A TWO-SCALAR LAW FOR ACTIVATION FUNCTIONS: HOW PEAK SLOPE AND PEAK CURVATURE GOVERN ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Which intrinsic, measurable properties of an activation function govern adversarial robustness? We answer: exactly two scalars, in provably disjoint roles. For L-layer fully connected networks with any twice-differentiable activation, we derive an exact path-summed decomposition of every diagonal element of the loss Hessian and prove that each element admits a *closed-form* upper bound *affine* in $M_2 = \max|\sigma''|$—the peak curvature—while every other activation-specific quantity in the bound reduces to the peak slope $M_1 = \max|\sigma'|$ and to the prediction residual $|f - y|$. This *two-scalar law* yields four a-priori, falsifiable predictions: (P1) once M_2 is large enough that fitting capacity is sufficient—so that the residual $|f - y|$ no longer inflates the bound—the normalized Hessian diagonal norm grows approximately monotonically with M_2; (P2) below this regime, the activation's fitting capacity becomes insufficient, which inflates the residual $|f - y|$ and drives a second rise of the same norm; (P3) in this same regime—where the residual no longer differs across variants—larger M_1 yields a provably larger bound, and correspondingly larger measured norms and faster robustness degradation, in one and the same ordering; (P4) activations of distinct functional forms share the same rise–peak–decline robustness profile along the common axis M_2. Testing the law requires decoupling the two scalars; the Recursive Curvature-Tunable Activation Family (RCT-AF) provides exactly this, with α scanning $M_2 = c_\beta \alpha$ while $M_1(\beta) \in \{1.0,\ 1.1,\ 1.281\}$ stays fixed per variant independently of α. All four predictions are confirmed across ResNet-18 and WideResNet-28-10, CIFAR-10, CIFAR-100, and TinyImageNet-200, and DAJAT, DKL, and TRADES under AutoAttack: robustness consistently attains its maximum in a common intermediate band of M_2 for every functional form and dataset, and large-M_2 degradation rates order by M_1. These results reduce the design of robust activation functions to two measurable scalars.

1 INTRODUCTION

Adversarial training remains the most effective defense against adversarial examples Madry et al. (2017); Zhang et al. (2019), yet its progress has been overwhelmingly *algorithmic*: worst-case minimization Madry et al. (2017), decomposed losses Zhang et al. (2019), misclassification awareness Wang et al. (2020), data augmentation Addepalli et al. (2022), and decoupled divergence Cui et al. (2024). Comparatively little is known about *architectural* determinants of robustness—and among them, activation functions occupy a foundational position. Empirical reports that smooth activations can outperform ReLU under attack Xie et al. (2020) stop short of a sharper, quantitative question, which this paper answers:

> *Which intrinsic, measurable properties of an activation function govern adversarial robustness, and in what division of labor?*

We identify these properties as the peak derivatives $M_1 \triangleq \max_z |\sigma'(z)|$ and $M_2 \triangleq \max_z |\sigma''(z)|$, and prove that they govern loss-landscape sharpness—and hence robustness—through strictly disjoint channels. Our central result (Theorem 3.2) shows that for an L-layer network and every activation

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.

with bounded first and second derivatives, each diagonal element of the loss Hessian admits the closed-form upper bound

$$\left| [\nabla^2_\theta \hat{L}]_{kk} \right| \;\leq\; C_{base} \;+\; M_2 \cdot |f - y|\, c_k^2\, \mathcal{C}^{(l)}_{struct}, \tag{1}$$

where $f - y$ is the prediction residual and $\mathcal{C}^{(l)}_{struct}$ is a structural coefficient aggregating weights and propagated signals along backward paths. Three properties of Eq. 1 structure everything that follows. First, the *explicit* activation dependence reduces to M_2 alone: σ'' enters every backward path linearly (Appendix A, Theorem A.5). Second, every *remaining* activation-specific quantity is confined to $\mathcal{C}^{(l)}_{struct}$, whose activation dependence reduces to M_1 alone and is provably increasing in M_1 at fixed network state (Corollary 3.3)—the coefficient aggregates only weights and first-derivative quantities, containing no factor of σ''. Third, the residual factor is itself an implicit function of training state: at small M_2 the activation is near-linear, fit quality degrades, and the residual inflates the very same bound. The two scalars thus occupy disjoint, individually provable positions: M_2 *sets the slope of curvature amplification;* M_1 *sets the height of the structural coefficient*—the two-scalar law.

The law yields four a-priori, falsifiable predictions (Proposition 3.6): **(P1)** once M_2 is large enough that fitting capacity is sufficient—so that the residual $|f - y|$ no longer inflates the bound—the normalized Hessian diagonal norm grows approximately monotonically with M_2; **(P2)** below this regime, the activation's fitting capacity is insufficient, the inflated residual factor raises the same bound, and the norm rises again; **(P3)** once M_2 is large enough that no variant is limited by insufficient fitting capacity—so that the residual no longer differentiates the variants—larger M_1 produces a provably larger bound, predicting larger measured norms and faster robustness degradation in one and the same ordering; **(P4)** distinct functional forms share the same qualitative robustness profile along the common axis M_2.

Testing the law requires an activation platform in which the two scalars vary *independently*. The Recursive Curvature-Tunable Activation Family (RCT-AF) provides exactly this: α scans the peak curvature $M_2 = c_\beta \alpha$ along a shared axis, while the recursion order β fixes $M_1(\beta) \in \{1.0,\, 1.1,\, 1.281\}$ *independently of* α (Sec. 4). At any matched M_2, the three variants differ in functional form and in M_1 but in nothing else—a controlled comparison tailor-made for the law. Both training regimes confirm the full contract. Under standard training, the diagonal norm is U-shaped in M_2: beyond the onset of cross-form separation, $M_2 \gtrsim 10$, where no variant is limited by insufficient fitting capacity, it orders consistently by M_1; below $M_2 \approx 4$ the ordering *reverses*—the smallest-M_1 variant exhibits the largest norm, and, at matched $M_2 < 1$, also the lowest clean accuracy—as the residual channel dominates when fitting capacity is insufficient. Under adversarial training, every variant attains its robustness peak within $M_2 \in [4, 10]$ across two architectures, three datasets, and three methods; the variants share the same rise–peak–decline profile (**P4**, established by elimination: no law phrased in the raw parameter α survives, and M_2 is the only scalar held invariant), while their large-M_2 degradation rates order by M_1 beyond $M_2 \approx 10$—in the same ordering as the *independently measured* norms under standard training, a dual-channel agreement that isolates the M_1-coefficient channel (**P3**).

Our contributions are threefold:

- **A two-scalar law.** An exact, path-summed decomposition of Hessian diagonal elements for arbitrary depth (Appendix A, Theorem A.5), a closed-form upper bound affine in M_2 (Theorem 3.2), and a structural coefficient provably monotone in M_1 at fixed network state, with depth-linear sensitivity (Corollary 3.3)—all with complete proofs.
- **A falsifiable contract and a matched instrument.** Four a-priori predictions; RCT-AF, which decouples M_1 from M_2, is the minimal controlled platform for testing them.
- **Consistent verification.** All four predictions confirmed; each observed phenomenon is matched to a specific term of the same bound (Appendix B, Table 1).

2 Related Work

2.1 Adversarial Robustness Evaluation

Adversarial examples Szegedy et al. (2013) initiated an arms race between gradient-based attacks Goodfellow et al. (2015); Madry et al. (2017) and defenses. Many early defenses relied on gradient

masking Athalye et al. (2018), motivating AutoAttack (AA) Croce & Hein (2020) as the community-standard, parameter-free evaluation ensemble. We adopt AA throughout so that our conclusions rest on rigorous, attack-agnostic measurements. Reliable evaluation settles *how* robustness should be assessed, but does not explain which architectural properties produce it—our work addresses the latter.

2.2 Adversarial Training Methods

Adversarial training has advanced along several algorithmic axes: worst-case minimization over input perturbations Madry et al. (2017); decomposed losses that balance clean and robust objectives, as in TRADES Zhang et al. (2019); misclassification-aware reweighting of hard examples, as in MART Wang et al. (2020); data augmentation that suppresses non-robust features, as in DAJAT Addepalli et al. (2022); decoupled divergence objectives, as in DKL Cui et al. (2024); and weight-space regularization through adversarial weight perturbation (AWP) Wu et al. (2020), which flattens the robust loss landscape and composes with each of the recipes above. These advances are complementary to ours: we hold the training objective fixed and ask how *architectural* degrees of freedom—the two peak derivatives of the activation—modulate the loss landscape and hence the robustness attainable under any of these recipes.

2.3 Smooth Activations and Robustness

Replacing ReLU with smooth activations such as GELU Hendrycks & Gimpel (2016) and Swish Ramachandran et al. (2017) was reported to improve robustness Xie et al. (2020), an effect attributed to healthier gradient propagation. However, prior studies compare particular activations holistically and never isolate *which measurable property* carries the benefit: neither the peak curvature M_2 nor the peak slope M_1 has been identified as an operative scale, connected mechanistically to a robustness metric, or disentangled from the functional form. Supplying precisely this missing link—as a provable division of labor between the two scalars—is the central contribution of this work.

2.4 Loss-Landscape Sharpness and Robust Generalization

Sharp minima are associated with poorer generalization Keskar et al. (2016); Sharpness-Aware Minimization optimizes for flatness Foret et al. (2020); in adversarial settings, flatter minima correlate with better robust generalization Wu et al. (2020). Existing sharpness studies treat flatness as an outcome to be optimized; none connects it *quantitatively* to properties of the activation function, nor identifies *which* measurable activation scalars govern the sharpness scale, and in what division of labor. This is precisely the question our two-scalar law answers (Sec. 3).

2.5 Curvature Bounds and Curvature-Based Certificates

A complementary line of work engages second-order quantities directly: curvature-regularized training Moosavi-Dezfooli et al. (2019), low-curvature activations against adversarial overfitting Singla et al. (2021), and efficient curvature estimation Srinivas et al. (2022); Singla & Feizi (2020). Most relevant to our setup, Entesari et al. Entesari et al. (2024) propagate provable upper bounds on the Lipschitz constant of the *input-space Jacobian* $D_x f$ layer by layer, yielding certified and attack radii and a curvature regularizer. Their machinery consumes the activation's curvature constant—equal to M_2 whenever σ'' exists—as a fixed per-layer input, with M_1 entering only as a slope-restriction constant; no activation-level law emerges, whereas our law makes the two scalars themselves the governing variables of the *loss Hessian* $\nabla^2_\theta \hat{L}$—the training geometry, distinct from the input–output map their bounds certify (Sec. 3). The prescriptive consequences also diverge: monotone certificate analysis recommends $M_2 \to 0$ at the activation level, whereas our residual channel proves that this regime is fitting-limited and re-inflates the very same bound—yielding the empirically confirmed optimal band $M_2 \in [4, 10]$ instead of monotone minimization (Secs. 5.2–5.3). The two perspectives are complementary: certificates deliver worst-case guarantees, our law explains measured behavior; they agree on the large-M_2 direction, while only the latter accounts for the small-M_2 regime.

3 A TWO-SCALAR LAW FOR ACTIVATION FUNCTIONS

All statements in this section are self-contained; complete derivations are given in Appendix A.

3.1 SETUP AND STANDING ASSUMPTION

Consider an L-layer fully connected network with forward pass $z^{(l)} = W^{(l)}h^{(l-1)} + b^{(l)}$, $h^{(l)} = \sigma(z^{(l)})$, $h^{(0)} = x$, and scalar linear output $f = z^{(L)}$, trained with the squared surrogate loss $\hat{L}(\theta) = \frac{1}{2}(f-y)^2$. Let θ_k be the parameter attached to neuron i of layer l, with $c_k = \partial z_i^{(l)}/\partial\theta_k$ (equal to $h_j^{(l-1)}$ for weights and 1 for biases), and let $\mathcal{P}\big((a,b) \rightsquigarrow (c,d)\big)$ denote the set of directed paths from neuron (a,b) to neuron (c,d) in the computation graph; a path $P = \big((l,i_l),(l{+}1,i_{l+1}),\dots,(r,i_r)\big)$ connecting (l,i) to (r,j) satisfies $i_l \equiv i$ and $i_r \equiv j$ (the complete notation table is given in Appendix A, Sec. A.1).

Assumption 3.1 (Activation regularity). $\sigma \in C^2(\mathbb{R})$ with finite constants $M_1 := \sup_z |\sigma'(z)| < \infty$ and $M_2 := \sup_z |\sigma''(z)| < \infty$. All RCT-AF variants satisfy this assumption (Sec. 4); ReLU-type kinks are excluded.

3.2 THE (M_1, M_2)-BOUND: DIVISION OF LABOR

Theorem 3.2 (Closed-form (M_1, M_2) upper bound). *Grant Assumption 3.1. For every parameter θ_k attached to neuron i of a* hidden *layer $l \le L-1$ (output-layer parameters, $l = L$, are degenerate: the curvature channel vanishes and the bound holds with equality; Appendix A, Remark A.6),*

$$\big|\,[\nabla^2_\theta \hat{L}]_{kk}\,\big| \;\le\; \underbrace{M_1^{2(L-l)}\, G_{abs}(l,i)^2\, c_k^2}_{\textit{base term}} \;+\; \underbrace{M_2\,|f-y|\,c_k^2\,\mathcal{C}^{(l)}_{struct}(M_1)}_{\textit{amplification term}}, \tag{2}$$

$$\mathcal{C}^{(l)}_{struct}(M_1) \;=\; \sum_{r=l}^{L-1}\sum_{j=1}^{n_r} M_1^{L-1+r-2l}\, G_{abs}(r,j)\,\big[\mathcal{A}_j^{(r)}(l,i)\big]^2, \tag{3}$$

where the two path sums depend only on the weights,

$$G_{abs}(a,b) := \sum_{P:\,(a,b)\rightsquigarrow(L,1)}\;\prod_{s=a}^{L-1}\big|W^{(s+1)}_{i_{s+1}i_s}\big|, \tag{4}$$

$$\mathcal{A}_j^{(r)}(l,i) := \sum_{P\in\mathcal{P}((l,i)\rightsquigarrow(r,j))}\;\prod_{s=l}^{r-1}\big|W^{(s+1)}_{i_{s+1}i_s}\big|. \tag{5}$$

Every factor of Eq. 2 is an activation scalar (M_1 or M_2), the prediction residual $|f-y|$, the parameter sensitivity c_k, or a weight-only path sum. At fixed network state, the bound is affine in M_2 with non-negative slope $|f-y|\,c_k^2\,\mathcal{C}^{(l)}_{struct}(M_1)$, and $\mathcal{C}^{(l)}_{struct}$ contains no dependence on M_2 *(Appendix A, Theorems A.15–A.16).*

Eq. 2 is the law's final form, and its three channels are visible on its face. The amplification term carries M_2 exactly once—the slope of curvature amplification; it carries $|f-y|$ exactly once—the residual, a training-state factor; and every remaining activation dependence sits in the powers of M_1: $M_1^{2(L-l)}$ in the base term and $M_1^{L-1+r-2l}$ in the amplification term, with exponents ranging from $L-1-l$ (at $r=l$) to $2(L-1-l)$ (at $r=L-1$). No factor of σ'' appears anywhere outside the explicit M_2: the σ'-mediated propagation along backward paths has been absorbed entirely into the M_1-powers and the weight-only path sums.

Loss scope. Theorem 3.2 is stated for the scalar squared surrogate, but its channel structure is loss-agnostic: for any twice-differentiable output loss ℓ, the same decomposition holds with $|f-y|$ replaced by the output-gradient norm $\|\partial\ell/\partial f\|_1$ (for cross-entropy, the total probability residual $\sum_c |p_c - y_c|$), while σ'' still enters linearly and the M_1-powers are unchanged. All experiments measure the training loss (cross-entropy); see Appendix A, Remark A.18.

Corollary 3.3 (Monotonicity in M_1 at matched M_2). *Suppose the* state factors *do not differentiate the variants: M_2 is large enough that every variant attains sufficient fitting capacity, so that comparisons*

at matched M_2 share a common residual $|f - y|$ and, for weight parameters, a common sensitivity $c_k = h_j^{(l-1)}$ (for bias parameters $c_k \equiv 1$; Appendix A, Theorem A.16(ii)). Then, at matched M_2 and fixed weights, the bound Eq. 2 is a polynomial in M_1 with non-negative coefficients: the base term carries the exponent $2(L - l)$, and the amplification term carries exponents $L - 1 + r - 2l$ ranging from $L - 1 - l$ to $2(L - 1 - l)$. Hence it is non-decreasing—generically strictly increasing—in M_1, and its M_1-sensitivity grows linearly with the depth above the parameter's layer (Appendix A, Theorem A.16).

Remark 3.4 (Disjoint channels). Eq. 2 is carried by three provably disjoint channels: the *M_2-slope channel*—the single explicit factor M_2 of the amplification term, the sole locus through which σ'' enters; the *M_1-coefficient channel*—the powers $M_1^{2(L-l)}$ and $M_1^{L-1+r-2l}$, contributed by the σ'-mediated propagation along backward paths, whose expansions contain no factor of σ'' (Appendix A); and the *residual channel*—the training-state factor $|f - y|$. For RCT-AF this division is directly testable, since $M_1(\beta)$ is fixed per variant while $M_2 = c_\beta \alpha$ sweeps a common axis (Sec. 4).

Remark 3.5 (The residual channel). The factor $|f - y|$ depends on the training state and, indirectly, on M_2: at small M_2 the activation approaches linearity, fit quality degrades, and the residual grows—as measured directly in Fig. 1b—inflating the amplification term even though M_2 itself is small. The bound supplies the multiplicative coupling between the residual and the curvature term; the residual growth itself is an empirical fact supplied by measurement (Sec. 5.3).

3.3 Measurement Target

To confront the law with measurements we summarize all diagonal elements by the *normalized Hessian diagonal L_2 norm*

$$\left\| \operatorname{diag}\left(\nabla_\theta^2 \hat{L}(\theta)\right)/p \right\|_2 = \sqrt{\frac{1}{p} \sum_{k=1}^{p} \left[\nabla_\theta^2 \hat{L}(\theta)\right]_{kk}^2}, \tag{6}$$

where p is the total number of model parameters. The norm is defined for whichever loss $\hat{L}$ is specified; in the experiments it is the training loss (cross-entropy), for which Remark A.18 guarantees that the three channels of the bound carry over. The norm is adopted for two reasons. First, it is a *direct sharpness measure*: each diagonal entry is the second derivative of the loss along one parameter axis at the converged point, so the norm aggregates squared curvatures over all directions—the smaller the norm, the flatter the landscape. Second, flatness is precisely the geometry on which robust generalization depends: flatter minima are associated with better generalization and, in adversarial settings, with better robust generalization Keskar et al. (2016); Foret et al. (2020); Wu et al. (2020); the norm therefore tracks the landscape quantity that robustness exploits. One scope convention: Theorem 3.2 bounds each diagonal element *per example*, whereas Eq. 6 aggregates over the test set; aggregating the per-example bounds preserves the channel structure—affine in M_2, M_1-powered coefficients, state factors entering through their data-set averages—so comparisons between Eq. 6 and the bound are made at the *channel level*, which is exactly the level at which P1–P4 are stated.

3.4 Falsifiable Predictions

Proposition 3.6 (Prediction contract)**.** *Under Assumption 3.1, the bound Eq. 2 supports the following conditional predictions.* ***(P1)*** *Once M_2 is large enough that fitting capacity is sufficient—so that the residual no longer inflates the bound—the amplification term of Eq. 2 is affine in M_2 (M_2-slope channel, Theorem 3.2); the normalized diagonal norm Eq. 6 is predicted to grow approximately monotonically with M_2 in this regime.* ***(P2)*** *Below this regime, the activation approaches linearity, its fitting capacity becomes insufficient, and the inflated residual factor raises the same bound (residual channel, Remark 3.5); the norm is predicted to exhibit a second rise.* ***(P3)*** *In this same regime—where the residual no longer differs across variants—larger M_1 yields a provably larger bound (M_1-coefficient channel, Corollary 3.3), predicting larger measured norms and, under adversarial training, faster robustness degradation, with the separation widening with depth.* ***(P4)*** *Activations of distinct functional forms follow the same qualitative robustness pattern along the common axis M_2—improving at small M_2, peaking within a common band, declining beyond it—since the explicit activation dependence of the bound reduces to M_2 alone. Rate differences at large M_2 belong to* ***P3****, not to violations of* ***P4****.*

Two conventions govern the tests of P1–P4 (Secs. 5.2–5.5). First, the *norm-to-robustness bridge*: predictions **P3**–**P4** concern robust accuracy, which inherits the norm predictions through the empirically established flatness–robustness correspondence Keskar et al. (2016); Foret et al. (2020); Wu et al. (2020); this bridge is empirical and is not part of Theorem 3.2. Second, the *depth-widening signature* in **P3** is a structural consequence of the exponents in Eq. 3; the present experiments compare architectures that differ in width and depth simultaneously (Appendix D.2) and do not isolate depth, so we report it as a falsifiable structural prediction rather than a confirmed finding (Sec. 5.4).

4 THE RCT-AF TESTBED

RCT-AF enters this paper as an *instrument*, not a contribution: the prediction contract of Sec. 3.4 requires an activation family in which M_1 and M_2 vary independently, and RCT-AF is engineered to that specification.

4.1 DEFINITION

The *base function* ($\beta = 0$) is a parameterized softplus Glorot et al. (2011),

$$\sigma_0(x) = \frac{1}{\alpha} \ln(1 + e^{\alpha x}) \,, \tag{7}$$

and the *recursion operator* is $T[\sigma](x) = \sigma'(x) \cdot x$, so that $\sigma_\beta(x) = T[\sigma_{\beta-1}](x)$ for $\beta \geq 1$. The first two recursions give

$$\sigma_1(x) = \frac{x}{1 + e^{-\alpha x}}, \tag{8}$$

$$\sigma_2(x) = \frac{\left(e^{2\alpha x} + (\alpha x + 1)e^{\alpha x}\right)x}{e^{2\alpha x} + 2e^{\alpha x} + 1}; \tag{9}$$

σ_1 recovers a parameterized Swish Ramachandran et al. (2017). Figure 3 (Appendix C) plots the resulting functional forms; the closed forms of σ'_β and σ''_β are collected in Appendix C.1.

4.2 DECOUPLING SPECIFICATION

The two properties that make RCT-AF a lawful instrument are read directly off its derivatives.

Property 1: M_1 is α-invariant. The first derivatives (Fig. 4, Appendix C) show that for each variant the peak slope is a constant of the recursion order alone,

$$M_1(0) = 1.0, \qquad M_1(1) \approx 1.1, \qquad M_1(2) \approx 1.281, \tag{10}$$

independent of α—the curves' heights do not change as α ranges over $\{1, 5, 10, 15, 20\}$. This follows from the scaling form $\sigma_\beta(x) = \alpha^{-1}\Phi_\beta(\alpha x)$, under which $\sigma'_\beta(x) = \Phi'_\beta(\alpha x)$. Property 1 is the experimental premise that renders matched-M_2 comparisons well posed: at matched M_2, the activation-specific quantities entering the bound Eq. 2 are the coefficient height M_1 and—through fit quality—the residual; once M_2 is large enough that every variant attains adequate fit, the residual no longer differentiates the variants and M_1 is the sole operative difference (Secs. 5.3–5.4).

Property 2: M_2 is exactly proportional to α. The second derivatives (Fig. 5, closed forms in Appendix C.1) show

$$M_2 \triangleq \max_x |\sigma''_\beta(x)| = c_\beta \, \alpha, \qquad c_\beta \in \{1/4,\, 1/2,\, 1\}, \tag{11}$$

with the same proportionality for every β. Sweeping α therefore scans M_2 along a *shared* axis for all three functional forms.

Together, Properties 1–2 define the decoupling: (α, β) maps onto (M_2, M_1) with orthogonal axes. The family spans the three qualitatively distinct regimes of the bound—*fitting-limited* ($M_2 \lesssim 4$; the defaults of GELU and Swish, $M_2 \leq 1$, lie deep inside this regime), *balanced* ($M_2 \in [4, 10]$), and *amplification-dominated* ($M_2 > 10$, with clean cross-form comparisons once $M_2 \gtrsim 10$)—so that every channel of the bound can be exercised. These ranges are deliberately not contiguous: they mark the regimes in which a single channel dominates, rather than a partition of the axis; between them lie transition regions in which the channels trade dominance.

5 EXPERIMENTAL VERIFICATION

We test the contract of Proposition 3.6 using two training regimes in complementary roles. *Standard training on clean data* supplies the mechanism-level measurements—Hessian diagonal norms and clean accuracy—verifying **P1** and **P2** (Secs. 5.2–5.3) and the norm channel of **P3** (Sec. 5.4). *Adversarial training* supplies the consequence-level measurements—AutoAttack robust accuracy—verifying the degradation channel of **P3** and the profile-sharing prediction **P4** (Secs. 5.4–5.5). Section D.1 confronts the resulting design principle with commonly used activations, and Section D.2 confirms consistency across settings.

5.1 SETUP

Adversarial-training suite. ResNet-18 He et al. (2016) on CIFAR-10 Krizhevsky et al. (2014) is the default configuration, and each experimental block varies a single axis from it. Adversarially trained results use three representative methods—DAJAT Addepalli et al. (2022), DKL Cui et al. (2024), and TRADES Zhang et al. (2019)—all with ℓ_∞ budget $\epsilon = 8/255$, original hyperparameters, and *adversarial weight perturbation* (AWP) Wu et al. (2020) enabled by default. The main-text adversarial results (Secs. 5.4–5.5) use DKL on CIFAR-10/ResNet-18; the cross-setting checks (Appendix D.2) vary one axis at a time: the training method (DAJAT, DKL, and TRADES on CIFAR-10/ResNet-18), the architecture (TRADES additionally on WideResNet-28-10 Zagoruyko & Komodakis (2016)), and the dataset (DKL on CIFAR-10, CIFAR-100, and TinyImageNet-200 Le & Yang (2015)). The activation comparison (Table 2, Appendix D.1) covers all three methods on all three datasets. Robustness is evaluated exclusively via AutoAttack (AA) Croce & Hein (2020).

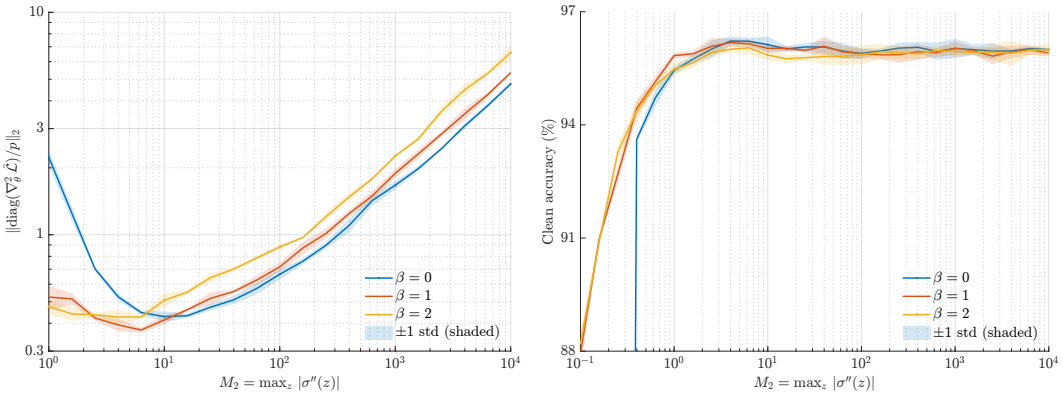
Controlled curvature sweep. In both regimes, the training recipe is fixed per run and only the activation profile varies through (α, β). By Eqs. Eq. 10–Eq. 11, sweeping α scans $M_2 = c_\beta \alpha$ along a *shared curvature axis* at fixed $M_1(\beta)$: in the adversarial-training sweep, the per-variant α ranges are chosen so that all three variants cover the common axis $M_2 \in [1, 50]$ (each variant evaluated at $\alpha = M_2/c_\beta$; Fig. 2b), while the standard-training sweep extends over a wide log-spaced curvature range to expose the full amplification-dominated regime (Fig. 1a). The sweep covers the fitting-limited ($M_2 < 4$), balanced ($M_2 \in [4, 10]$), and amplification-dominated ($M_2 > 10$) regimes; within the latter, matched $M_2 \gtrsim 10$ realizes the condition under which Corollary 3.3 applies—every variant attains sufficient fitting capacity and the state factors no longer differentiate the variants. Every cross-form comparison is therefore made *at matched* M_2, where the bound isolates the M_1-carrying quantities—the coefficient height and, through fit quality, the residual.

Measurement. Hessian diagonal norms Eq. 6 are computed on the full test set after standard training, for the training loss (cross-entropy; Remark A.18), via signed square-root finite differences of gradients—the signed square root suppressing floating-point cancellation in the difference; clean accuracy is evaluated on the same set. Every configuration is trained over five independent runs, and all results are reported as means $\pm$ standard deviations over these runs. Throughout, we write $M_2^\star \approx 10$ for the empirical *onset of cross-form separation*, read off the convergence of the clean-accuracy curves (Fig. 1b); all large-M_2 statements below refer to $M_2 \geq M_2^\star$.

5.2 **P1**: MONOTONIC AMPLIFICATION AT LARGE M_2

Figure 1a plots the normalized Hessian diagonal norm, measured after standard training, against M_2 for all three variants. For every functional form the curve is U-shaped, attaining its minimum within $M_2 \in [4, 10]$; the right branch realizes **P1**. Beyond $M_2 \approx 10$, the norm rises consistently and monotonically for $\beta = 0, 1, 2$ alike—matching the directional forecast of Theorem 3.2, whose amplification term $M_2 \, |f - y| \, c_k^2 \, \mathcal{C}_{struct}^{(l)}(M_1)$ is affine in M_2 at fixed network state. The realized growth on the log–log axes is a sublinear power law rather than the bound's unit slope, as expected: the bound holds the state factors fixed, whereas along a trained sweep the residual, the sensitivities, and the weight-only path sums all vary. At matched $M_2 \gtrsim 10$, where no variant is fitting-limited, the curves additionally order by M_1; this ordering is the province of **P3** (Sec. 5.4).

Consistent with **P1**, the rising right branch follows the direction anticipated by the bound Eq. 2.



(a) Normalized Hessian Diagonal Norm vs. M_2

(b) Clean accuracy vs. M_2

Figure 1: Standard training on CIFAR-10 with ResNet-18 (SGD), $\beta = 0, 1, 2$; means $\pm$ std over five runs, full test set, training loss (cross-entropy). **(a)** Normalized Hessian diagonal norm Eq. 6 vs. M_2: U-shaped for every functional form, minimum in the band $M_2 \in [4, 10]$. Right branch ($M_2 > 10$): consistent monotone rise (M_2-slope channel, **P1**); at matched $M_2 \gtrsim 10$ the curves order by M_1 (**P3**). Left branch ($M_2 < 4$): second rise (residual channel, **P2**); the ordering *reverses*—$\beta = 0$, smallest M_1 but weakest fitting capacity, shows the largest norm as underfitting inflates the residual. **(b)** Clean accuracy vs. M_2: sharp decline below $M_2 \approx 4$, with $\beta = 0$ lowest at matched $M_2 < 1$—direct evidence that the small-M_2 residual is form-dependent. The full channel-level reading is given in Secs. 5.2–5.4.

5.3 **P2**: Residual-Driven Rise at Small M_2

The left branch—below $M_2 \approx 4$—exhibits a second rise, and the bound explains it through its own residual factor (Remark 3.5): at small M_2 the activation approaches linearity, fit quality degrades, the residual grows, and the amplification term is inflated even though M_2 is small. The measurements supply both halves of the mechanism: the norm rises toward small M_2 for all three variants (Fig. 1a), and clean accuracy declines markedly in the same region (Fig. 1b)—direct evidence of the underfitting that drives the residual.

The rise is moreover form-dependent in exactly the direction the residual channel requires. In this regime $\beta = 0$—the variant with the *smallest* M_1—shows the *largest* norm, together with the lowest clean accuracy (at matched $M_2 < 1$, its clean accuracy is clearly below that of $\beta = 1, 2$): the most linear variant underfits most and inflates the residual the most. The β-ordering thus reverses between the two regimes—largest-norm variant $\beta = 0$ at small M_2, smallest at large M_2—a crossover that discriminates the two channels: the M_1-coefficient channel alone would predict the opposite ordering at small M_2.

Consistent with **P2**, the empirical U-shape is carried by two distinct terms of one inequality—a curvature-driven right branch and a residual-driven left branch, the latter dominant precisely where fitting capacity fails—with the minimum in between.

5.4 **P3**: M_1-Ordered Amplification and Degradation at Matched M_2

Theorem 3.2 and Corollary 3.3 predict that, once no variant is fitting-limited, the bound at matched M_2 increases with M_1; by Eq. Eq. 10 the variants realize $M_1 \in \{1.0,\ 1.1,\ 1.281\}$, so matched-$M_2$ comparisons read directly as M_1 comparisons. We test this through two independently measured channels, in the regime where the prediction applies: matched M_2 at or above the onset of separation $M_2^\star \approx 10$, large enough that every variant attains sufficient fitting capacity and residual differences across variants are negligible ($M_2 \gtrsim 10$; Fig. 1b).

Norm channel (standard training). In the right branch of Fig. 1a—at matched $M_2 \gtrsim 10$—the measured norm increases with β, i.e., with M_1: 1.28-variant above 1.1-variant above 1.0-variant, throughout the regime $M_2 \gtrsim 10$.

Degradation channel (adversarial training). In Fig. 2(b), beyond $M_2 \approx 10$ the three variants decline at visibly different rates: larger β—larger M_1—degrades faster under AutoAttack. The ordering of degradation rates coincides with the ordering of the measured norms.

The agreement of two independently measured quantities—landscape curvature under standard training, robust accuracy under adversarial training—in one and the same M_1 ordering constitutes strong evidence for the M_1-coefficient channel: in the closed-form bound Eq. 2, M_1 enters through the exponents $2(L-l)$ and $L-1+r-2l$, which are independent of M_2 and attach to weight-only path sums (Corollary 3.3)—so larger M_1 raises the bound, and is tracked by larger measured norms and faster degradation alike, without touching the M_2 slope. Three qualifications keep the claim precise. First, the ordering is guaranteed for the *bound* and tracked empirically by the measurements, realized values possibly falling below the bound through sign cancellation. Second, larger β simultaneously endows σ'' with more local extrema ($2\beta + 1$), which affects only the tightness of $|\sigma''| \leq M_2$—the provable channel is M_1 (Appendix A, Remark A.22). Third, the depth-widening signature of **P3** is not isolated by the present experiments: the architecture comparison of Appendix D.2 varies width and depth simultaneously and uses a single variant, so we report it as a falsifiable structural prediction for future work, not a confirmed finding.

Consistent with **P3**, at matched M_2 in this regime, both the measured norm and the robustness degradation rate increase with M_1, in the ordering prescribed by Corollary 3.3.

5.5 **P4**: Shared Robustness Profile Across Functional Forms

Figure 2 establishes **P4** by elimination. Panel (a) plots AutoAttack robust accuracy against the raw recursion strength α for $\beta = 0, 1, 2$ under DKL on CIFAR-10/ResNet-18: because identical α induces three distinct peak curvatures ($M_2 = c_\beta \alpha$), the curves separate markedly, and no law phrased in α can account for the data. Panel (b) places all three variants on the shared curvature axis $M_2 \in [1, 50]$—each evaluated at its own $\alpha = M_2/c_\beta$—so the comparison is at matched M_2 across the full range: the three variants exhibit the same qualitative response—robustness improves monotonically while $M_2 < 4$, attains its maximum within the shared band $M_2 \in [4, 10]$, and declines thereafter. The curves share the profile but not the pointwise values: at large M_2 they fan out in the M_1 ordering, which is the province of **P3**, not a violation of **P4**. Between the two panels the elimination is complete: α fails to explain the data, and the only quantity held invariant across the divergent functional forms is the explicit scalar of the bound, M_2.

Consistent with **P4**, distinct functional forms aligned by M_2 share the same rise–peak–decline profile and the same optimal band.

Appendix D.1 confronts the resulting design principle with commonly used activations across three training methods and three datasets, and Appendix D.2 confirms cross-setting consistency.

6 Conclusion

We showed that the influence of an activation function on adversarial robustness is carried by two measurable scalars in provably disjoint roles: $M_2 = \max |\sigma''|$ sets the slope of curvature amplification on the loss Hessian, $M_1 = \max |\sigma'|$ sets the height of the structural coefficient, and the training-state residual couples the two regimes into a single bound. The resulting two-scalar law generated four a-priori, falsifiable predictions, all confirmed across two architectures, three datasets, and three adversarial training methods under AutoAttack. Its prescription is actionable: tune M_2 into the intermediate band where fitting capacity is sufficient yet curvature amplification has not yet sharpened the landscape—there the functional form becomes secondary, and the cheapest form sustaining adequate fit ($\beta = 0$ or $\beta = 1$) is preferred. More broadly, the law reduces activation design for robustness from an empirical search to a two-parameter design space with provable structure.

AI use statement

We did not use generative AI tools for any task with required disclosure: generating synthetic data, formulating research ideas or mathematical propositions, assisting in proof derivation or verification, designing methodologies or experiments, implementing code, and interpreting results were all performed entirely by the authors. We used generative AI tools only for tasks with recommended

disclosure, namely polishing the language and editing the manuscript for clarity and grammar. We have reviewed and edited all AI-assisted text. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

Complete, self-contained proofs of all theoretical claims (Assumption 3.1, Theorem 3.2, Corollary 3.3, and Proposition 3.6, built on the exact decomposition of Theorem A.5 and the closed-form bound of Theorem A.17) are provided in Appendix A, together with a full notation table (Sec. A.1) and closed-form expressions for the RCT-AF derivatives (Sec. C.1), from which the decoupling properties (Eqs. 10–11) can be verified directly. The experimental protocol—architectures, adversarial-training methods, hyperparameters, AutoAttack evaluation, and the Hessian-diagonal-norm measurement—is specified in Sec. 5.1; all results are reported as means $\pm$ standard deviations over five independent runs. Extended cross-architecture, cross-method, and cross-dataset validation is given in Appendix D, and the computational-overhead analysis in Appendix E.

Ethics statement

This work studies how intrinsic, measurable properties of activation functions govern adversarial robustness, with the goal of improving the design and reliability of adversarially trained models. All experiments use public benchmarks (CIFAR-10, CIFAR-100, TinyImageNet-200) under the standard ℓ_∞ threat model ($\epsilon = 8/255$), involve no human subjects or sensitive data, and release no new datasets. The proposed design principle is defensive in nature; we do not anticipate ethical risks beyond those already accepted in the study of adversarial robustness.

References

Sravanti Addepalli, Samyak Jain, et al. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems*, 35:1488–1501, 2022.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, July 2018. URL https://arxiv.org/abs/1802.00420.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. *Advances in Neural Information Processing Systems*, 37:74461–74486, 2024.

Taha Entesari, Sina Sharifi, and Mahyar Fazlyab. Compositional curvature bounds for deep neural networks. In *International Conference on Machine Learning (ICML)*, 2024.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 and CIFAR-100 datasets. 2014. Available at: `http://www.cs.toronto.edu/~kriz/cifar.html`.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *CVPR*, 2019.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In *ICML*, 2020.

Vasu Singla, Sahil Singla, Soheil Feizi, and David Jacobs. Low curvature activations reduce overfitting in adversarial training. In *ICCV*, 2021.

Suraj Srinivas, Kenji Matoba, Hima Lakkaraju, and François Fleuret. Efficient training of low-curvature neural networks. *Advances in Neural Information Processing Systems*, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rklOg6EFwS`.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.