. Projecting the unconstrained pretrain onto the caps moves its gap from 1.12×10^6 to 1.86×10^5. That is a factor of six, against the five orders of magnitude separating it from the capped regime. The same projection applied to a capped pretrain is a no-op, 34.44 before and after. That locates the six-fold in the bias-dependent gating term rather than in the projection itself. It is also why the weight-only projector of Algorithm 1 falls short of the covariant one Proposition 3 describes. Repair comes from training instead. One hundred steps of capped continuation from the projected handoff bring the gap to 8.4. That is indistinguishable from a capped-from-scratch reference, and the two arms track each other thereafter. This check runs on a five-class disjoint split with 4,096 samples rather than on CIFAR-10 proper. Its accuracies are therefore not comparable with those reported elsewhere.

Table 19: Gap G across the two stages: the instantaneous projection at handoff, then capped continuation training. The capped pretrain is the control, on which the projection is a no-op. One run per arm.

Arm	Before projection	After projection	Capped continuation 100	500	2000
Unconstrained pretrain	1.12×10^6	1.86×10^5	8.4	8.0	8.6
Capped pretrain (control)	34.44	34.44	8.0	8.1	8.8

P REGISTERED ABLATION, SINGULAR-VALUE CLIPPING

Clipping individual singular values leaves the scalar-rescaling family of Proposition 3 and changes the function. The ablation is registered as pure post-processing on the saved endpoint weights of the six grid corners. We report it for the unconstrained Muon corner on two seeds, the case the prediction concerns. Clip each layer's singular values to the cap profile, then record clean accuracy and G before and after. Scalar rescaling preserves the function and the gap exactly. The registered prediction was that the clip, unlike it, buys a smaller certified bound by changing what the network computes. That change comes at an accuracy cost. The outcome is the strongest form of that prediction. Clipping the unconstrained endpoint weights to the cap profile takes the certified bound from 7.7×10^8 to 38.5. It takes the gap from 3.2×10^6 to 5.5×10^3. It also takes clean accuracy from 0.878 to 0.100, which on ten classes is chance. The second seed agrees, 0.873 to 0.100 with the gap at 3.9×10^3. Non-scalar surgery does buy a smaller bound, and it buys it by destroying the classifier. It still leaves the gap three orders of magnitude above one. Tightness survives only when it is trained.

Q STRESS TESTS, SCALE, AND THE DIAL IN FULL

This appendix carries the figures and tables behind Section 3.5 and the registered depth-10 verdict discussed in Section 3.1.

Capped SGD at looser budgets. This test was registered before the runs as four exhaustive branches. Capped SGD at $c = 4$ and $c = 8$, two seeds each, dies to chance accuracy 0.100 at

every endpoint. Stable rank falls from 65 to between 1.05 and 1.27 by step 250, the same signature window as $c = 2$. The witnessed Jacobian is zero from that step. Weight decay then shrinks the dead network, to L_{global} of 4.8×10^{-4} at $c = 4$ and 0.065 to 0.070 at $c = 8$. So no endpoint is at its cap, and the $(c/2)^7$ arithmetic never applies. Capped Muon from the same initialization at the cap trains to 0.809 and 0.843 at these budgets (Table 22). At $c = 2$ capped SGD stalled at 0.24 to 0.27. Four and eight times looser, it collapses to chance. The step-0 diagnostic of Appendix I says why the two deaths differ. One curiosity is reported rather than scored. At $c = 4$ the dead constant network certifies chance, with verified accuracy 0.100 at every radius. A bound of 5×10^{-4} makes any margin certifiable. That is certified accuracy without a function, and it is the same warning the $c = 0.5$ row carries.

Horizon. Table 20 extends the unconstrained shared-scale arms to 40,000 steps on three seeds. SGD's gap is flat past 8,000, while AdamW's and Muon's keep rising. So the grading widens with the horizon, and every mildness claim carries its horizon.

Table 20: Unconstrained gap along a 40,000-step horizon, geometric means over three seeds, and the factor gained after step 8,000.

Optimizer	8,000	20,000	40,000	factor after 8,000
SGD	4.7×10^3	5.0×10^3	5.5×10^3	1.2
AdamW	8.5×10^4	6.3×10^5	2.3×10^6	27.6
Muon	3.6×10^6	2.3×10^7	6.6×10^7	18.3

The gap under a tighter bound. Table 21 tracks the gap on a three-layer MLP of width 128 on flattened CIFAR-10. It uses both the product bound and the LipSDP bound of Fazlyab et al. (2019). The LipSDP bound's ratio to the product bound runs from 0.36 at initialization to 0.02 at the unconstrained Muon endpoint. The grading and the cap's effect are the same under both.

Table 21: Gap under the product bound and under LipSDP on the MLP slice, at initialization and at the last logged step. Values are mean $\pm$ standard deviation over two seeds. Capped SGD's second seed was logged only to step 1,000.

Arm	G_{prod} at 0	at end	G_{sdp} at 0	at end	last step
Unconstrained Muon	8.3 ± 0.1	$5{,}067 \pm 484$	2.9 ± 0.1	122.6 ± 16.7	8,000
Unconstrained SGD	8.3 ± 0.1	41.1 ± 1.6	2.9 ± 0.1	12.6 ± 0.9	8,000
Capped Muon	8.4 ± 0.1	9.3 ± 0.7	3.0 ± 0.0	3.6 ± 0.1	8,000
Capped SGD	8.4 ± 0.1	11.2 ± 1.6	3.0 ± 0.0	5.5 ± 0.6	8,000 and 1,000

Weight decay. Unconstrained Muon's endpoint gap moves through 4.6×10^6, 3.2×10^6, and 2.7×10^3 at decoupled weight decay 0, 5×10^{-4}, and 0.1. These are geometric means over two seeds, with clean accuracy flat at about 0.88. Strong decay is thus the one unconstrained setting that tames it, and Table 9 carries it as a route. Its endpoint weights were not saved, so no certificate was computed and its verified accuracy is not reported.

The cap as a dial. The budget sweep behind Section 3.7 is in Fig. 8 and Table 22. The certified bound is the product of the seven weight-layer norms times $1/32$ from the four average pools. Three 2×2 pools have norm $1/2$, and the global 4×4 pool has $1/4$. A layer the projector has touched is at $0.999\,c$, its margin. So wherever every layer binds, the bound is $(0.999\,c)^7/32$. That is 3.97 at $c = 2$ and 508 at $c = 4$, matching the measured values to within 0.1%. The step-0 network at $c = 8$, scaled to its cap exactly, prints $8^7/32 = 65{,}536$. That is the product structure made visible. Over the surviving budgets, gap and clean accuracy rise together. The gap grows by a factor of about 2,500, and clean accuracy from 0.185 to 0.843. Table 22 carries the verified column that Section 3.7 reads.

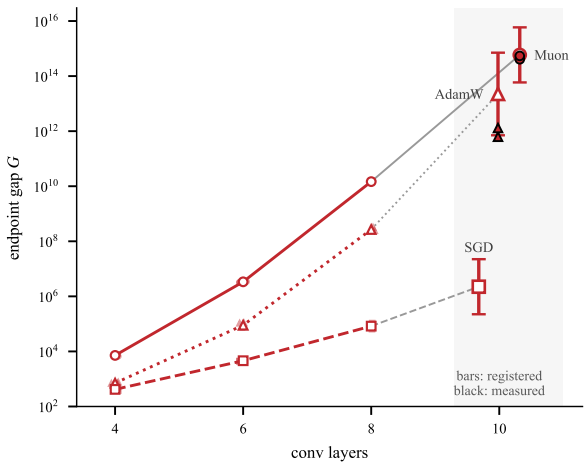

Figure 7: **The registered depth-10 prediction, and the verdict.** Measured endpoint gap G for unconstrained runs at depths 4, 6, and 8. Open marks are seeds, and lines join geometric means. The pre-registered prediction bands at depth 10 were drawn before any depth-10 run existed. The registration (2026-08-23) states $10^{14.77\pm1}$ for Muon, $10^{6.35\pm1}$ for SGD, and $10^{13.35\pm1.5}$ for AdamW. Its primary model is quadratic in depth, with the log-linear alternative registered as a model comparison. The measured points, filled marks, fall at $10^{14.61}$ and $10^{14.73}$ for Muon, 0.1 decades from the primary center. For AdamW they fall at $10^{11.8}$ and $10^{12.12}$, the first seed slightly below the band. Both lean toward the log-linear alternative the registration named as the competing model. We record this as a partial miss. SGD diverged at the declared learning rate and is reported as diverged rather than drawn. This panel draws the registered model and bands as written and fits nothing. Both were committed before the depth-10 runs. Red bars are the registered bands, each with an open mark at its center. Black-edged filled marks are the measured depth-10 seeds.

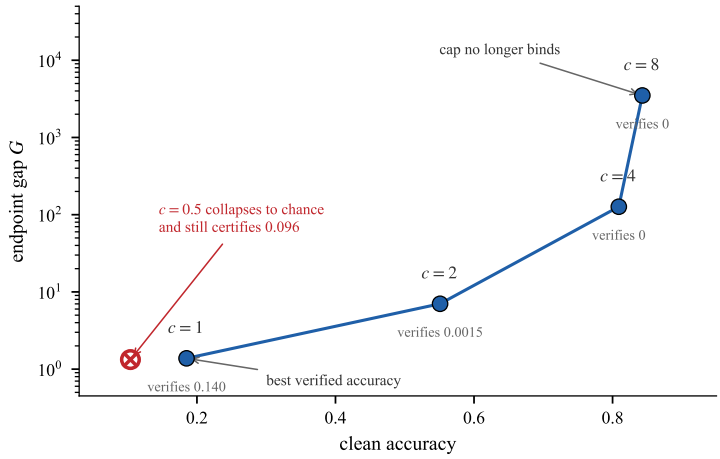

Figure 8: **The cap as a dial.** Clean accuracy against endpoint gap G, traced by the per-layer budget c, for capped Muon at the shared scale. Points are geometric means over two seeds. Verified accuracy at $\varepsilon = 36/255$ on the full test set is printed beside each point. The frontier is monotone over the surviving budgets, so tightness is bought with accuracy and no setting dominates another. Verified accuracy is not monotone along it. It peaks at $c = 1$, away from the operating point used elsewhere in this paper. The $c = 0.5$ run is drawn off the curve because it collapsed to chance. It still certifies 0.096, more than the operating point does. A verified-accuracy number means nothing without the clean accuracy beside it.

Table 22: **The cap as a dial, exact values.** Endpoint gap, clean accuracy, and verified accuracy at $\varepsilon = 36/255$ against the per-layer budget c. Rows are capped Muon at the shared scale, as geometric means over two seeds. The $c = 2$ row is the four-seed grid mean. The two corner seeds of Table 11 read 7.36 and 7.03, reported there as 7.20 ± 0.23. The $c = 0.5$ row is at chance accuracy and its numbers are reported, not interpreted. At $c = 8$ the cap no longer binds.

Budget c	G	Clean	Verified	Note
0.5	1.33	0.104	0.096	collapsed to chance
1	1.38	0.185	**0.140**	best verified accuracy
2	7.02	0.551	0.0015	used elsewhere in this paper
4	126.8	0.809	0	
8	3,499	0.843	0	cap not binding

Gating across the budget. The linearity check of Section 3.6 measures gating as the per-unit entropy of each unit's activation pattern. It is zero for a unit that never switches and near $\ln 2$ for one that switches with the input. It is recorded on the sweep endpoints over the full CIFAR-10 test set. Table 23 reports the per-layer mean of the pattern entropy and the fraction of units active. It also gives the fraction silent on every input and A as measured in the sweep, over two seeds. The 64-witness set gives the same entropy and active fraction to within 0.01, but it overstates silent units. A unit silent on 64 inputs need not be silent on ten thousand. On the unconstrained network, the small set calls 14 percent of units dead. The full test set finds 1 percent.

Table 23: Gating statistics over the test set across the cap budget, for capped Muon at the shared scale. The unconstrained arm is the reference, and values are mean $\pm$ standard deviation over two seeds. Entropy near zero means every unit is always on or always off, and near $\ln 2 \approx 0.69$ it means fully input-dependent gating.

Budget c	entropy	active	dead	A	reading
0.5	0.06 ± 0.00	0.56 ± 0.01	0.42 ± 0.01	0.94 ± 0.00	linear, collapsed
1	0.09 ± 0.00	0.75 ± 0.01	0.20 ± 0.01	0.93 ± 0.00	linear
2	0.42 ± 0.01	0.70 ± 0.00	0.01 ± 0.00	1.00 ± 0.02	gated, tight
4	0.50 ± 0.01	0.51 ± 0.03	0.05 ± 0.02	0.98 ± 0.06	gated, tight
8	0.26 ± 0.02	0.15 ± 0.00	0.43 ± 0.01	4.6 ± 1.1	pruned
unconstrained	0.37 ± 0.00	0.17 ± 0.00	0.01 ± 0.00	$5{,}667 \pm 1{,}189$	gated, loose

ImageNet-64 in full. The arms behind the Section 3.5 paragraph. Endpoint gaps at 20,000 steps are 5.0×10^8 for SGD, 5.5×10^{12} for AdamW, and 7.4×10^{18} for Muon. These are geometric means over two seeds, with the pairs four and six orders apart. The registered ratio between SGD and Muon was two to four orders. Both capped arms fail to train. Capped Muon reaches accuracy 0.03 and capped SGD 0.008, against 0.23 unconstrained and 0.001 chance. Their gaps are 190 and 687 against the registered ceiling of 10^2, misses by factors of 1.9 and 6.9. Capped SGD collapses to stable rank 2, capped Muon holds stable rank near 61. The spectral sphere reaches clean accuracy 0.057 and 0.072 at a gap of 1.9×10^6. That gap is more than twelve orders below unconstrained Muon, and below unconstrained SGD. The registered bar was accuracy within eight points of the unconstrained arm, and it misses by eight to nine points.

The cap at a thousand classes, with a temperature. A second registration asked whether the lever that buys accuracy at CIFAR scale rescues the cap here. Stage A ran five cells of capped Muon over temperature and budget, one seed each. A registered rule selected the best live cell, and Stage B replicated it at a second seed. The cells at $\tau = 40$ run at a tenth of the declared rate, as at CIFAR scale. The cells at $\tau = 1$ and 10 run at the declared rate. So this grid does not separate temperature from rate. The bar for a live cell was clean accuracy 0.12 at step 20,000. Three of five cells cleared it (Table 24). The selected cell, $\tau = 40$ at budget 4, reaches 0.386 and 0.380. That is above every unconstrained arm. The best of them is SGD at 0.317, and Muon reaches 0.227 and 0.237. Its gap

is 12.7 and 12.6 orders below Muon's, against a registered requirement of ten. At budget 2 the same temperature reaches 0.225 on its one seed. That matches Muon's unconstrained accuracy under a certified bound of 1.7×10^3 against 8.0×10^{21}, eighteen orders apart. The untempered control shares that bound to five figures and reaches 0.03. The two live cells at $\tau = 40$ are two ends of one trade. The looser budget buys sixteen points of accuracy at a bound five thousand times larger.

Three things are reported and not interpreted. No cell verifies any image at any radius, under either certificate convention. The exact split against the untempered control is not scored, by registration. The control is dead, and a live network is not compared with a dead one. We print the measured factors all the same. At budget 4 and the declared rate, $\tau = 1$ trains to 0.134 and $\tau = 10$ dies at 0.036. That is a real basin or a single-seed accident, and one seed per cell cannot say which. The best cell is at $\tau = 40$ and the lower rate, so it does not extend that comparison. Unconstrained SGD is both the tightest and the most accurate unconstrained arm here, 0.317 at a gap of 5.0×10^8. Unconstrained accuracy is thus passed at this scale by the cap with a temperature and a lower rate. A certificate that verifies anything there remains open.

Table 24: Capped Muon on ImageNet-64 at 20,000 steps, by logit temperature and budget. Two-seed rows give mean $\pm$ standard deviation. The others are one seed. M and A are the exact split, reported and not scored. Verified accuracy is zero in every row at every radius. A cell is live at clean accuracy 0.12 or above. The cells at $\tau = 40$ run at a tenth of the declared rate. Unconstrained rows are geometric means of L_{global} and G over two seeds.

τ	c	Seeds	Clean	L_{global}	G	M / A	
1	2	2	0.031 ± 0.003	1.67×10^3	191 ± 2	223 / 0.86	dead, the control
10	2	1	0.033	1.67×10^3	852	1.7×10^3 / 0.50	dead
40	2	1	0.225	1.67×10^3	1.4×10^3	8.2×10^3 / 0.17	live
1	4	1	0.134	9.70×10^6	7.8×10^4	3.3×10^4 / 2.34	live
10	4	1	0.036	4.84×10^6	1.8×10^6	1.3×10^6 / 1.39	dead
40	4	2	0.383 ± 0.004	9.06×10^6	1.65×10^6	3.3×10^6 / 0.54	live, selected
Unconstrained SGD			0.317 ± 0.001	1.3×10^{11}	5.0×10^8		
Unconstrained AdamW			0.195 ± 0.001	1.9×10^{15}	5.5×10^{12}		
Unconstrained Muon			0.232 ± 0.007	8.0×10^{21}	7.4×10^{18}		

R THE PREDICTION LEDGER

This appendix documents every prediction the text relies on, with its measured outcome, in the order it was made. The four rows of Table 25 are drawn from here. The release ledger holds 77 rows. Table 26 prints forty-four predictions, some of which fold two ledger rows into one. The rest are sub-predictions the text does not use. Five of those deserve naming as misses. The first is the swap endpoint bands, which were too tight for the horizon. The second is the cap sweep's accuracy bands, which the dial exceeded in both directions. The third is the self-swap control's movement clause, a calibration error against Muon's own drift. The fourth is the sphere's accuracy partition at its faithful radius, refuted in the favorable direction. The fifth is the capped arms' gap ceiling at a thousand classes, missed by factors of 1.9 and 6.9 (Appendix Q). The rest are confirmed sub-clauses of predictions printed here, among them the witness-scarcity bound and the certified-radius check. The others are arms recorded as not scored, having collapsed below their liveness floors.

Table 25: **Four registered predictions and their outcomes.** Each row was committed to the code history before the run that tests it. The first was committed after SGD and Muon had run and before AdamW had. The release ledger holds 77 rows. Table 26 prints forty-four rows, twenty-two confirmed in full. Appendix R names five further misses among the rows it does not print.

Registered claim	Registered	Measured	Verdict
Unconstrained gap growth orders SGD $<$ AdamW $<$ Muon	ordering only	$3\times$, $60\times$, $2{,}300\times$ geometric means over seeds	CONFIRMED
Endpoint gap at depth 10, outside the trained grid	Muon $10^{14.77\pm1}$ AdamW $10^{13.35\pm1.5}$	Muon $10^{14.61}$, $10^{14.73}$ AdamW $10^{11.80}$, $10^{12.12}$	MUON CONFIRMED ADAMW MARGINAL
Switching Muon to SGD mid-run deflates the gap by at most 0.2 decades	≤ 0.2 decades	1.67 and 1.70 decades	REFUTED
Muon-to-SGD gap ratio at ImageNet-64	2 to 4 orders	10 orders	MISSED BY SIX

Table 26: **The registered predictions the text relies on.** Direction and magnitude were written down before measurement in every row. Every row was committed to print before its outcome was known, including the two mechanistic hypotheses the data refuted. The rows state two exceptions. The Mixer entry is not witnessed by the code history. SGD's half of the mechanism sign was known when the row was registered. Distances are quoted in decades, factors of ten.

Prediction	Registered	Measured	Verdict
Unconstrained inflation orders SGD $<$ AdamW $<$ Muon	before any AdamW run existed	$3\times < 60\times < 2{,}300\times$, geometric means over seeds, two datasets	confirmed
Skips relocate looseness from gating into composition under caps	before the residual runs	capped residual M 1,972–2,523 against A 0.35–0.37	confirmed
Covariant projection cannot de-inflate, and the weight-only projector does not repair	before the warm-start runs	the function is unchanged by proof, and the weight-only projector moves G by a factor of six against five orders (Appendix O)	confirmed
The growth persists under the tighter LipSDP bound	before the solves	unconstrained G_{sdp} 2.9 to 110.8 and 3.0 to 134.4, capped flat under both bounds	confirmed
Mixer ordering Muon $>$ AdamW $>$ SGD on the all-linear column	ledger entry only, not witnessed by the code history	at the declared rate Muon diverges, reported, and at its grid-favored rate Muon 4.6×10^6 is below AdamW 6.3×10^6, both above SGD 3×10^5	not confirmed, the ordering is scoped to the convolutional families
Depth-10 gap $10^{14.77\pm1}$ Muon, $10^{6.35\pm1}$ SGD, $10^{13.35\pm1.5}$ AdamW	nine hours before any depth-10 run, witnessed by the code history	Muon $10^{14.61}$/$10^{14.73}$, 0.1 decades from the primary center, AdamW $10^{11.80}$/$10^{12.12}$, the first seed slightly below the band and both leaning toward the registered log-linear alternative, SGD diverged at the declared learning rate (Fig. 7)	split, Muon confirmed and AdamW marginal
Caps compose with a trained margin	before the composition runs	verified accuracy 0.0015 to 0.052/0.056, G unchanged	confirmed
The gap is fixed by the region training entered early	before the swap runs	every swapped run moves more than 1.6 decades after the swap	refuted
Swapped runs re-aim halfway to the destination within 1,000 steps	before the swap runs	about half upward, about a fifth downward	not supported, re-aiming is rate-limited and asymmetric

Table 26, continued

Prediction	Registered	Measured	Verdict
The collapse generalizes to an optimizer outside the grid	before the spectral-sphere runs	$G = 751$ at the matched budget, 3.7 orders below unconstrained Muon	confirmed
That optimizer's reference implementation holds its sphere to within 12%	before the runs	per-layer ratios span 1.07 to 1.49	refuted
A milder optimizer does not deflate an inflated gap	before the swap runs	switching to SGD moved it down 1.67 and 1.70 decades against a registered ceiling of 0.2	refuted
Bound arithmetic $L_{\text{global}} = (c/2)^7 \times 3.97$ where caps bind	before the sweep	within 0.1% at every binding budget	confirmed
Gap monotone in the budget, no interior optimum on the frontier	before the sweep	$1.38 \to 7.02 \to 126.8 \to 3{,}499$, no violating pair	confirmed
The grading holds at a thousand classes	before the ImageNet-64 runs	5.0×10^8, 5.5×10^{12}, 7.4×10^{18}, no seed overlap	confirmed
The Muon-to-SGD gap ratio there lies within two to four orders	before those runs	ten orders, the effect amplifies with scale	missed by six orders
Capped arms train at that scale, within 15 points of unconstrained	before those runs	the Muon half misses, 0.031 against 0.232, twenty points down against a bar of fifteen; the SGD half holds, stable rank 2 at 0.008; both arms dead	refuted
The spectral sphere trains at that scale, within 8 points of unconstrained	before those runs	0.057 and 0.072 against 0.23, both seeds alive, gap 1.9×10^6	alive but short, the bar at 0.15 missed by 8 to 9 points
A looser budget rescues capped SGD, scored on four exhaustive branches	before the $c \in \{4, 8\}$ runs	dies to chance at both budgets on both seeds, stable rank 1.05 to 1.27 by step 250, the budget-insufficient branch	budget eliminated as the cause
Temperature or rate alone rescues capped SGD and AdamW at $c = 2$, five exhaustive branches	before the factorial	neither alone; both together, SGD 0.49 and AdamW 0.52 against a bar of 0.40	mixed, the inversion scoped to the declared rate and temperature
Rate alone rescues capped SGD or AdamW once the horizon reaches 8,000 steps, seven exhaustive branches	before the extension	every rate-only cell settles or collapses below the matched bar of 0.404, the fastest collapsing after the shorter window	mixed confirmed, the horizon caveat discharged
The tempered capped-SGD rescue certifies something, scored against a 100-image bar	before the certificate pass	neither the tempered cell nor its untempered control verifies above the bar, while the gap rises about tenfold through misalignment	control also vacuous, the zero is not the temperature's
Capped Muon's accuracy price stands under a logit temperature	same registration	0.54 to 0.77 at $\tau = 40$ and a tenth of the rate, L_{global} unchanged at 3.97, G 7 to 58, all of it in M with the gating entropy unchanged, and verified accuracy zero	the price moves into misalignment and the recovered accuracy is uncertified
Capped AdamW on a dense network at $c = 2$ dies as on the convolutional stack	before the dense runs, three branches	trains, 0.46 and 0.48 with no rank collapse, level with capped Muon	the death is convolution-specific
Capped AdamW at $c \in \{4, 8\}$ dies as SGD does	addendum before the runs	trains, 0.55 and 0.82 at stable rank 4 and 20	the death is confined to the operating budget

Table 26, continued

Prediction	Registered	Measured	Verdict
Mechanism, sign: SGD's collapse step falls with the budget and AdamW's rises	after the capped-SGD budget runs and before the AdamW ones, so SGD's half was known	SGD 292/289, 6/4, 1/1 at $c = 2, 4, 8$; AdamW 268/248 at $c = 2$ and beyond the window at 4 and 8	confirmed
Mechanism, size: SGD's per-step update in operator norm crosses half the budget before collapse at every budget, Muon's never exceeds 0.1	same registration	SGD 0.16/0.21 at $c = 2$, 4.2/6.8 at 4, 74/70 at 8; Muon at most 0.046; AdamW never crosses	partial, confirmed at $c \geq 4$, the risk the registration named at $c = 2$
Mechanism, rank: update stable rank below 2 for SGD, below 5 for AdamW, above 20 for Muon	same registration	1.0 to 1.5, 1.3, and 88 to 134	confirmed
Annealing the budget from 8 to 2 rescues capped AdamW	before the anneal runs	stable rank 1.91 and clean accuracy 0.30, against the registered floors of 2 and 0.40	refuted
Annealed SGD collapses under every schedule	same registration	stable rank 1.05 and 1.35, clean accuracy at chance on both seeds	confirmed
GloRo's recipe reproduces in our code, within two points of its published 0.584	before the reproduction runs	kept-epoch verified accuracy 0.5875 over three seeds, range 0.5866 to 0.5881	confirmed
Muon lifts verified accuracy on that recipe	both directions written before the runs	verified accuracy falls to 0.414 against 0.586, and the gap rises from 2.79 to 10.4	refuted
The cap at least halves that Muon arm's gap, at no cost in verified accuracy	same registration	gap 2.35 against 10.4, a ratio of 0.23; verified accuracy 0.537 against 0.414, still 0.049 below the published recipe; the unscored check under Adam reads a ratio of 0.56 and does not halve	confirmed
Muon verifies more than SGD on each retrained construction, three-seed ranges disjoint	before the construction runs	SOC 0.230 to 0.242 against 0.188 to 0.209, Björck 0.229 to 0.255 against 0.183 to 0.204	confirmed
G lies in $[1, 1.15]$ on every by-construction endpoint	same registration	1.019 to 1.064 over the twenty-four endpoints of the three constructions	confirmed
Clean accuracy falls short of the native recipes, and the shortfall tracks the protocol	same registration	the three constructions spread 0.035 under Muon and 0.030 under SGD, against shortfalls of 0.46 and 0.51	confirmed
Muon verifies more than SGD on SLL, whose guarantee is algebraic, three-seed ranges disjoint	before the SLL runs	0.238 to 0.260 against 0.185 to 0.204, each optimizer at its declared rate	confirmed
G lies in $[1, 1.15]$ on every SLL endpoint	same registration	1.025 to 1.041 over six endpoints	confirmed
The ordering holds on all four constructions, ranges disjoint on each	same registration	holds on all four, each at the permutation floor of $p = 0.05$, with no joint value claimed	confirmed
The cap with a temperature trains at a thousand classes, clean accuracy at least 0.12 on both seeds	before the scale runs	0.386 and 0.380 at $\tau = 40$ and budget 4, above the unconstrained 0.227 and 0.237	confirmed

Table 26, continued

Prediction	Registered	Measured	Verdict
Its gap is at least ten orders below unconstrained Muon's	same registration	12.7 and 12.6 orders below, on the two seeds	confirmed
The exact split of the rescued cell against the untempered control	same registration, not scored by design	the control is dead; measured M 2.4×10^6 and 4.3×10^6, A 0.70 and 0.39	reported
Mechanism, isotropic deposit: unconstrained Muon raises stable rank	before the stable-rank runs	stable rank falls, about 65 to 25	refuted
Mechanism, spike growth: the rank fall tracks the gap growth	before the stable-rank runs	SGD falls furthest, to about 11, while its gap grows about $700\times$ less than Muon's	refuted

S PUBLISHED CONSTRUCTIONS UNDER ONE PROTOCOL

Four published 1-Lipschitz constructions are retrained here under the shared protocol. Three orthogonalize their weights. They are the Cayley anchor (Trockman & Kolter, 2021), skew-orthogonal convolutions (Singla & Feizi, 2021), and GroupSort with Björck orthonormalization (Anil et al., 2019). The fourth, SDP-based Lipschitz layers or SLL (Araujo et al., 2023), does not. Its residual layer is 1-Lipschitz for any weight, by an algebraic rescaling. Each fixes its certified bound by design, up to the unit cap on stem and head. The runs are CIFAR-10 at width 64, six construction convolutions, 8,000 steps, three seeds per optimizer. The learning rate is chosen per optimizer by a 1,000-step screen on a held-out split of the training set. Every endpoint certifies at the same bound to four figures, 0.03119. The four constructions agree to within 0.003 percent. That agreement is measured, and only the match between optimizers within a construction was registered. It is what lets the thesis be tested with the bound held fixed.

One route needs reconciling with its own literature. Our Cayley anchor reaches $G = 1.02$ at clean accuracy 0.27 to 0.29 under the shared protocol. Native recipes report about 0.75 (Trockman & Kolter, 2021). The registered diagnostics locate the delta. Tripling the budget to 24,000 steps moves the anchor to 0.29–0.31 across two seeds at G 1.015–1.016. The native learning rate moves it to 0.29–0.30 across two seeds at G 1.015–1.019. Removing the unit cap on the dense head drives both seeds to chance accuracy at G of 2.9 and 4.3. The gap near one is construction-invariant under every lever we measured except the head cap. The accuracy delta is architectural and protocol-borne. One lever we did not measure, they report themselves. In their Table 3, Trockman & Kolter (2021) find that input normalization alone raises the bound to about 4.1 in the native recipe. The anchor is the thesis operating on an architecture we did not design, and its cost is the shared protocol, not the construction.

Muon reaches higher clean accuracy and verifies more of the fixed bound than SGD, on all four constructions (Table 27). The three-seed ranges are disjoint under both certificate conventions. The exact one-sided permutation test gives $p = 0.05$ on each construction. That is its floor at three seeds a side. These are four results at that floor, and we claim no joint value. The four share seeds, data order and protocol, so they are not independent draws. Across the thirty endpoints G runs 1.019 to 1.064. The witnessed Jacobian reaches 94 to 98 percent of the certified bound. The meter reads near one where the bound is tight by design. Projected gradient descent at nine tenths of the certified radius flips no certified image, under either convention. It attacks the first thousand test inputs, as in Appendix C. In the orthogonal construction layers Muon holds a mean stable rank of 26 to 35, against SGD's 2 to 10. On SLL the two are closer, 8.5 against 5.9. That is the rank signature of Section 3.3, on architectures the cap never touches. With the bound held fixed, training decides how much of it is usable.

SLL is the construction that separates the effect from orthogonalization. Its layers never orthogonalize, and Muon still verifies more than SGD, 0.250 against 0.196 per pair. So what training moves here is not a property of orthogonal layers. Its guarantee rests on a diagonal rescaling $\boldsymbol{T}$ with $\|\boldsymbol{W}\boldsymbol{T}^{-1}\boldsymbol{W}^\top\|_2 \leq 1$, by Gershgorin's theorem. That norm is the share of its algebraic budget a layer uses. Under Muon it reads 0.976 to 0.988 on every layer. Under SGD it reads 0.687 to 0.993.

The per-layer Lipschitz bound reads exactly 1.0 under both. So the bound cannot see the difference between the two, and this margin can. It was not registered and scores nothing.

A bound fixed by construction is only as sound as its implementation. So each construction layer is checked at run time on its forward pass, not on its weights. The check takes the layer's Jacobian at the operating point and reads its top singular value. Over sixty checkpoints on SOC and Björck, under both optimizers, the worst reading is 1.0000001 and none fails. On SLL the gate ran inside training, on thirty checkpoints across the six arms. Every one passes, at a worst reading of 0.999999. The Cayley endpoints are the exception, since their weights were not retained. That construction is orthogonal exactly by design.

Table 27: Four published constructions under two optimizers, three seeds each, at one certified bound. Values are mean ± standard deviation over the seeds, at $\varepsilon = 36/255$. The last two columns are the two certificate conventions of Table 9. Every endpoint certifies at $L_{\text{global}} = 0.0312$. Rates are the screened ones. The SOC declared-rate cells are in the text.

Construction	Optimizer	Clean	Verified, product	Verified, per-pair
Cayley	Muon	0.286 ± 0.008	0.218 ± 0.000	0.267 ± 0.005
Cayley	SGD	0.236 ± 0.003	0.173 ± 0.007	0.221 ± 0.003
SOC	Muon	0.252 ± 0.007	0.182 ± 0.005	0.234 ± 0.007
SOC	SGD	0.213 ± 0.013	0.144 ± 0.015	0.197 ± 0.011
Björck	Muon	0.259 ± 0.015	0.187 ± 0.006	0.241 ± 0.013
Björck	SGD	0.206 ± 0.011	0.152 ± 0.012	0.193 ± 0.011
SLL	Muon	0.267 ± 0.014	0.199 ± 0.008	0.250 ± 0.011
SLL	SGD	0.208 ± 0.011	0.154 ± 0.015	0.196 ± 0.010

Five things are registered with this result, and none is repaired after the fact. Four of the six rate screens peaked at the edge of their grid. Those arms are reported unbracketed. On SOC we also ran the declared rate, where the ordering holds and the ranges stay disjoint. Both SLL arms ran at their declared rates. SGD's screen chose its rate directly, and Muon's fell to it by the registered tie rule. So SLL has no second rate, and nothing there separates an optimizer from its rate. The registered contrast is between training pipelines, each optimizer at its own rate. SOC and Björck are approximate where the Cayley anchor is exact. The effect is therefore shown on approximate constructions and not tested against exactness. SOC and Björck were chosen on implementation cost inside the window. SLL was chosen because its guarantee does not rest on orthogonality. Each registration stated its choice before any result. LiResNet (Hu et al., 2023) and LOT (Xu et al., 2022) are quoted from their papers and were not retrained here.

The accuracy shortfall of the anchor is a property of the protocol, and the three further constructions say the same. Under Muon the four spread 0.035 in clean accuracy, against a shortfall of 0.46 from the native recipes' 0.75. Under SGD they spread 0.030 against 0.51. The constructions differ from each other far less than all of them differ from their own published numbers.

T THE TOKEN-MIXING FAMILY

A token-mixing architecture without convolutions shows the same cap effect. We label it exploratory, because its registration is not witnessed by the code history. At the declared rate, unconstrained Muon diverges on that family. The capped arm survives at accuracy 0.69 and $G = 2.6 \times 10^4$. At the rate its own grid prefers, the unconstrained arms reach 3.5–6.1×10^6 against 1.3–1.4×10^4 capped. The ordering there does not reproduce, so the optimizer grading is claimed for the convolutional families only.