ACTIVATION-MASK STABILITY AND ADVERSARIAL VULNERABILITY: A MEASUREMENT-COUPLING CASE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

A statistic measured *at* an adversarial example can appear extraordinarily predictive of that example's adversarial distance even when it carries little independent information, simply because the measurement point is chosen by the same search that defines the distance. We demonstrate this for activation-mask stability $D_M(x, \delta) = \|M(x + \delta) - M(x)\|_0/N$ in ReLU networks. Across five sparsity-regularization strengths and three datasets (MNIST, FashionMNIST, CIFAR-10, plus a PGD-AT CIFAR-10 control), per-example activation sparsity $S(x)$ is a weak-to-null correlate of adversarial distance $R(x)$ ($r \in [-0.20, 0.31]$), whereas D_M measured *at* the minimal perturbation found by DeepFool is strong ($r = 0.91$–0.98). But $R(x) = \|\delta\|_2$ is exactly the distance to that same point, so the two are coupled by construction. Using the same 20 checkpoints and no new experiments beyond what we already ran, we test this three ways: two δ-free alternatives — fixed-radius mask sensitivity $D^{\epsilon}_M(x)$ and clean-input boundary distance $A_{FC}(x)$ — both collapse relative to the attack-point measurement ($r = 0.10$–0.17 and near-zero to -0.23, respectively); and a crossing-density decomposition $C(x) = D_M/R$ testing whether D_M is merely proportional to distance traveled, which holds on MNIST ($r = -0.04$) but not the other three datasets ($r = -0.11$ to -0.35). Together these support reading most, but not all, of the attack-point correlation as measurement coupling rather than an independent relationship of comparable size, with a smaller residual ($|r| \approx 0.1$–0.35) surviving every way we decoupled it and reproducing under an independent CW-L2-style attack. We frame this as a case study of one measurement-coupling mechanism in one small ReLU architecture, not a general claim about activation sparsity or mask-stability metrics at large.

1 MOTIVATION

A natural intuition connects activation sparsity to adversarial robustness: a sparser hidden representation activates fewer computational paths, so a perturbation has fewer routes through which to affect the output. Prior work supports a version of this story, and a closely related literature explains robustness via the stability of a network's activation path rather than sparsity alone (Section 2), all of it pitched at the level of trained models — e.g., comparing mean sparsity or mean stability against mean robust accuracy *across* models. Our question is narrower and sits one level down: *within* a single trained model, does an individual example's sparsity $S(x)$ predict that same example's adversarial distance $R(x)$? This per-example question is logically independent of the model-level one — a model-level correlation between mean sparsity and robust accuracy is compatible with sparsity carrying no information about *which* examples within that model are vulnerable, and our results below should be read as a claim about the latter, not a challenge to the former.

We ask a more specific question, motivated by the fact that two activation patterns can have identical sparsity while encoding completely different computations:

$$M_1 = (1, 1, 1, 0, 0, 0, 0, 0), \qquad M_2 = (0, 0, 0, 1, 1, 1, 0, 0)$$

are both 62.5% sparse but geometrically distinct. Because a ReLU network is piecewise linear, an adversarial perturbation can succeed in two distinct ways: by moving within the current 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.

region (exploiting the local linear classifier), or by flipping which units are active and moving into a different linear region entirely. This suggests that what should matter for robustness is not *how many* units are off, but *how easily the set of active units changes* under a small perturbation.

2 Related Work

Activation sparsity and adversarial robustness. Guo et al. (2018) report that sparser hidden representations correlate with improved adversarial robustness at the model level, and Peng et al. (2023) offer a theoretical account connecting activation sparsity to flat minima and robustness. Krithivasan et al. (2020) show the opposite-direction effect: adversarial perturbations can be constructed specifically to *destroy* sparsity, e.g. to mount energy/latency attacks. All of this evidence is at the level of trained models (mean sparsity vs. mean robust accuracy); our per-example question — does an individual example's sparsity predict that same example's adversarial distance, within one fixed model — sits one level below it and is not in tension with these findings (Section 1).

Activation paths, linear regions, and robustness geometry. Jiang et al. (2023) argue for explaining robustness via the stability of a network's activation path rather than sparsity alone, a reframing this paper's results are broadly consistent with (Section 6); our distinctive contribution is diagnosing a specific measurement hazard in testing that kind of claim, not the reframing itself. More broadly, the piecewise-linear structure of ReLU networks has been used to relate local decision-boundary geometry to robustness (Fawzi et al., 2016), and to derive certified distance-to-boundary guarantees from local Lipschitz bounds (Hein & Andriushchenko, 2017) or extreme-value-based estimators (Weng et al., 2018). Our clean-input boundary distance $A(x)$ (Section 5.6) is a much cruder, uncertified relative of this idea, used here only to probe circularity rather than to provide guarantees — and, as Appendix B.3 and Appendix D show, even this crude version runs into real estimation difficulties (extreme-value dilution when minimizing over many units) that a more careful treatment would need to address.

Minimal-distortion attacks and adversarial distance estimation. We use DeepFool (Moosavi-Dezfooli et al., 2016) as our primary $R(x)$ estimator and cross-check with a simplified CW-ℓ_2 attack (Carlini & Wagner, 2017) (Section 5.8); more rigorous ensembles of parameter-free attacks such as AutoAttack (Croce & Hein, 2020) exist and would strengthen a full validation of $R(x)$ beyond the spot check we run here (Section 6).

Measurement-conditional statistics and endogenous selection bias. The core methodological concern in this paper — that a statistic evaluated at a point chosen by the same procedure that defines the outcome can inflate the apparent relationship between them — is a specific instance of a broader phenomenon known in the causal-inference and statistics literature as conditioning on a collider, or endogenous selection bias (Elwert & Winship, 2014). To our knowledge this general risk has not previously been made concrete and quantified for activation-based robustness measures specifically; doing so, and showing how much of a widely-plausible-looking correlation it can account for in one setting, is this paper's primary contribution.

3 Definitions

For a fixed input x and hidden layer activations $a(x)$, define the binary activation mask

$$M(x) = \mathbf{1}[a(x) > 0].$$

Per-example activation sparsity, pooled across all layers, is

$$S(x) = \frac{\#\{i : a_i(x) = 0\}}{N}.$$

Given an adversarial perturbation δ found for x, the mask instability is the normalized Hamming distance between the clean and adversarial masks,

$$D_M(x, \delta) = \frac{\|M(x+\delta) - M(x)\|_0}{N}.$$

Adversarial vulnerability is the minimal perturbation norm required to flip the predicted label,

$$R(x) = \min_{\delta} \|\delta\|_2 \quad \text{s.t.} \quad \arg\max f(x+\delta) \neq \arg\max f(x).$$

We approximate $R(x)$ and δ per example with DeepFool (Moosavi-Dezfooli et al., 2016), an iterative method designed to find a near-minimal ℓ_2 perturbation, rather than a fixed-budget attack such as PGD, so that $R(x)$ is a genuine per-example distance rather than a threshold-crossing indicator.

4 EXPERIMENTAL SETUP

Model. A single architecture throughout: three 3×3 conv layers (32/64/64 channels, ReLU, two 2×2 max-pools) followed by one 128-unit ReLU fully-connected layer and a linear classifier head. This keeps every hidden unit a clean ReLU with an unambiguous on/off mask, at the cost of architectural diversity (Section 6).

Sparsity sweep. For each dataset we train five models differing only in an ℓ_1 penalty on post-ReLU activations,

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \sum_l \|h_l(x)\|_1, \qquad \lambda \in \{0,\ 10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1}\},$$

trained with Adam (MNIST, FashionMNIST: 8 epochs; CIFAR-10: 15 epochs), producing models whose mean activation sparsity spans roughly 50–92% (Table 1).

Adversarial-training control. All models above are undefended. Because CIFAR-10 in particular ends up uniformly fragile (Section 5.4), we add one more sweep: five CIFAR-10 models trained identically but with 7-step PGD-ℓ_∞ adversarial training (Madry et al., 2018) ($\epsilon = 8/255$, $\alpha = 2/255$) replacing clean images during the forward/backward pass, same 15-epoch budget. This is a direct within-dataset control for whether the sparsity/mask-stability relationship depends on the model being undefended.

Per-example metrics. For 500 correctly-classified test images per model (each model's own first 500 correctly-classified images in a fixed iteration order, not a literal fixed set shared across λ; Appendix A) we run untargeted DeepFool (up to 50 iterations, overshoot 0.02) to obtain δ, $R(x) = \|\delta\|_2$, and the resulting $D_M(x, \delta)$ (pooled across all four ReLU layers). We also record clean-image softmax confidence and logit margin, and a standard fixed-budget PGD-ℓ_∞ robust accuracy (MNIST, FashionMNIST: $\epsilon = 0.1$; CIFAR-10: $\epsilon = 8/255$; 20 steps) as a sanity-checked, independent measure of overall model robustness.

Table 1: Model sweep: clean accuracy, PGD robust accuracy (fixed-budget sanity check), mean per-example sparsity/mask instability, and the per-example Pearson correlation of each with adversarial distance $R(x)$ (across the 500 evaluated examples), at each λ. CIFAR-10-AT is the adversarially-trained control; all other rows are undefended models.

Dataset	λ	Clean acc.	PGD robust acc.	mean $S(x)$	$\text{corr}(S, R)$	mean $D_M(x, \delta)$	$\text{corr}(D_M, R)$
MNIST	0	0.992	0.671	0.592	−0.153	0.037	0.971
MNIST	0.0001	0.990	0.728	0.666	−0.097	0.047	0.972
MNIST	0.001	0.989	0.608	0.646	−0.133	0.061	0.974
MNIST	0.01	0.992	0.619	0.727	−0.007	0.057	0.977
MNIST	0.1	0.995	0.720	0.922	−0.065	0.025	0.977
FashionMNIST	0	0.919	0.000	0.501	−0.087	0.039	0.924
FashionMNIST	0.0001	0.917	0.000	0.552	−0.108	0.030	0.964
FashionMNIST	0.001	0.914	0.000	0.543	0.310	0.032	0.963
FashionMNIST	0.01	0.918	0.000	0.593	0.180	0.044	0.942
FashionMNIST	0.1	0.926	0.000	0.715	−0.201	0.039	0.963
CIFAR-10	0	0.723	0.002	0.596	0.221	0.012	0.925
CIFAR-10	0.0001	0.719	0.001	0.552	0.058	0.012	0.934
CIFAR-10	0.001	0.727	0.002	0.598	0.249	0.012	0.910
CIFAR-10	0.01	0.731	0.004	0.582	0.129	0.012	0.929
CIFAR-10	0.1	0.740	0.000	0.663	0.019	0.012	0.915
CIFAR-10-AT	0	0.545	0.320	0.627	0.255	0.020	0.937
CIFAR-10-AT	0.0001	0.539	0.316	0.652	0.273	0.019	0.946
CIFAR-10-AT	0.001	0.530	0.332	0.657	0.234	0.022	0.928
CIFAR-10-AT	0.01	0.559	0.325	0.691	0.117	0.019	0.927
CIFAR-10-AT	0.1	0.530	0.332	0.743	0.133	0.018	0.959

Near-zero PGD robust accuracy on CIFAR-10 and FashionMNIST is expected: these are undefended models with no adversarial training, evaluated at a standard ϵ. This column is a sanity check that the models are ordinary undefended classifiers, not the metric of interest — the per-example analysis below does not depend on it. The CIFAR-10-AT models trade clean accuracy for a large jump in PGD robust accuracy ($\approx 0 \rightarrow \approx 33\%$), confirming that adversarial training moved these models out of the uniformly-fragile regime of their undefended CIFAR-10 counterparts. Mean $D_M(x, \delta)$ on undefended CIFAR-10 is also strikingly flat across λ (0.012 throughout); this tracks mean $R(x)$,

which is similarly flat for the same models (0.26–0.32, essentially unchanged by λ) at a roughly constant mean ratio ($D_M/R \approx 0.039$–0.045 across λ) — consistent with these undefended models being uniformly fragile regardless of sparsity regularization (Section 5.4), rather than D_M and R being decoupled from one another.

5 Results

5.1 Sparsity does not correlate with per-example vulnerability; mask stability does

Table 1 reports, for every (λ, dataset) pair, the Pearson correlation of $S(x)$ and of $D_M(x, \delta)$ with $R(x)$ across the 500 evaluated examples. Figure 1 shows the same result as a bar chart, and Figure 2 shows the pooled scatter of $D_M(x, \delta)$ against $R(x)$.

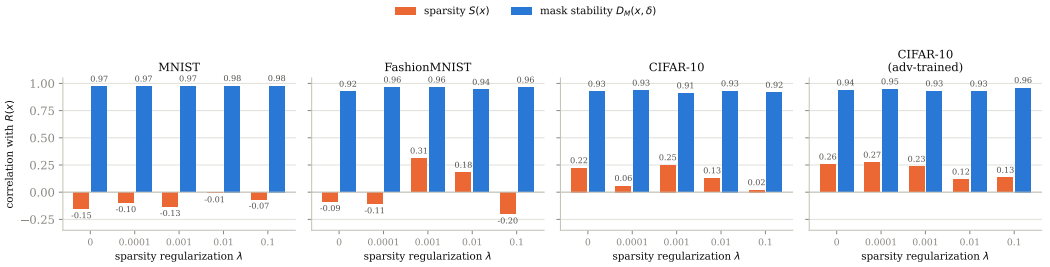


Figure 1: Per-example correlation with adversarial distance $R(x)$, at each sparsity-regularization strength λ. Sparsity $S(x)$ (orange) is a weak and sign-inconsistent predictor; mask stability $D_M(x, \delta)$ (blue) is strong and stable across every λ, on all three datasets and on the CIFAR-10-AT adversarially-trained control.

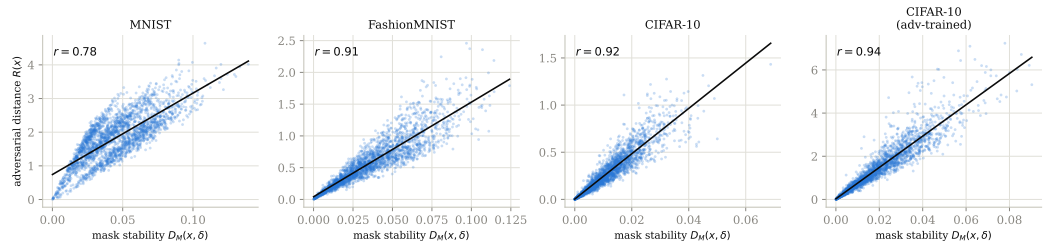


Figure 2: Pooled per-example mask instability $D_M(x, \delta)$ against adversarial distance $R(x)$ (all λ, $n = 2500$ per dataset/model). Examples whose active-unit set is stable under the DeepFool perturbation require a much larger perturbation to flip the prediction.

The pattern is uniform: $\mathrm{corr}(S, R)$ stays within $[-0.20, 0.31]$ and changes sign across λ within every dataset (most strikingly on FashionMNIST, where it swings from -0.20 to $+0.31$ across the sweep), while $\mathrm{corr}(D_M, R)$ stays in a narrow band (0.91–0.98) regardless of how sparse the model is on average, which dataset it was trained on, or whether the model was adversarially trained. Sparsity level is, at best, a weak and unreliable predictor of which individual examples are close to their decision boundary; mask stability is a strong and remarkably consistent one. We do not have a rigorous explanation for the sign flips themselves; a plausible speculation is that the ℓ_1 penalty changes *which* units are near zero, not just how many, and since $S(x)$ only counts the latter, the population-level relationship between sparsity and vulnerability could depend on incidental properties of which examples end up on which side of the many individual units' thresholds at a given λ — a hypothesis we have not tested.

5.2 Pooled statistics: rank correlation and bootstrap confidence intervals

Table 1 reports Pearson r separately for each of the 20 individual models (one per dataset $\times$ λ), each computed on its own $n = 500$ examples. To give a single, less linear-assumption-dependent summary per dataset, and to attach uncertainty to it, Table 2 reports both Pearson and Spearman

rank correlation pooled across all five λ values within a dataset ($n = 2500$), together with bootstrap 95% confidence intervals (5000 resamples) for the pooled Pearson correlations.

Table 2: Pooled correlations with $R(x)$ across all five λ within each dataset ($n = 2500$). Bootstrap 95% CIs (5000 resamples, resampling examples with replacement) are reported for the pooled Pearson correlations. Pooled values are lower than the per-λ range in Table 1 because pooling mixes λ groups with different mean S/D_M/R, the same effect noted for the margin correlations in Section 5.4.

Dataset	Pearson $r(S, R)$	Spearman $\rho(S, R)$	Pearson $r(D_M, R)$	Spearman $\rho(D_M, R)$
MNIST	-0.087 [-0.121, -0.051]	-0.069	0.776 [0.761, 0.789]	0.749
FashionMNIST	-0.011 [-0.046, $+0.025$]	$+0.019$	0.905 [0.898, 0.913]	0.937
CIFAR-10	$+0.002$ [-0.036, $+0.039$]	$+0.030$	0.921 [0.912, 0.929]	0.955
CIFAR-10-AT	$+0.100$ [$+0.058$, $+0.141$]	$+0.084$	0.935 [0.928, 0.943]	0.969

The pooled picture matches the per-λ picture: every S-R CI either straddles zero (FashionMNIST, CIFAR-10) or excludes it by only a small margin in an inconsistent direction across datasets (MNIST negative, CIFAR-10-AT positive), while every D_M-R CI is far from zero and the Spearman correlations, which do not assume a linear relationship, are as strong as or stronger than the Pearson ones. This rules out a purely linear-modeling artifact as the source of the D_M-R relationship, though it does not by itself address the endogeneity concern raised in Section 5.5. The 2500 pooled examples are drawn from only 5 related models (not 2500 independent replications), a subtlety we take up with a cluster bootstrap in Appendix A: the central D_M-R finding survives it, but it overturns MNIST's apparently significant S-R correlation above, which turns out to be an artifact of the naive per-example bootstrap.

5.3 Separating iteration-count effects from perturbation-length coupling

Because DeepFool accumulates δ across iterations, a natural concern is that both $R(x)$ and $D_M(x, \delta)$ simply scale with the number of iterations the attack needed, producing a spurious correlation between them. We check this by computing the partial correlation of D_M and R after regressing both on the DeepFool iteration count. Raw iteration counts correlate only weakly with R ($r = 0.15$–0.27 on MNIST, $r = 0.61$–0.66 on FashionMNIST, $r = 0.56$–0.67 on CIFAR-10, $r = 0.62$–0.72 on CIFAR-10-AT), and the partial correlation of D_M with R controlling for iteration count remains at 0.97–0.98 (MNIST), 0.87–0.96 (FashionMNIST), 0.85–0.92 (CIFAR-10), and 0.88–0.94 (CIFAR-10-AT) — essentially unchanged from the raw correlation in every case. The relationship is not an artifact of how long the attack ran. A stricter version of this check, using the attack's cumulative ℓ_2 path length rather than iteration count, turns out to be uninformative for a different reason — path length and $R(x)$ are themselves nearly collinear ($\mathrm{corr} > 0.99$ in every dataset) — which we detail in Appendix A.

5.4 The effect is not fully explained by logit margin

Logit margin is an obvious confound: a large-margin example is far from the decision boundary and might simply also be far from any activation boundary, making D_M a redundant restatement of margin. Table 3 reports $\mathrm{corr}(D_M, \text{margin})$ and the partial correlation of D_M with R after controlling for margin (per-λ ranges and a joint regression with sparsity and confidence are given in Appendix A).

Table 3: Mask stability vs. logit margin: raw correlation between D_M and margin, and the partial correlation of D_M with $R(x)$ after controlling for margin (pooled across λ).

Dataset	$\mathrm{corr}(D_M, \text{margin})$	partial $\mathrm{corr}(D_M, R \mid \text{margin})$
MNIST	0.575	0.637
FashionMNIST	0.783	0.736
CIFAR-10	0.881	0.610
CIFAR-10-AT	0.741	0.852

On MNIST and FashionMNIST, D_M and margin are correlated but the partial relationship between D_M and R stays close to the raw per-λ correlation. On undefended CIFAR-10, D_M and margin are more entangled, and controlling for margin reduces the partial correlation of D_M with R substantially more than on the other two datasets. This is consistent with the undefended CIFAR-10 models being uniformly fragile (mean DeepFool iterations $\approx$ 2.3–2.5, PGD robust accuracy

≈ 0), leaving less room for mask stability to add information beyond confidence and margin. The CIFAR-10-AT control directly tests this explanation: adversarial training raises PGD robust accuracy from ≈ 0 to $\approx 33\%$ (Table 1), moving the model out of the uniformly-fragile regime, and the partial correlation of D_M with R controlling for margin rises accordingly — essentially back to the MNIST/FashionMNIST range despite CIFAR-10-AT still being a harder dataset than either. This supports reading the undefended-CIFAR-10 margin entanglement as a symptom of that specific model's fragility rather than a general limitation of D_M as a predictor.

5.5 A CAVEAT: D_M IS MEASURED AT THE POINT THE ATTACK FINDS

The checks in Sections 5.4 and 5.2 rule out two specific alternative explanations (attack bookkeeping, logit margin), but they do not rule out a more basic concern with how $D_M(x, \delta)$ is defined. δ here is not an arbitrary perturbation: it is *the perturbation DeepFool used to reach the decision boundary*, so $D_M(x, \delta)$ is the mask change accumulated over a path whose length is, by definition, $R(x) = \|\delta\|_2$. A perturbation that must travel farther to flip the label has mechanically more distance over which to cross activation boundaries, so part of the D_M–R relationship could be a near-tautological consequence of *how* D_M is measured rather than evidence that mask stability is an independent predictor of vulnerability. Two examples that need the same number of DeepFool iterations (ruled out in Section 5.3) can still have perturbations of very different ℓ_2 norm, and nothing in that check controls for perturbation length specifically (Appendix A).

Resolving this requires decoupling D_M from R by measuring mask sensitivity *independently of the adversarial search that produces* $R(x)$ — for instance, (i) sampling mask changes under many random directions at several *fixed* radii ϵ and asking whether that fixed-radius sensitivity predicts $R(x)$, or (ii) estimating, from the clean input alone, the distance to the nearest activation-region boundary (e.g. via the per-unit linearization $d_i(x) \approx |z_i(x)|/\|\nabla_x z_i(x)\|_2$ for pre-activation z_i) and correlating that purely-clean-input quantity against $R(x)$. Section 5.6 runs both experiments directly on the same 20 trained checkpoints.

5.6 TESTING THE CAVEAT DIRECTLY: CLEAN-INPUT AND FIXED-RADIUS ALTERNATIVES

Using the same checkpoints and the same 500 per-model examples (selected identically to Section 4), we compute two quantities that never reference δ or $R(x)$:

- **Fixed-radius random-direction sensitivity $D_M^\epsilon(x)$.** For a radius ϵ fixed *per dataset* (not per example, so it cannot inherit any individual example's $R(x)$), chosen from that dataset's overall $R(x)$ scale in Table 1 (MNIST: $\epsilon \in \{0.05, 0.1, 0.25, 0.5\}$; FashionMNIST: $\epsilon \in \{0.01, 0.025, 0.05, 0.1\}$; CIFAR-10: $\epsilon \in \{0.002, 0.005, 0.01, 0.02\}$; CIFAR-10-AT: $\epsilon \in \{0.01, 0.025, 0.05, 0.1\}$), we sample $K = 16$ random unit-norm directions u_k and average the pooled mask-Hamming-distance at $x + \epsilon u_k$ over the K directions.
- **Clean-input boundary distance $A(x)$.** For a pre-activation $z(x)$ in any layer, unit i's linearized distance to its own activation boundary is $d_i(x) = |z_i(x)|/\|\nabla_x z_i(x)\|_2$. For the 128-unit FC layer we take the exact minimum $A_{FC}(x) = \min_i d_i(x)$ over all 128 units (128 backward passes, batched over all 500 examples; this architecture has no batch normalization, so batching does not affect per-example gradients). Each convolutional layer has tens of thousands of units, too many for an exact per-unit backward pass at this pilot's scale, so we approximate each conv layer's minimum with a subsample of units and report $A_{\text{all}}(x)$, the minimum across all four layers, as an estimate of the true global boundary distance; Appendix B tests this approximation's sensitivity to the sampling scheme and budget in detail.

The result is unambiguous: decoupling D_M from $R(x)$ collapses the correlation. $A_{FC}(x)$ is weakly but consistently positively correlated with $R(x)$ (pooled $r = 0.10$–0.17; bootstrap 95% CIs exclude 0 in every dataset), the correctly-signed direction — examples farther from an FC-layer activation boundary need a larger perturbation — but nowhere near the strength of the attack-point measurement. $D_M^\epsilon(x)$ is noisier: on MNIST every fixed radius gives a correlation statistically indistinguishable from 0; on FashionMNIST the sign is inconsistent across radii and only the smallest ϵ gives a CI excluding 0; on CIFAR-10 and CIFAR-10-AT the correlation is consistently negative (the expected sign: more sensitivity, smaller R) and grows monotonically in magnitude with ϵ, from $r \approx -0.12$

Table 4: Decoupling D_M from $R(x)$: pooled Pearson correlation with $R(x)$ ($n = 2500$) for the attack-point $D_M(x, \delta)$ (reproduced from Table 2 for reference) versus two quantities that never use δ: the clean-input FC-layer boundary distance $A_{FC}(x)$, and the fixed-radius sensitivity $D_M^{\epsilon}(x)$ at the four per-dataset radii listed above, smallest (ϵ_1) to largest (ϵ_4).

Dataset	$D_M(x,\delta)$	$A_{FC}(x)$	$D_M^{\epsilon_1}(x)$	$D_M^{\epsilon_2}(x)$	$D_M^{\epsilon_3}(x)$	$D_M^{\epsilon_4}(x)$
MNIST	0.776	0.165	−0.030	−0.029	−0.021	−0.012
FashionMNIST	0.905	0.139	−0.183	−0.043	+0.002	+0.026
CIFAR-10	0.921	0.139	−0.116	−0.156	−0.197	−0.215
CIFAR-10-AT	0.935	0.099	−0.158	−0.188	−0.216	−0.231

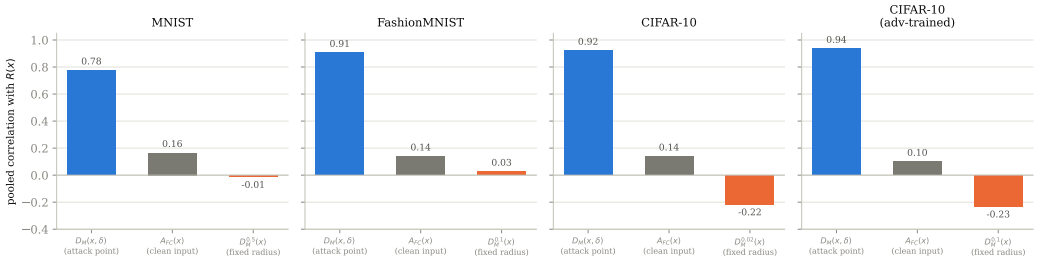


Figure 3: Pooled correlation with $R(x)$ for the attack-point $D_M(x, \delta)$ versus the two decoupled, δ-free alternatives (clean-input $A_{FC}(x)$ and fixed-radius $D_M^{\epsilon}(x)$ at the largest radius tested per dataset). Decoupling from the adversarial search collapses the correlation from $r > 0.9$ to $|r| \lesssim 0.2$–0.3.

to $r \approx -0.22$, all CIs excluding 0, but still far short of the $r > 0.9$ seen with the attack-point measurement.

We read this as confirming, not refuting, the concern raised above: most of the strength of the original $D_M(x, \delta)$–$R(x)$ correlation appears attributable to the shared measurement point rather than to an independent relationship between activation-region geometry and adversarial distance of comparable magnitude. A smaller residual relationship does appear to survive decoupling: $A_{FC}(x)$ is positive with a CI excluding 0 in every dataset, and $D_M^{\epsilon}(x)$ on CIFAR-10/CIFAR-10-AT strengthens with ϵ in the expected direction, so the qualitative claim that activation-region geometry carries *some* information about adversarial vulnerability beyond what sparsity offers is not eliminated. We call this residual, not necessarily causal or geometrically fundamental: it is consistent with a real contribution from activation-boundary geometry, but also with $A_{FC}(x)$ and $R(x)$ both reflecting shared latent factors (e.g. margin, confidence, or input complexity) that we have only partially controlled for (Section 5.4). Its practical size, once decoupled from $R(x)$ itself, is closer to $|r| \approx 0.1$–0.3 than to $r \approx 0.9$, and the "strong predictor" framing of the earlier sections should be read as characterizing the entangled measurement specifically, not the underlying geometric relationship.

We ran three further robustness checks on this decoupling — replacing the isotropic random directions in $D_M^{\epsilon}(x)$ with a single gradient-aligned direction; testing whether extending $A(x)$ to the convolutional layers (via uniform-random sampling, importance-guided sampling, layer normalization, and different subsample budgets) helps or hurts; and verifying the results are not an artifact of the 300-unit subsample budget — all detailed in Appendix B. In short: gradient-aligned directions meaningfully strengthen the CIFAR-10 signal (up to $\sim 25\%$ larger in magnitude) but do not rescue MNIST; layer-normalizing the combined boundary-distance aggregation helps consistently, while importance-sampling and budget choice help on some datasets and not others. None of these variants overturns the qualitative collapse in Table 4.

5.7 A CROSSING-DENSITY DECOMPOSITION: IS D_M JUST DISTANCE TRAVELED?

Section 5.5 argued informally that part of the D_M–R relationship could be a near-tautological consequence of how D_M is measured. We can state a sharper version of that concern. Suppose activation boundaries are crossed along the DeepFool path at some local rate $\lambda(x)$ (boundary crossings per unit ℓ_2 distance, which may itself vary across examples). To first order, the number of mask flips accumulated over a perturbation of length $R(x)$ is then $N_{\text{flips}}(x) \approx \lambda(x)\, R(x)$, so that, since $D_M(x, \delta) = N_{\text{flips}}(x)/N$,

$$D_M(x, \delta) \;\approx\; \frac{\lambda(x)}{N}\, R(x).$$

If $\lambda(x)$ were roughly constant across examples, $D_M(x, \delta)$ would be close to *proportional* to $R(x)$ — a mechanical consequence of counting more crossings over a longer path, carrying essentially no information about x beyond $R(x)$ itself. If instead $\lambda(x)$ varies systematically with vulnerability, more of that correlation reflects real per-example geometry: vulnerable examples would not just be closer to the boundary, they would also cross activation boundaries at a different *rate* along the way. This is directly testable with data we already have: define the crossing density $C(x) = D_M(x, \delta)/R(x)$, an estimate of $\lambda(x)/N$, and correlate it against $R(x)$.

Table 5: Crossing-density decomposition. $\mathrm{corr}(D_M, R)$ is reproduced from Table 2 for reference. $\mathrm{corr}(C, R)$ is the pooled correlation ($n = 2500$) of $C(x) = D_M(x, \delta)/R(x)$ with $R(x)$, with bootstrap 95% CI; "per-λ range" gives the range of $\mathrm{corr}(C, R)$ computed separately within each of the five λ models before pooling.

Dataset	$\mathrm{corr}(D_M, R)$	pooled $\mathrm{corr}(C, R)$	per-λ range of $\mathrm{corr}(C, R)$
MNIST	0.776	-0.042 [-0.089, $+0.003$]	$[-0.51, +0.51]$
FashionMNIST	0.905	-0.109 [-0.146, -0.072]	$[-0.33, +0.27]$
CIFAR-10	0.921	-0.346 [-0.380, -0.311]	$[-0.41, -0.29]$
CIFAR-10-AT	0.935	-0.219 [-0.257, -0.180]	$[-0.27, -0.20]$

The result differs by dataset, and is more informative than either would be alone. On MNIST, $C(x)$ is essentially uncorrelated with $R(x)$ once pooled ($r = -0.042$, CI including 0) — consistent with a largely mechanical, roughly-constant-rate relationship there, though the per-λ correlations swing wildly in sign ($+0.51$ to -0.51), mirroring the sign-inconsistency we saw for sparsity itself (Section 5.1), which we read as λ-group confounding of the pooled estimate rather than clean evidence of zero information on MNIST specifically. On FashionMNIST, CIFAR-10, and CIFAR-10-AT, $C(x)$ is negatively and consistently correlated with $R(x)$ ($r = -0.11$, -0.35, and -0.22 respectively, all CIs excluding 0): more vulnerable examples (smaller R) cross activation boundaries at a higher rate per unit distance, not merely over a shorter distance. On CIFAR-10 in particular, this crossing-density correlation (-0.346) is larger in magnitude than either decoupled measure from Section 5.6 (A_{FC}: $+0.139$; D_M^ϵ: -0.116 to -0.215).

This sharpens, rather than resolves, the picture from Section 5.6: the entangled correlation is not purely a tautological consequence of counting more crossings over a longer path on three of our four datasets, though it may be closer to that on MNIST specifically. $C(x)$ still uses δ (it is a normalized version of the entangled measurement, not a δ-free quantity like $A_{FC}(x)$ or $D_M^\epsilon(x)$), so it does not settle the circularity question on its own; we see it as a useful complement to Section 5.6's δ-free checks, not a replacement for them. We also checked whether $C(x)$ computed only at the endpoint is representative of the whole attack path, by re-running DeepFool while logging the local crossing rate at every intermediate iteration: it shows no systematic trend along the path, and averaging it over the whole path gives essentially the same correlation with $R(x)$ as the endpoint alone (Appendix B).

5.8 Cross-checking with a different attack: CW-L2

Every result so far uses DeepFool to approximate $R(x)$ and δ. Since DeepFool is a first-order local method, a natural concern is that the collapse in Section 5.6 is itself specific to DeepFool rather than a property of $R(x)$ in general. We spot-check this on a 200-example subset for the $\lambda = 0$ model on each dataset, using a lightweight batched CW-L2 attack (a fixed grid over the confidence constant c rather than a full adaptive binary search; Appendix C), giving an independent δ_{cw}, $R_{\mathrm{cw}}(x) = \|\delta_{\mathrm{cw}}\|_2$, and attack-point $D_M(x, \delta_{\mathrm{cw}})$.

Table 6: Cross-checking the circularity finding with CW-L2 instead of DeepFool, on 200 examples per dataset ($\lambda = 0$ only). "entangled" is $\mathrm{corr}(D_M(x, \delta), R(x))$ for each attack; "decoupled" is $\mathrm{corr}(A_{FC}(x), R(x))$, using the same $A_{FC}(x)$ values against each attack's own $R(x)$. CW succ. is the fraction of the 200 examples where the grid attack found a successful perturbation.

Dataset	CW succ.	$\mathrm{corr}(R_{\mathrm{df}}, R_{\mathrm{cw}})$	entangled (DF)	entangled (CW)	decoupled A_{FC} (DF)	decoupled A_{FC} (CW)
MNIST	0.955	0.837	0.972	0.977	0.202	0.174
FashionMNIST	1.000	0.863	0.912	0.846	0.050	0.008
CIFAR-10	1.000	0.846	0.925	0.908	0.123	0.107
CIFAR-10-AT	0.975	0.951	0.935	0.933	0.292	0.238

Three things support reading the circularity finding as attack-agnostic rather than DeepFool-specific. First, DeepFool and CW-L2 broadly agree on which examples are more or less vulnerable:

$\mathrm{corr}(R_{\mathrm{df}}, R_{\mathrm{cw}}) = 0.84$–$0.95$ across all four datasets, even though the two methods optimize differently. Second, the entangled correlation is strong under *both* attacks and at similar magnitude: 0.85–0.98 regardless of which attack supplies δ and $R(x)$. Third, the decoupled $A_{FC}(x)$ correlation stays in the same modest range under both: 0.05–0.29 (DeepFool) versus 0.01–0.24 (CW-L2) on this subset — consistently weaker than the entangled measurement, under either attack. This is only a spot check — one λ per dataset, 200 examples, and a simplified grid-search CW-L2 rather than a full adaptive attack — but it weighs against the collapse in Section 5.6 being an artifact of DeepFool's specific linearization.

6 Discussion

Six findings are worth separating. First, sparsity alone is not a reliable per-example correlate of adversarial distance — small and sign-inconsistent across datasets and regularization strengths; a claim about *within-model, per-example* variation, not in tension with model-level findings that mean sparsity and robustness co-vary across trained models (Section 1). Second, mask stability computed *at the adversarial point* is strong and strikingly consistent, robust to bookkeeping, margin, and rank correlation (Section 5.2). Third, the CIFAR-10-AT control suggests margin's apparent explanation of D_M's signal on undefended CIFAR-10 was a symptom of uniform fragility, not a real limit: robustify the model and the signal returns. Fourth, and this reframes the first three: D_M is measured at the adversarial point itself, so it is coupled by construction to $R(x)$, and neither bookkeeping nor margin rules this out. Testing this directly (Section 5.6) with two δ-free alternatives, the correlation collapses from $r > 0.9$ to $|r| \approx 0.1$–0.3. Fifth, that collapse is itself informative: the decoupled correlations are small but often non-zero and correctly signed — a bootstrap CI excluding 0 in 14 of 20 combinations (Appendix D) — so the geometry sketched in Section 1 retains some support, though not the earlier "strong" framing. Sixth, the crossing-density decomposition (Section 5.7) sharpens this: $C(x)$ is uncorrelated with $R(x)$ on MNIST (largely mechanical), but negatively and consistently correlated on FashionMNIST, CIFAR-10, and CIFAR-10-AT — vulnerable examples cross boundaries at a systematically higher rate, not only over a shorter distance.

We think this distinction — a real but modest effect, reported at 3–5$\times$ its decoupled size — is the paper's most important methodological point: any metric defined at the point an attack lands risks this inflation, and checks like iteration count or margin do not catch it alone. Our results are consistent with a weaker version of the reframing suggested by Jiang et al. (2023): activation-region geometry carries *some* real information beyond sparsity, but is a modest contributor, not a dominant one. Appendix D sketches a heuristic and toy mixture model for why $A(x)$ would be expected to be real but weak, then tests its specific numerical prediction directly and finds it does not hold as stated — a useful negative result about the *mechanism*, even though the underlying weak-but-real correlation still stands. Appendix B details three further checks that make us more confident the collapse reflects something real about D_M and $R(x)$ rather than an artifact of our specific test setup.

Limitations. These are pilot-scale results motivating a larger, better-controlled study, not a settled finding. The most important qualification, stated above, is that $D_M(x, \delta)$ is coupled to $R(x)$ by construction. Beyond this: a single small, mostly-undefended architecture; $R(x)$ spot-checked against only one alternative attack; a sample not fully independent across λ; only one sparsity-inducing mechanism; and decoupled measures with their own design choices whose sensitivity we characterize but do not fully resolve. Appendix E expands each of these seven points.

7 Conclusion

Per-example activation sparsity is a weak, sign-inconsistent correlate of adversarial vulnerability, while activation-mask stability computed *at* the adversarial perturbation is strong and robust to every check we tried. But that measurement is coupled by construction to the very distance it correlates with, and decoupling the two collapses the correlation from $r > 0.9$ to $|r| \approx 0.1$–0.35. An attack-point measurement of this form can substantially inflate the apparent strength of a mask-stability/vulnerability relationship, as it does in our own results; a smaller, statistically distinguishable effect survives every way we decoupled it, and reproduces under an independent attack and a cluster bootstrap (Appendix A); we make no claim about other papers' specific measurement definitions. The next step is no longer *whether* $D_M(x, \delta)$ predicts $R(x)$, but characterizing the smaller, decoupled effect and scaling the architecture past this toy 4-layer CNN — a case study of one mechanism, not a general verdict on activation-sparsity or mask-stability research.

Reproducibility Statement

All architecture, training, and sparsity-sweep details needed to reproduce our models are given in Section 4 (the fixed 4-layer CNN, optimizer, λ grid, and adversarial-training recipe); attack settings (DeepFool, fixed-radius ϵ grids, CW-L2) are given in Section 5 and Appendix C; the per-example selection rule, bootstrap procedure, and layer-wise breakdown are given in Appendix A. All 20 models use the same fixed architecture and optimizer settings, one seed per configuration (Appendix E). Code, trained checkpoints, and per-example result files sufficient to regenerate every table and figure will be released publicly upon publication.

AI Use Statement

An AI coding and writing assistant was used throughout this project: implementing and debugging experiment and analysis scripts, running the reported computations, and drafting and revising manuscript text across multiple rounds, including in response to reviewer-style critiques of earlier drafts. All experimental design choices, interpretation of results, and decisions about what to report were made by the authors, who reviewed all AI-assisted code and text before inclusion and take full responsibility for the correctness and integrity of the reported findings.

References

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53, 2014.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse DNNs with improved adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yiding Jiang et al. On explaining neural network robustness with activation path. In *International Conference on Learning Representations (ICLR)*, 2023.

Sarada Krithivasan et al. Sparsity turns adversarial: Energy and latency attacks on deep neural networks. *arXiv preprint arXiv:2006.08020*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ori Peng et al. A theoretical explanation of activation sparsity through flat minima and adversarial robustness. *arXiv preprint arXiv:2309.03004*, 2023.

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 (ICLR)*, 2018.

A ADDITIONAL PER-EXAMPLE DIAGNOSTICS

This appendix collects diagnostics that refine or extend checks in the main text but are not needed to follow the core argument.

A.1 LAYER-WISE BREAKDOWN OF MASK INSTABILITY

$D_M(x, \delta)$ as defined in the main text pools activation changes across all four ReLU layers (three conv layers and one FC layer). Because a natural question is whether instability in early or late layers carries more of the signal, we also recorded the per-layer instability $D_M^{(l)}(x, \delta)$ separately for each layer l and correlate each against $R(x)$ (pooled across λ); Figure 4 shows the result.

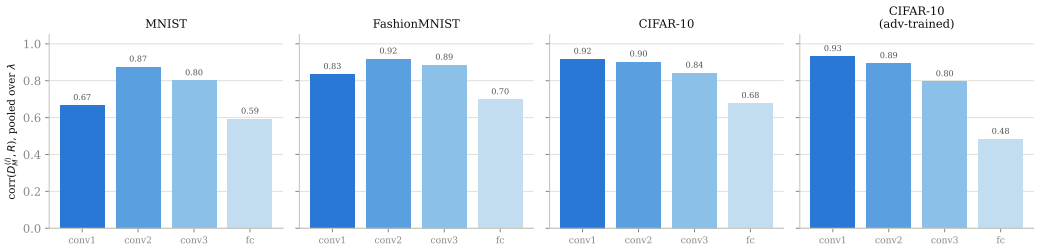


Figure 4: Correlation of per-layer mask instability $D_M^{(l)}(x, \delta)$ with $R(x)$, pooled across λ ($n = 2500$ per dataset). conv1–conv3 are the three convolutional ReLU layers; fc is the 128-unit fully-connected layer.

The consistent pattern across all four datasets/models is that the convolutional layers (conv1–conv3) carry a stronger signal than the final 128-unit fully-connected layer, whose correlation with $R(x)$ is consistently the lowest of the four (e.g. 0.59 on MNIST, 0.70 on FashionMNIST, 0.68 on CIFAR-10, 0.48 on CIFAR-10-AT, versus 0.80–0.93 for the strongest conv layer in each case). One reading is that instability in the earlier, higher-dimensional conv layers is a better proxy for how much the input-to-logit computation changed than instability in the narrow 128-unit FC layer, where a single flipped unit is a much coarser event. We report this as a descriptive pattern in this single architecture, not a general claim about layer depth; it would need to be checked in networks with more than four ReLU layers before drawing conclusions about "early vs. late" more broadly.

A.2 CUMULATIVE PATH LENGTH: A SHARPER BOOKKEEPING CHECK

Iteration count (Section 5.3) says nothing about how far each step moved, so a stricter version of the bookkeeping concern asks about the attack's cumulative ℓ_2 path length — the sum of per-iteration step norms — rather than how many steps it took. We recomputed DeepFool on all 20 models while also recording this quantity. The result is that, in this setting, path length and $R(x)$ are nearly the same quantity: $\mathrm{corr}(\text{path length}, R) > 0.99$ in every dataset and every λ, with path length exceeding the net displacement $R(x)$ by only 2–8% on average (mean ratio 1.02 on CIFAR-10 to 1.08 on FashionMNIST; up to $1.5\times$ in the most curved individual cases). DeepFool's trajectory is close to a straight line to the decision boundary here, not a meandering search that happens to land at distance $R(x)$.

This changes how to read a partial-correlation check built on path length. Because path length and $R(x)$ are themselves almost collinear, regressing D_M and R on path length and correlating the residuals is close to asking what remains of $\mathrm{corr}(D_M, R)$ after removing nearly all of R's own variance — a numerically fragile quantity, not a clean bookkeeping control. Consistent with that, the resulting partial correlations are unstable across λ (MNIST: 0.11–0.33; FashionMNIST: -0.01–0.23; CIFAR-10: -0.17–0.21; CIFAR-10-AT: 0.21–0.55), unlike the stable 0.85–0.98 we get controlling for iteration count. We do not read this instability as evidence against D_M; we read it as evidence that path length is not a useful *independent* control here, precisely because it is so tightly coupled to $R(x)$ itself. If anything, this reinforces rather than undermines the circularity concern developed in Section 5.5: $D_M(x, \delta)$ is computed at the end of a path whose length is, empirically, almost exactly $R(x)$, so "mask change accumulated over the attack's path" and "mask change measured at distance

$R(x)$" are nearly the same statement in this setting, not two independent facts that happen to agree. The claim that the entangled correlation "is not an artifact of attack bookkeeping" should therefore be read as resting on the iteration-count check specifically, not on path length, which is too collinear with $R(x)$ to serve as an independent control.

A.3 Extended margin analysis

Table 7 gives the per-λ ranges underlying the pooled correlations in Table 3.

Table 7: Per-λ ranges for Table 3: raw $\mathrm{corr}(D_M, \text{margin})$ and partial $\mathrm{corr}(D_M, R \mid \text{margin})$, computed separately within each of the five λ models before pooling.

Dataset	per-λ $\mathrm{corr}(D_M, \text{margin})$	per-λ partial $\mathrm{corr}(D_M, R \mid \text{margin})$
MNIST	0.69–0.83	0.92–0.96
FashionMNIST	0.77–0.90	0.70–0.91
CIFAR-10	0.87–0.91	0.55–0.67
CIFAR-10-AT	0.71–0.79	0.84–0.92

The pooled values in Table 3 are lower than these per-λ ranges (most visibly for CIFAR-10, least for MNIST) because pooling mixes λ groups with different mean margins; CIFAR-10 is least affected because λ has little effect on that dataset's margins.

A standardized linear regression of $R(x)$ on $S(x)$, $D_M(x, \delta)$, confidence, and margin jointly (Table 8) confirms the main-text reading: D_M carries the largest standardized coefficient in every case, sparsity's coefficient is small or sign-flips relative to its raw correlation (a sign of confounding with λ-group membership rather than a genuine effect), and confidence contributes little once margin is included except on CIFAR-10-AT, where it picks up part of what margin explains elsewhere.

Table 8: Standardized OLS regression $R(x) \sim S(x) + D_M(x, \delta) + \text{confidence}(x) + \text{margin}(x)$, pooled across λ ($n = 2500$ per dataset).

Dataset	β_S	β_{D_M}	β_{conf}	β_{margin}	R^2
MNIST	$+0.319$	$+0.561$	$+0.020$	$+0.551$	0.832
FashionMNIST	-0.046	$+0.682$	-0.024	$+0.306$	0.859
CIFAR-10	$+0.017$	$+0.541$	-0.010	$+0.438$	0.891
CIFAR-10-AT	$+0.103$	$+0.798$	$+0.142$	$+0.050$	0.903

A.4 Cluster bootstrap: accounting for 5 models, not 2500 independent examples

The bootstrap CIs in Table 2 resample individual examples, which is appropriate for uncertainty *within* a model but treats the 2500 pooled examples as exchangeable observations. They are not: the five λ values within a dataset are five related models sharing an architecture and training procedure, not five independent replications, so the true unit of replication here is closer to the model than the example. Table 9 repeats the pooled correlations with a cluster bootstrap instead: each of the 5000 resamples draws 5 λ-models with replacement (keeping all 500 examples within each drawn model intact) rather than resampling examples directly.

As expected, cluster CIs are wider — sometimes substantially so, since there are only 5 clusters to resample from. This matters for one specific claim: MNIST's S-R correlation, which looked significantly non-zero under the naive per-example bootstrap ($[-0.121, -0.051]$), is no longer distinguishable from 0 under the cluster bootstrap ($[-0.169, +0.029]$) — the naive bootstrap had overstated precision there by treating 500 examples from one model as if they added four more independent data points beyond the first. CIFAR-10-AT's small positive S-R correlation survives the cluster bootstrap, narrowly ($[+0.078, +0.224]$), so unlike the other three datasets we cannot rule out a small non-null sparsity effect specific to that adversarially-trained model. What does not change is

Table 9: Cluster (per-model) bootstrap 95% CIs versus the naive per-example bootstrap CIs from Table 2, for the same pooled correlations ($n = 2500$, 5 clusters). The cluster bootstrap accounts for the 5 λ models being the actual unit of replication.

Dataset	$r(S, R)$	example-CI	cluster-CI	$r(D_M, R)$	example-CI	cluster-CI
MNIST	-0.087	[-0.121,-0.051]	[-0.169,$+0.029$]	0.776	[0.761,0.789]	[0.682,0.929]
FashionMNIST	-0.011	[-0.046,$+0.025$]	[-0.083,$+0.183$]	0.905	[0.898,0.913]	[0.894,0.948]
CIFAR-10	$+0.002$	[-0.036,$+0.039$]	[-0.073,$+0.180$]	0.921	[0.912,0.929]	[0.912,0.928]
CIFAR-10-AT	$+0.100$	[$+0.058$,$+0.141$]	[$+0.078$,$+0.224$]	0.935	[0.928,0.943]	[0.928,0.948]

the central positive finding: every D_M-R cluster CI remains clearly bounded away from 0, including MNIST's widest case ($[0.682, 0.929]$). The strong attack-point correlation is not an artifact of treating 2500 correlated examples as if they were independent; the weak sparsity correlation, on the other hand, was partly an artifact of exactly that in at least one dataset.

A.5 Example-set overlap across λ

"Per model" in the main-text description of the 500 evaluated examples is deliberate, not incidental: we iterate the test set in a fixed order and keep each model's own first 500 correctly-classified images, so the specific images differ across the five λ models within a dataset whenever their clean accuracy differs. The overlap is large but not complete: across the five MNIST models (clean accuracy 99%+) the images agree on 490 of 500 indices; across the five CIFAR-10 models ($\approx 72\%$ clean accuracy) they agree on only 336 of 500, and CIFAR-10-AT on 354 of 500. Pooling across λ (Table 2 onward) therefore pools over partially, not fully, overlapping example sets, which is a further reason to read the five λ models as related but non-identical samples rather than five draws from one fixed population of examples.

B Robustness of the decoupled measures

B.1 Crossing density along the attack path, not just at the endpoint

The crossing density $C(x) = D_M(x, \delta)/R(x)$ in Section 5.7 is computed only at the attack's endpoint. We check whether this is representative of the whole path by re-running DeepFool on all 20 models with instrumentation that logs, at every iteration k, the cumulative displacement so far and the pooled mask-Hamming-distance from the clean mask at that intermediate point, giving a local crossing-rate estimate $\lambda_k(x) = D_M^{(k)}(x)/\|\delta_k\|_2$ at every step along the path (the last logged step coincides with the endpoint $C(x)$ already reported). This requires no new training, only re-running the attack with extra logging; DeepFool's unbatched, sequential nature makes this slower than the batched experiments elsewhere in this paper, so we report it as a standalone check rather than folding it into every other table.

Table 10: Does crossing density change systematically along the path, and does sampling it only at the endpoint matter? "≥ 3 steps" is the fraction of examples whose path has at least 3 logged points (needed for a meaningful within-path comparison). "Median ratio" is $\lambda_{\text{last}}/\lambda_{\text{first}}$ among those examples (≈ 1 indicates no systematic trend). The two rightmost columns are pooled correlations with $R(x)$ ($n = 2500$) using only the endpoint step versus the mean of λ_k over the whole logged path, from this same re-instrumented run (so compare the two columns to each other, not to Table 5, which used the original, separately-run DeepFool pass).

Dataset	≥ 3 steps	median ratio	$\text{corr}(\lambda_{\text{endpoint}}, R)$	$\text{corr}(\lambda_{\text{path mean}}, R)$
MNIST	0.982	1.001	-0.042	-0.050
FashionMNIST	0.992	0.990	-0.114	-0.105
CIFAR-10	0.378	0.998	-0.346	-0.346
CIFAR-10-AT	0.416	1.010	-0.219	-0.232

Two things follow. First, crossing density shows no systematic trend along the path: the median last-to-first ratio is within 1% of exactly 1 in every dataset except CIFAR-10-AT (1.010), and the fraction

of examples where it increases rather than decreases is close to 50% everywhere (MNIST 51%, FashionMNIST 41%, CIFAR-10 46%) except CIFAR-10-AT (69%, a mild tendency to increase). This is consistent with $\lambda(x)$ being reasonably well approximated as a single per-example constant over the (typically short) DeepFool path, rather than the endpoint being a special, unrepresentative point on it. Second, and consequently, using the path-averaged λ instead of the endpoint-only value barely changes anything: the two rightmost columns agree to within 0.01 in three datasets (exactly, on CIFAR-10) and 0.013 on the fourth (CIFAR-10-AT). The circularity concern's response to this measure is therefore not sensitive to sampling only the endpoint — there was nothing special hiding earlier in the path that the endpoint-only version was missing.

CIFAR-10 and CIFAR-10-AT have far fewer multi-step examples (38–42% with ≥ 3 steps, versus $\geq$ 98% for MNIST/FashionMNIST): the pilot's undefended and lightly-adversarially-trained CIFAR-10 models are uniformly fragile enough that DeepFool typically succeeds in 1–2 steps, leaving little of a "path" to sample beyond the endpoint for most examples there.

B.2 Direction choice: gradient-aligned vs. isotropic random

The $D^{\epsilon}_{M}(x)$ in Section 5.6 averages over $K{=}16$ isotropic random directions, which in a high-dimensional input space rarely align with any structurally relevant direction (e.g. the local decision-boundary normal). We test whether a direction chosen to be locally relevant, but still independent of δ and $R(x)$, reveals more signal: for each clean example we compute the single unit-norm direction $u = \nabla_x \mathrm{CE}(f(x), y)/\|\nabla_x \mathrm{CE}(f(x), y)\|_2$ (the ascending-loss direction at the true label) and evaluate mask sensitivity at $x + \epsilon u$ for the same per-dataset radii, replacing K random directions with this one gradient-aligned direction.

Table 11: Fixed-radius mask sensitivity using isotropic-random directions (reproduced from Table 4) versus a single gradient-aligned direction, pooled Pearson correlation with $R(x)$ ($n = 2500$) at the smallest (ϵ_1) and largest (ϵ_4) radius per dataset.

Dataset	random ϵ_1	grad. ϵ_1	random ϵ_4	grad. ϵ_4
MNIST	-0.030	-0.015	-0.012	-0.021
FashionMNIST	-0.183	-0.179	$+0.026$	$+0.019$
CIFAR-10	-0.116	-0.130	-0.215	-0.272
CIFAR-10-AT	-0.158	-0.121	-0.231	-0.245

The gradient-aligned direction is not uniformly stronger, but it is consistently at least as strong on three of four datasets, and meaningfully stronger on one. On CIFAR-10, the gradient-aligned correlation exceeds the isotropic-random one at every radius we tested (e.g. at the largest radius, $-0.215 \rightarrow -0.272$, roughly a 25% increase in magnitude), consistent with the local loss gradient pointing toward directions that are more informative about nearby activation-boundary crossings than a random direction. On FashionMNIST and CIFAR-10-AT the two are close, with no consistent winner across radii. On MNIST both remain near-null, so aligning the direction does not rescue the signal there — consistent with our reading in Appendix D that MNIST's activation regions are simply large enough, relative to $R_{\mathrm{intra}}(x)$, that $A(x)$ rarely binds regardless of which direction is probed.

B.3 Extending clean-input boundary distance to convolutional layers

$A_{FC}(x)$ only queries the final hidden layer, so a natural question is whether the three convolutional layers carry additional signal that a FC-only measure misses, and whether a *uniform-random* 300-unit subset per layer is the right way to query them. We test two sampling variants: the uniform-random subset described in the main text, and an importance-sampled variant where each example nominates its own 300 candidate units per conv layer — those with the smallest $|z_i(x)|$, a cheap (no backward pass needed), necessary-but-not-sufficient condition for a unit to be close to its boundary — rather than one fixed random subset shared by the whole batch. Table 12 reports both.

The result is mixed rather than a clean win. Importance sampling clearly helps on FashionMNIST: A_{all} rises from 0.029 (random) to 0.144 (importance), now matching $A_{FC}(x)$ alone (0.139), and its conv1 distance alone (0.166) becomes the single strongest clean-input correlate we report for that

Table 12: Pooled Pearson correlation of the per-layer clean-input boundary distance with $R(x)$ ($n = 2500$), for both the uniform-random and the importance-sampled 300-unit conv-layer subsets. fc is exact (128 units); all/all-imp are the minimum across all four layers under each sampling scheme.

	uniform-random			importance-sampled					
Dataset	conv1	conv2	conv3	conv1	conv2	conv3	$A_{FC}(x)$	all (rand)	all (imp)
MNIST	-0.013	-0.054	$+0.054$	$+0.036$	-0.017	$+0.061$	$+0.165$	$+0.014$	$+0.001$
FashionMNIST	$+0.092$	-0.051	-0.007	$+0.166$	$+0.064$	$+0.018$	$+0.139$	$+0.029$	$+0.144$
CIFAR-10	$+0.019$	$+0.104$	$+0.096$	$+0.079$	$+0.055$	$+0.089$	$+0.139$	$+0.056$	$+0.072$
CIFAR-10-AT	$+0.090$	$+0.071$	$+0.087$	$+0.107$	$+0.095$	$+0.063$	$+0.099$	$+0.106$	$+0.105$

dataset. On CIFAR-10 it gives a smaller but real improvement ($0.056 \rightarrow 0.072$); on CIFAR-10-AT the two are indistinguishable (0.106 vs. 0.105). But on MNIST it makes things *worse*: A_{all} falls from an already-weak 0.014 to an essentially null 0.001, even though the individual importance-sampled conv distances there are, if anything, slightly better-behaved than their random-subset counterparts (conv1 flips from -0.013 to the correctly-signed $+0.036$; conv2's wrong-signed -0.054 shrinks to -0.017). We read this as evidence that the bottleneck is not primarily *which* units are sampled, but the *aggregation step*: deliberately seeking the smallest-$|z_i(x)|$ units guarantees an even smaller numerator than random sampling would, so on a very high-dimensional layer (MNIST's conv1 alone has 25,088 units) some sampled unit will have a tiny $|z_i(x)|$ in essentially every example regardless of whether that example is actually vulnerable, and taking a minimum over the combined pool inherits whichever layer's sampling noise is largest. Where this effect is small relative to genuine signal (FashionMNIST, apparently), importance sampling helps; where it dominates (MNIST), it hurts.

Sensitivity to the subsample budget. We fixed 300 units per conv layer above without justifying that number; we test its effect on the λ=0 model per dataset by rerunning both sampling schemes at 100 and 1000 units per layer.

Table 13: Pooled correlation of $A_{\text{all}}(x)$ with $R(x)$ as the conv-layer subsample budget varies (single λ=0 model per dataset, $n = 500$). "Importance (any budget)" reflects that the importance-sampled value was numerically *identical* across all three budgets for every dataset (to displayed precision), not merely similarly correlated.

Dataset	random, 100	random, 300	random, 1000	importance (any budget)
MNIST	$+0.181$	$+0.034$	-0.042	$+0.078$
FashionMNIST	-0.070	$+0.015$	-0.130	$+0.035$
CIFAR-10	$+0.040$	$+0.008$	$+0.032$	$+0.053$
CIFAR-10-AT	$+0.072$	$+0.131$	$+0.163$	$+0.079$

The two sampling schemes respond completely differently to budget. The uniform-random $A_{\text{all}}(x)$ is unstable: on MNIST and FashionMNIST it flips sign across budgets, and nowhere does it change monotonically as the budget grows, even though each individual example's random-subset minimum can only shrink (or stay the same) as more units are added — the *correlation* with $R(x)$ depends on how much each example's minimum shrinks, which varies unpredictably example-to-example. The importance-sampled $A_{\text{all}}(x)$, in contrast, is exactly the same value, to displayed precision, at 100, 300, and 1000 units, for every dataset: the unit with smallest $d_i(x)$ among the 300 or 1000 smallest-$|z_i(x)|$ candidates is (numerically) already found among the smallest 100 in every example we checked, so additional budget contributes nothing further. This directly answers the concern that our $A(x)$ results might be an artifact of the specific budget chosen: they plausibly are, for the uniform-random variant, which is one more reason to treat $A_{\text{all}}(x)$ under uniform-random sampling as unreliable on its own; the importance-sampled variant is, by this test, robust to budget specifically because it targets informative units directly rather than relying on how much of the unit space a random draw happens to cover.

Layer-normalizing the aggregation. Different layers have very different raw distance scales — e.g. on CIFAR-10-AT, the mean A_{FC} distance (0.128) is roughly 8–55× the mean conv-layer distance (0.002–0.016) — so an unnormalized minimum across layers is dominated by whichever layer happens to have the smallest raw scale, not necessarily the layer with the most information. We test this directly (no new computation needed: we already have every layer's per-example minimum) by dividing each layer's distances by that layer's own pooled mean before taking the combined minimum, $A_{\text{all}}^{\text{norm}}(x) = \min_l A_{\text{layer}\,l}(x)/\bar{A}_{\text{layer}\,l}$, and by replacing the hard minimum with a softer LogSumExp aggregation at temperature τ, $-\tau \log\big(\frac{1}{4}\sum_l \exp(-A_{\text{layer}\,l}^{\text{norm}}(x)/\tau)\big)$, which down-weights (rather than fully discards) layers whose minimum is driven by noise. Table 14 reports both against the unnormalized $A_{\text{all}}(x)$ (uniform-random sampling) and $A_{FC}(x)$ alone.

Table 14: Layer normalization consistently improves the combined clean-input measure over the raw (unnormalized) minimum, though it does not consistently beat $A_{FC}(x)$ alone. $A_{\text{all}}^{\text{norm}}$: hard minimum over layer-normalized distances. Soft-min (τ=1): LogSumExp aggregation of the same normalized distances. Bootstrap 95% CI in brackets for $A_{\text{all}}^{\text{norm}}$.

Dataset	$A_{FC}(x)$	$A_{\text{all}}(x)$ (raw)	$A_{\text{all}}^{\text{norm}}(x)$	soft-min (τ=1)
MNIST	0.165	0.014	0.064 [0.028,0.100]	0.075
FashionMNIST	0.139	0.029	0.061 [0.021,0.101]	0.073
CIFAR-10	0.139	0.056	0.094 [0.051,0.138]	0.150
CIFAR-10-AT	0.099	0.106	0.110 [0.065,0.155]	0.159

Layer normalization helps, consistently, in every dataset — unlike importance sampling, which helped on some and hurt on others. Normalizing roughly doubles to quadruples the raw $A_{\text{all}}(x)$ correlation (e.g. CIFAR-10: $0.056 \to 0.094$), and the softer LogSumExp aggregation helps further still on CIFAR-10 and CIFAR-10-AT ($0.094 \to 0.150$; $0.110 \to 0.159$, respectively, now exceeding $A_{FC}(x)$ alone on both). On MNIST and FashionMNIST, though, even the best normalized/soft-min combined measure (0.075, 0.073) remains below $A_{FC}(x)$ alone (0.165, 0.139): the FC layer still carries more signal on those two datasets than anything we could extract from the (subsampled) conv layers, normalized or not. This confirms the diagnosis above — the unnormalized minimum was indeed the main bottleneck, not an inherent lack of signal in the conv layers — while showing that a better aggregation rule is necessary but not sufficient: even after fixing it, the combined clean-input measure is not uniformly better than the simpler FC-only one.

C CW-L2 CROSS-CHECK: IMPLEMENTATION DETAILS

Rather than DeepFool's single-step linearization, the CW-L2 spot check in Section 5.8 optimizes $\|\delta\|_2^2 + c \cdot f(x+\delta)$ (the standard Carlini–Wagner (Carlini & Wagner, 2017) objective, with f the margin-violation hinge loss) via Adam under a tanh reparameterization, at a small fixed grid of four c values rather than a full adaptive per-example binary search — a cheaper approximation of CW-L2, not a state-of-the-art minimal-distortion attack, and we report it as such. For each example we keep the smallest successful perturbation found across the grid, giving an independent δ_{cw}, $R_{\text{cw}}(x) = \|\delta_{\text{cw}}\|_2$, and attack-point $D_M(x, \delta_{\text{cw}})$. The fixed-radius $D_M^{\epsilon}(x)$ shows the same qualitative pattern as $A_{FC}(x)$ under this cross-check; we omit the detailed numbers here for space.

D A HEURISTIC ACCOUNT OF WHY $A(x)$ IS REAL BUT WEAK

An affine-region heuristic. We do not have a proof, but a simple heuristic is consistent with what we observe. Within the current activation region, f is affine, so to first order the distance to the decision boundary behaves like a standard linear-classifier margin: $R_{\text{intra}}(x) \approx \text{margin}(x)/\|\nabla_x \text{margin}(x)\|_2$, the same linearization DeepFool itself uses at each step. If this affine approximation held everywhere, $R(x)$ would equal $R_{\text{intra}}(x)$ regardless of $A(x)$, and $A(x)$ would carry no information about $R(x)$. But the approximation is only valid *within* the current region, i.e. for perturbations smaller than $A(x)$; informally, $R(x) \approx \min(A(x), R_{\text{intra}}(x))$, since the true decision boundary can only be farther than $R_{\text{intra}}(x)$ predicts once the perturbation is large enough to reach a different linear region. $A(x)$ is therefore relevant only when it is the *binding*

constraint — when the nearest activation boundary is closer than the affine estimate of the decision boundary. For examples where $R_{\text{intra}}(x)$ already binds (the affine region is stable enough over the relevant distance), $A(x)$ is uninformative about $R(x)$ by this argument, and mixing both regimes in one correlation would produce exactly the diluted-but-nonzero, correctly-signed pattern we observe: a real effect in the (unknown) fraction of examples where $A(x)$ binds, washed out by the examples where it does not. This also offers a candidate explanation for why $D_M^\epsilon(x)$ and $A(x)$ are noisier on MNIST than on CIFAR-10/CIFAR-10-AT (Table 4): if activation regions are typically larger relative to $R_{\text{intra}}(x)$ on MNIST (plausible given MNIST's larger median $R(x)$; Table 1), $A(x)$ binds for a smaller fraction of examples there, further diluting the correlation. We flag this as an informal argument to be tested, not a derivation we have verified — doing so rigorously would require directly measuring, per example, which of $A(x)$ and $R_{\text{intra}}(x)$ is smaller, which we have not done.

A toy two-regime model of the attenuation. The argument above can be made quantitative under simplifying assumptions, giving a rough sense of how large the observed $|r| \approx 0.1$–0.2 is consistent with. Let $Z \in \{0, 1\}$ indicate which term binds in $R(x) \approx \min(A(x), R_{\text{intra}}(x))$, with $P(Z{=}1) = p$ the fraction of examples where $A(x)$ binds. Suppose that, conditional on $Z{=}1$, $R(x) = A(x)$ exactly (so $\text{Cov}(A, R \mid Z{=}1) = \text{Var}(A \mid Z{=}1)$), and conditional on $Z{=}0$, $A(x)$ and $R(x)$ are independent ($\text{Cov}(A, R \mid Z{=}0) = 0$), with both conditional means and variances of A and R approximately equal across regimes (i.e. the two regimes differ in *which* variable determines $R(x)$, not in the overall scale of A or R). The law of total covariance, $\text{Cov}(A, R) = E_Z[\text{Cov}(A, R \mid Z)] + \text{Cov}(E[A \mid Z], E[R \mid Z])$, then gives $\text{Cov}(A, R) \approx p \cdot \text{Var}(A)$ (the second term vanishing under the equal-moments assumption), so

$$\text{corr}(A, R) \;\approx\; p \cdot \frac{\sigma_A}{\sigma_R},$$

which is $\approx p$ itself if $A(x)$ and $R(x)$ have comparable scale. Under this toy model, an observed pooled correlation of $|r| \approx 0.1$–0.2 for $A_{FC}(x)$ (Table 4) is consistent with the FC-layer boundary being the binding constraint for roughly 10–20% of examples — not with $A_{FC}(x)$ being a weak signal spread thinly across all examples. We stress every assumption here (perfect coupling within a regime, no cross-regime confound in the means, comparable variances) is a simplification chosen for a clean closed form, not something we have verified holds in our data; we present this as a plausibility argument for why a small pooled correlation is compatible with a real, fully-predictive relationship in a minority regime, not as a validated decomposition of the observed correlation into a specific p.

Testing the toy model directly: it does not hold as literally stated. The toy model above makes a testable prediction: the fraction of examples with $A_{FC}(x) < R_{\text{intra}}(x)$ should be of the same order as the observed correlation (10–20%). We test this directly, computing $R_{\text{intra}}(x) = \text{margin}(x)/\|\nabla_x \text{margin}(x)\|_2$ from the clean input for every example in every model (one batched backward pass per model; no retraining) and comparing it against the already-computed $A_{FC}(x)$.

Table 15: Empirically testing the toy model's binding-fraction prediction. $P(A_{FC} < R_{\text{intra}})$ is the pooled fraction of examples where $A_{FC}(x)$ is numerically smaller than $R_{\text{intra}}(x)$ ($n = 2500$). "Among these, mean $R(x)$ / mean $A_{FC}(x)$" tests the toy model's core assumption that $R(x) = A_{FC}(x)$ in this regime — if it held, this ratio would be ≈ 1.

Dataset	$P(A_{FC} < R_{\text{intra}})$	mean $R(x)$ / mean $A_{FC}(x)$, among these
MNIST	0.999	61.7
FashionMNIST	0.997	73.0
CIFAR-10	0.982	32.9
CIFAR-10-AT	0.947	12.9

The result refutes the toy model's specific numerical prediction. Far from binding for 10–20% of examples, $A_{FC}(x)$ is the numerically smaller term for 95–99.9% of examples in every dataset — essentially always. Worse for the toy model, its core assumption fails even among these "binding" examples: mean $R(x)$ is 13–73$\times$ larger than mean $A_{FC}(x)$ there, not approximately equal as "$R(x) = A_{FC}(x)$ when A_{FC} binds" requires. We think the likely explanation is a version of the same extreme-value effect diagnosed in Appendix B.3: $A_{FC}(x) = \min_i d_i(x)$ is a minimum over

128 individual unit-boundary distances, so by chance *some* unit is almost always closer than the whole-network margin-based estimate $R_{\text{intra}}(x)$, regardless of whether crossing that specific unit's boundary actually changes the affine piece that determines the decision margin. "$A_{FC}(x)$ is numerically the smaller of the two quantities" is therefore a much weaker, much more common condition than "the nearest unit boundary is the one that actually governs $R(x)$," and the two should not be conflated. We take this as evidence that the two-regime model's mechanism, though intuitive, is not the right description of what is actually happening; the small residual correlation of $A_{FC}(x)$ with $R(x)$ we report in the main text is real (Table 4), but this direct test does not support explaining its magnitude via a literal binding fraction, and we withdraw that specific numerical interpretation. What remains standing is only the weaker, qualitative point that a minority-regime story is *one* way a small correlation could coexist with a locally strong relationship; this experiment shows that story is not simply verified by comparing $A_{FC}(x)$ to $R_{\text{intra}}(x)$ unit-by-unit.

E FULL LIMITATIONS

Section 6 summarizes these; we expand each point here.

- **The attack-point $D_M(x, \delta)$ is coupled to $R(x)$ by construction, and decoupling it lowers the correlation substantially (Section 5.6).** This is the most important qualification to every number in Sections 5.1–5.4: once mask sensitivity is measured without reference to δ, its correlation with $R(x)$ drops from $r > 0.9$ to $|r| \approx 0.1$–0.3. This drop reproduces qualitatively under an independent CW-L2-style attack on a 200-example, single-λ subset (Section 5.8), so it is unlikely to be purely a DeepFool artifact, but we have not re-run the full 20-model sweep with a second attack. Our clean-input measure is also only a partial test: exhaustive per-unit computation is tractable only for the 128-unit FC layer, and neither uniform-random nor importance-guided subsampling of the conv layers reliably beats $A_{FC}(x)$ alone, though layer-normalizing the aggregation helps consistently (Appendix B). Exhaustive per-unit computation on the convolutional layers, or on a narrower architecture where it is tractable, remains untested.
- **Single architecture.** All results use one small 3-conv/1-FC CNN. We have not yet tested whether the effect holds for deeper CNNs (ResNet, WideResNet), or transfers to architectures without a clean ReLU on/off mask (ViT/Swin use GELU-MLP blocks and softmax attention, neither of which has an unambiguous binary activation mask in the same sense). Smooth activations (GELU, Swish) do not have a hard on/off boundary to define $M(x)$ against, so D_M and $A(x)$ as defined here do not transfer directly; the natural analogue would replace the binary mask with a soft one (e.g. thresholding $|z_i(x)|$ against a small $\tau > 0$, or working directly with $\nabla_x a_i(x)$ as a continuous local-sensitivity measure), which we have not implemented or tested.
- **Mostly undefended models.** Only one dataset (CIFAR-10) has an adversarially-trained counterpart; MNIST and FashionMNIST are undefended only. The single CIFAR-10-AT model uses a fairly weak adversarial-training recipe (7-step PGD, 15 epochs, no learning-rate schedule) that trades off a large amount of clean accuracy ($0.72 \to 0.53$–0.56) for its robustness gain, so it should be read as a directional control, not a fully tuned robust model.
- $R(x)$ **is a DeepFool approximation**, not the true minimal ℓ_2 distance. We cross-checked the main circularity finding against a simplified CW-L2-style attack, but only on a 200-example, $\lambda = 0$-only subset (Section 5.8); we have not repeated the full sweep with CW-L2, AutoAttack, or a properly adaptive CW-L2. We also found that DeepFool's cumulative path length is almost collinear with $R(x)$ itself (Appendix A), which means it cannot serve as a second, independent bookkeeping control the way iteration count does.
- **Scale and independence of the sample.** Each model is evaluated on 500 test examples with no repeated seeds; these are each model's own first 500 correctly-classified images in a fixed order, not a literal fixed set shared across λ, and the overlap across the five λ models ranges from $490/500$ on MNIST down to $336/500$ on CIFAR-10. More fundamentally, the five λ values per dataset are five related models, not independent replications. We partially address this with a cluster (per-model) bootstrap (Appendix A): the central D_M-R finding survives it, but the CIs are necessarily much wider than the naive per-example bootstrap, and with only 5 clusters even the cluster bootstrap is a coarse fix, not a substitute

for genuinely independent replications (e.g. multiple random seeds per λ, which we did not train).

- **Regularization scheme.** Only ℓ_1-on-activations sparsity was tested; other mechanisms for inducing sparsity (top-k, k-WTA, explicit thresholding) might decouple sparsity and mask stability differently.
- **The decoupled measures have their own design choices.** The fixed-radius sensitivity uses $K = 16$ isotropic random directions and a per-dataset grid of four radii; a gradient-aligned direction strengthened the signal meaningfully on CIFAR-10 but did not rescue MNIST (Appendix B), so direction choice matters on at least one dataset. The crossing-density measure $C(x) = D_M(x, \delta)/R(x)$ (Section 5.7) still uses δ and is computed only at the attack's endpoint, not sampled along the path; whether crossing density is roughly constant along the path or itself varies as the search proceeds remains untested.