SAME PRIVACY GUARANTEE, DIFFERENT ROBUSTNESS: HOW GRADIENT CLIPPING SHAPES ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Trustworthy machine learning requires addressing multiple threats, including privacy leakage and adversarial perturbations. Differential privacy and adversarial robustness provide complementary forms of protection against these risks. Combining techniques for these goals within the same training procedure can change how a model learns. In private gradient-based training, bounding each example's contribution calibrates privacy noise but can also alter the balance between classification and robustness. The loss coefficients and privacy budget alone therefore do not specify the update used to learn these goals. We compare two ways of limiting gradient contributions: bounding classification and robustness gradients jointly or separately, with the same objective and privacy calibration. We analyze the bounded updates before noise is added, then use controlled fine-tuning experiments in two image-classification settings to test their robustness consequences under hard and smooth clipping. Our analysis shows that preserving the balance between loss components within each example need not preserve the combined update across examples. Across both settings and clipping maps, matching magnitudes to joint bounding while retaining componentwise directions improves held-out AutoAttack accuracy; restoring directions alone is less consistent. Joint exceeds Componentwise by over three percentage points in one setting, but their ordering changes with the setting and attack radius. Contribution magnitudes affect learned robustness even when per-example gradient directions are preserved.

1 INTRODUCTION

Differential privacy limits what a trained model reveals about any individual training record. Adversarial robustness concerns maintaining correct predictions under worst-case perturbations of limited magnitude. Training for both goals requires understanding how privacy mechanisms interact with robustness objectives (Hayes et al., 2022; Bu & Zhang, 2023).

TRADES combines classification loss on the original input with a weighted regularizer penalizing prediction differences under perturbation (Zhang et al., 2019). Each record therefore contributes a task gradient plus a weighted robustness gradient.

Differentially private stochastic gradient descent (DP-SGD) rescales gradients exceeding a norm bound, sums the clipped vectors, and adds Gaussian noise (Abadi et al., 2016). The bound limits each record's influence and sets the noise scale.

For a composite loss, *joint clipping* bounds the summed gradient, as in standard DP-SGD. A different design goal is to cap each objective's contribution independently: a large task gradient should not determine how strongly the robustness gradient is attenuated. *Componentwise clipping* implements this by assigning separate bounds before summation. Both rules admit the same overall bound and noise level, but this added control need not preserve the balance specified by the loss.

We study the bounding step of private training, separating its optimization effects from noise injection. Our central distinction is between balance *within* an example and weighting *across* examples. Separate bounds change the relative weights of loss terms; correcting those weights still need not recover Joint, because example-dependent magnitudes can change the summed direction. The ques-

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.

tion is therefore not just whether clipping changes a gradient, but whether restoring each example's intended balance restores the training update.

Geometric analyses establish clipping bias (Chen et al., 2020; Qian et al., 2021); Esipova et al. (2023) separate batch-level direction and norm effects in DP fairness. PCGrad modifies conflicting task directions; GradNorm adapts task weights through gradient norms (Yu et al., 2020; Chen et al., 2018). We study *same-record* components on *shared parameters*, bounded before aggregation: their construction changes component balance and example weighting under a common sensitivity bound. This differs from term-wise privacy for different data dependencies (Takahashi et al., 2020) and clipping disjoint parameter groups (He et al., 2023). Hard clipping and AUTO-S (Bu et al., 2023) separate these effects from hard saturation.

Our contributions are:

- **From component balance to example weighting.** We derive how restoring the relative weights of loss components can still reweight examples and rotate their aggregate. Controlled training shows robustness consequences even when each example's direction is preserved, under both hard and smooth bounding.
- **Quantifying distortion beyond loss coefficients.** We characterize when shared branch limits cannot preserve component balance and quantify the directional distortion remaining after those limits are chosen. Fixed-state measurements further reveal aggregate differences that a global learning-rate rescaling cannot remove.
- **A constructive role for magnitude correction.** Replacing componentwise magnitudes with Joint magnitudes while retaining componentwise directions improves AutoAttack accuracy in both settings and clipping maps, across all three training seeds. This establishes magnitude as a useful intervention axis, rather than alignment with Joint as a universal selection criterion.

2 Problem Statement

We separate three choices in private training: the loss that defines the gradients, the rule that bounds each record's contribution, and the sampling and noise used to release their sum. Holding the loss and privacy calibration fixed lets us study the bounding rule.

2.1 Composite objectives and gradients

Let θ denote the trainable model parameters. Each training example consists of an input x_i and its label y_i. We write $\ell_{\mathrm{task},i}(\theta)$ for the classification loss and $\ell_{\mathrm{rob},i}(\theta)$ for the robustness regularizer. A coefficient $\beta \geq 0$ controls the weight of the regularizer. The loss for this example is

$$\ell_i(\theta) = \ell_{\mathrm{task},i}(\theta) + \beta \ell_{\mathrm{rob},i}(\theta). \tag{1}$$

In our experiments, the task loss is cross-entropy and the robustness regularizer is the Kullback–Leibler (KL) divergence between predictions on an original input and its adversarially perturbed counterpart, as in TRADES (Zhang et al., 2019). An iterative attack constructs the perturbed input x_i^{adv}, which is held fixed when differentiating the training loss (gradient definitions in Appendix A.1).

We denote the task and robustness gradients by g_i and r_i, respectively. Multiplying r_i by β gives the weighted robustness contribution h_i; adding it to g_i gives the composite gradient s_i:

$$g_i = \nabla_\theta \ell_{\mathrm{task},i}, \qquad r_i = \nabla_\theta \ell_{\mathrm{rob},i}, \qquad h_i = \beta r_i, \qquad s_i = g_i + h_i. \tag{2}$$

2.2 Joint and componentwise queries

We call each record's bounded contribution a *per-record query*. Here $\|\cdot\|$ is the Euclidean norm. Define $\mathrm{clip}_C(x) = \min(1, C/\|x\|)x$ for $x \neq 0$, with $\mathrm{clip}_C(0) = 0$. For a total bound $C > 0$, fractions $q \in (0, 1)$ and $1-q$ specify robustness and task norm limits, not realized update proportions. Joint (J) and componentwise (U) hard-clipping queries are

$$J_i = \mathrm{clip}_C(g_i + h_i), \tag{3}$$

$$U_i = \mathrm{clip}_{(1-q)C}(g_i) + \mathrm{clip}_{qC}(h_i). \tag{4}$$

The robustness branch already includes β; equivalently, the unweighted r_i has threshold qC/β when $\beta > 0$.

AUTO-S uses $\Phi_\gamma(x) = x/(\|x\| + \gamma)$, where $\gamma > 0$ controls attenuation and prevents division by zero. Its output norm is below one. To keep the same total bound, our componentwise extension gives the task and robustness terms maximum norms $1 - q$ and q:

$$J_i = \frac{g_i + h_i}{\|g_i + h_i\| + \gamma}, \tag{5}$$

$$U_i = (1 - q)\frac{g_i}{\|g_i\| + (1 - q)\gamma} + q\frac{h_i}{\|h_i\| + q\gamma}. \tag{6}$$

The triangle inequality gives $\|U_i\| \leq C$ for hard clipping and $\|U_i\| < 1$ for AUTO-S, matching Joint's bounds. Scaling each branch's γ by its share also makes both derivatives I/γ at zero: the small-gradient limit is $(g_i + h_i)/\gamma$, independent of q. This enforces branch caps without changing the linearized objective as q varies.

2.3 The private release and comparison

At step t, each training record is included independently in the batch $\mathcal{B}_t$ with a fixed probability (Poisson sampling). Let Q_i denote the chosen bounded contribution, J_i or U_i, and let σ be the noise multiplier. The optimizer receives

$$\widehat{G}_t = \frac{1}{L}\left(\sum_{i \in \mathcal{B}_t} Q_i + \xi_t\right), \qquad \xi_t \sim \mathcal{N}(0, \sigma^2 B^2 I), \tag{7}$$

where L is a fixed averaging denominator, rather than the sampled batch size, and $B = C$ for hard clipping or $B = 1$ for AUTO-S. Noise is added once, before dividing by L; momentum and weight decay act on the privatized update.

Conditional on the privatized history, a bounded query depends only on its record and data-independent randomness, giving add/remove sensitivity at most B. Sampled-Gaussian composition and post-processing cover the final model (Mironov, 2017; Mironov et al., 2019). Matching sampling, σ and steps matches the certified upper bound, not the queries (Appendix A.2).

Same certificate, different updates. For $g = (3, 0)$, $h = (0, 1)$, $C = 1$ and $q = 1/2$, Joint gives $J = (3, 1)/\sqrt{10}$ and Componentwise $U = (1/2, 1/2)$. The constructions share a unit bound and Gaussian-noise calibration but differ in direction: the common privacy certificate remains valid without specifying the optimizer.

Our *parameter-matched* comparisons fix initialization, differentiation convention, β, total bound, bounding map, learning rate (LR), sampling, steps and optimizer. Only query construction changes. Noise distributions match; realized draws need not. Separately selected hyperparameters answer a practical question but do not isolate query construction.

3 Mathematical Comparison

The analysis proceeds from individual records to their aggregate: Theorem 2 quantifies per-record distortion; Lemma 3 and Equation (15) characterize reweighting after balance correction. Fix a model state and write $a = \|g\|$, $b = \|h\|$, $s = g + h$. These pre-noise results do not order final accuracies. Table 2 collects the notation.

3.1 Relative weighting and allocation

Joint bounding applies a common scalar to g and h, preserving their relative coefficients and the direction of nonzero s. Componentwise bounding instead has $U = \alpha_g g + \alpha_h h$. For nonzero hard branches,

$$\alpha_g = \min\left(1, \frac{(1 - q)C}{a}\right), \qquad \alpha_h = \min\left(1, \frac{qC}{b}\right). \tag{8}$$

With $h = \beta r$ and $\beta > 0$,

$$U = \alpha_g(g + \beta^{\text{eff}} r), \qquad \frac{\beta^{\text{eff}}}{\beta} = \frac{\alpha_h}{\alpha_g}. \tag{9}$$

The ratio β^{eff}/β measures the record-specific change in robustness weight. It can vary across records and training. This vector identity need not define a scalar-loss gradient. For linearly independent g, h, unequal coefficients change the summed direction.

For AUTO-S, the coefficients are $\alpha_g = (1-q)/(a + (1-q)\gamma)$ and $\alpha_h = q/(b + q\gamma)$; Equation (9) still applies. We characterize exactly which allocation preserves the coefficients. A hard branch is *saturated* when its norm reaches or exceeds its assigned bound.

Lemma 1 (When component coefficients are preserved)**.** *Let $a, b > 0$. Under hard clipping, if $a \geq (1-q)C$ and $b \geq qC$, then $\alpha_g = \alpha_h$ if and only if*

$$q = q^* := \frac{b}{a+b}. \tag{10}$$

For the AUTO-S construction in Equation (6)*, the same coefficient-equality condition holds without a saturation assumption.*

One fixed q cannot preserve coefficients across different b/a in these regimes. Both unsaturated hard coefficients already equal one; the coefficient proof and collinear exceptions appear in Appendix B.1.

Our next result quantifies what one shared allocation cannot correct. Fix n gradient pairs with $a_i, b_i > 0$ and $s_i \neq 0$. Let P_i project perpendicular to s_i. Define the *off-axis contribution energy*

$$E(q) = \frac{1}{n}\sum_i \|P_i U_i(q)\|^2, \qquad w_i = \|P_i(h_i/b_i - g_i/a_i)\|^2. \tag{11}$$

These weights measure sensitivity to allocation, not example importance in training. Set $\bar{w} = n^{-1}\sum_i w_i > 0$, let q_w be the w_i-weighted mean of q_i^*, and V_w their weighted variance.

Theorem 2 (A heterogeneity floor for fixed allocation)**.** *Let $I \subset (0,1)$ be a nonempty closed interval on which every hard branch saturates. Then*

$$E(q) = C^2\bar{w}\big[V_w + (q - q_w)^2\big], \quad q \in I. \tag{12}$$

The minimum is $C^2\bar{w}[V_w + \operatorname{dist}(q_w, I)^2]$. For AUTO-S, set $\widetilde{w}_i = w_i\{a_i b_i/[(a_i+\gamma)(b_i+\gamma)]\}^2$. Without saturation assumptions, $E(q) \geq \widetilde{w}\, V_{\widetilde{w}}$ for every $q \in (0,1)$, when the total weight is positive.

The hard identity separates correctable allocation error from heterogeneity: positive V_w prevents any shared allocation in I from eliminating off-axis energy. The soft bound gives an analogous obstruction without saturation. Collinear pairs have zero weight; if all weights vanish, $E(q) = 0$. This is not a batch-error or accuracy bound: perpendicular contributions can cancel. Appendix B.2 gives the proof and edge cases.

3.2 Preserving branch balance need not recover Joint

Choosing the balance-preserving allocation separately for each record removes coefficient distortion, but we show that it need not recover Joint. It produces a *radial control*, denoted R: a rescaling of s that preserves its direction when $s \neq 0$ but can change the record's weight.

Lemma 3 (Balance-preserving allocation yields a radial rule)**.** *For $a, b > 0$, setting $q = q^*$ yields*

$$U_{\text{hard}}(q^*) = \min\left(1, \frac{C}{a+b}\right) s =: R_{\text{hard}}, \tag{13}$$

$$U_{\text{AUTO-S}}(q^*) = \frac{s}{a+b+\gamma} =: R_{\text{AUTO-S}}. \tag{14}$$

Both vectors satisfy the original total norm bound and preserve the direction of nonzero s, but need not equal J.

Radial (R) uses $a + b$; Joint uses $\|g + h\|$, which cancellation can reduce. Restoring relative coefficients can therefore leave unequal record weights. The rules agree for aligned branches, and under hard clipping when neither attenuates the sum (Appendix B.3).

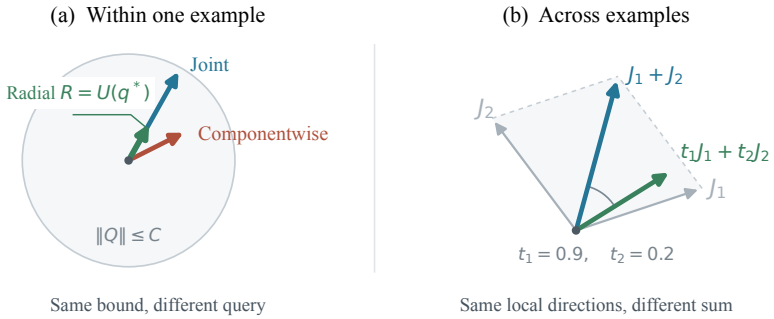

Figure 1: **Two ways the update can change.** Synthetic illustrations. (a) With the same hard norm bound, Componentwise (U) can have a different direction from Joint (J). Choosing the norm-proportional allocation q^* gives the radial query (R), which has J's direction but can be shorter. (b) Rescaling each record's vector by a positive factor t_i preserves its direction but can change the direction of their sum.

3.3 From component balance to example weighting

Write $R_i = t_i J_i$, with $t_i > 0$ for nonzero contributions; if $J_i = R_i = 0$, assign any finite t_i. For a fixed nonempty batch of n records, bars denote arithmetic means. Define

$$v = \frac{1}{n}\sum_i (t_i - \bar{t})(J_i - \bar{J}). \qquad \bar{R} = \bar{t}\,\bar{J} + v. \tag{15}$$

The covariance v links record scales to Joint contributions. Its component perpendicular to nonzero $\bar{J}$ prevents scalar LR recovery; scale variation alone does not force rotation (Appendix B.4). Even for parallel means, LR also rescales noise: at fixed noise calibration, identical one-step Gaussian distributions require equal means and LR (Appendix B.5). This does not rule out successful trajectory-level retuning.

3.4 Projection, radial response, and interventions

Joint hard clipping is a Euclidean projection; AUTO-S instead has a smooth, increasing radial response. These local properties do not order final accuracies (Appendices B.6 and B.7). Comparing both maps tests whether the observed separation requires hard saturation.

To test the roles of direction and magnitude in training, we construct two additional queries from J and U at the same model state:

$$D = \|U\|\,\mathrm{unit}(J), \qquad M = \|J\|\,\mathrm{unit}(U), \tag{16}$$

Here $\mathrm{unit}(x) = x/\|x\|$ for $x \neq 0$ and zero otherwise. Direction-restored (D) uses Joint directions with Componentwise norms; magnitude-restored (M) reverses this pairing. Both respect the common bound and noise calibration, recomputing J/U at their own parameters. When hard Joint saturates, $M = C\,\mathrm{unit}(U)$: magnitude restoration then coincides with fixed-norm normalization. The controls test direction and magnitude changes, not additive accuracy contributions (Appendix B.7).

4 Experiments

The analysis motivates two tests: whether balance-preserving reweighting changes the batch update, and whether retaining directions while changing magnitudes affects trained robustness. Public probes test the first at shared parameters; matched private fine-tuning from robust public pretraining tests the second. Separate tuning assesses utility within the search.

4.1 Experimental setup

Models and datasets. We use ResNet-18 pretrained on public CIFAR-100 for CIFAR-10, and ConvNeXt-T pretrained on synthetic Synset Signset Germany (SSG) for GTSRB (Krizhevsky, 2009;

He et al., 2016; Stallkamp et al., 2012; Liu et al., 2022; Sielemann et al., 2024). The public development sets contain 5,000 and 4,020 images, respectively; the official test sets contain 10,000 and 12,630. Within each setting, runs share initialization and train the final block/stage and classifier with Nesterov SGD. Natural-DP and non-private classification baselines omit the robustness regularizer (Appendix C).

Privacy calibration. Each fixed private run targets $(\varepsilon, \delta) = (4, 10^{-5})$ under image-level add/remove adjacency. Rényi accounting certifies $\varepsilon < 4$; matched rules share the norm bound, absolute noise, sampler and step count (Appendix A.2).

Hyperparameter search. Hard grids cross $C \in \{.1, .3, 1\}$ and LR $\in \{.025, .05, .1\}$; Componentwise also varies $q \in \{.25, 1/3, .5, .75\}$. Following β selection in the CIFAR study, matched comparisons use $\beta = 1$. For AUTO-S, Natural and Joint search three smoothing parameters, $\gamma \in \{.01, .1, 1\}$, with three LRs each. To isolate the effect of gradient decomposition, the Componentwise grid crosses four allocations with the same three LRs as Joint at a shared $\gamma = .1$. This provides an exact parameter-matched Joint control for every Componentwise configuration (Appendix C.4).

Selection and confirmation. We select configurations on the public development sets using training seed 42. In the CIFAR hard search, clean accuracy and 20-step projected gradient descent (PGD) provide a low-cost screen before final selection with standard AutoAttack (AA) (Croce & Hein, 2020). The larger PGD–AA discrepancies in GTSRB motivate full AA coverage of all 57 hard and 30 AUTO-S candidates (Appendix C.4). Hyperparameters remain fixed across confirmation runs; tables identify the training seeds.

Robustness evaluation. We evaluate final checkpoints on the full official test sets, reporting accuracy on unperturbed inputs (clean) and after AA: ℓ_∞ radius $8/255$ on $[0, 1]$ inputs, attack seed 42, FP32 and evaluation mode. Equal pixel amplitudes need not imply equal perceptual difficulty across resolutions (Appendix C.5).

Uncertainty. All $\pm$ values denote sample standard deviations across training seeds; differences pair the same seeds. Bootstrap intervals resample paired examples, or GTSRB dev's 134 tracks. They exclude training-seed variation and selection uncertainty.

4.2 Matched bounding rules produce different held-out robustness

On CIFAR test, selected hard Joint improves AA over natural-DP by $7.87 \pm .05$ points ($22.16 \pm .24\%$ to $30.03 \pm .28\%$), with a clean change of $-1.59 \pm .15$ points ($68.83 \pm .23\%$ to $67.24 \pm .08\%$). Both use the same robust public initialization: the robustness gain comes from fine-tuning, not a different initialization.

For CIFAR, the parameter-matched Joint and Componentwise pairs use $\beta = 1$: hard $C = 1$, LR .025, U $q = 1/3$; AUTO-S $\gamma = .1$, LR .0151027086, U $q = .25$. This hard Joint matches Componentwise's bound and LR, not the selected Joint above. Table 1 summarizes all five rules on fresh seeds; Appendix D.1 retains the initial cohort and clean accuracies.

Table 1: **Matched test results across five bounding rules.** AutoAttack (AA) accuracy at $8/255$, mean $\pm$ sample SD (%), seeds 45–47; parameters and random streams match within rows. J: Joint; U: Componentwise; D: direction-restored; M: magnitude-restored; R: norm-proportional radial. M exceeds U and J exceeds R throughout; J–U changes sign. Clean and AA at $4/255$: Appendix D.

Setting / map	J	U	D	M	R
CIFAR / Hard	$29.78 \pm .17$	$26.56 \pm .04$	$27.38 \pm .21$	$28.27 \pm .10$	$27.64 \pm .14$
CIFAR / AUTO-S	$31.19 \pm .11$	$27.48 \pm .05$	$29.57 \pm .14$	$28.39 \pm .12$	$29.00 \pm .16$
GTSRB / Hard	$13.18 \pm .25$	$13.70 \pm .22$	$12.28 \pm .36$	$14.26 \pm .12$	$12.70 \pm .26$
GTSRB / AUTO-S	$13.74 \pm .78$	$14.02 \pm .12$	$13.31 \pm .90$	$14.71 \pm .13$	$13.41 \pm .75$

On CIFAR, Joint has higher AA in every seed of both cohorts, with mean clean differences below one point. Fresh hard/AUTO-S test gaps are $3.22 \pm .14$/$3.71 \pm .08$ points, versus $3.30/3.16$ on dev. Separately selected hard models yield a $3.81 \pm .03$-point test gap, but their differing bounds and LRs preclude attributing it solely to query construction.

The gap under AUTO-S shows that hard clipping's flat response to large norms is not necessary for a separation. However, the hard and AUTO-S comparisons also differ in q and LR. Their gaps cannot be subtracted to isolate the effect of changing the bounding map.

4.3 Public trajectories: a near-unit median can conceal distortion

Figure 2 measures the effective-weight ratio on 512 public examples at five steps. For hard Componentwise with $q = 1/3$, its median moves from 19.09 to .92, yet final 10th–90th percentiles span .073–1.68. A near-unit median therefore hides record-level distortion. Final medians at $q = .5$ and .75 are 5.97 and 45.86.

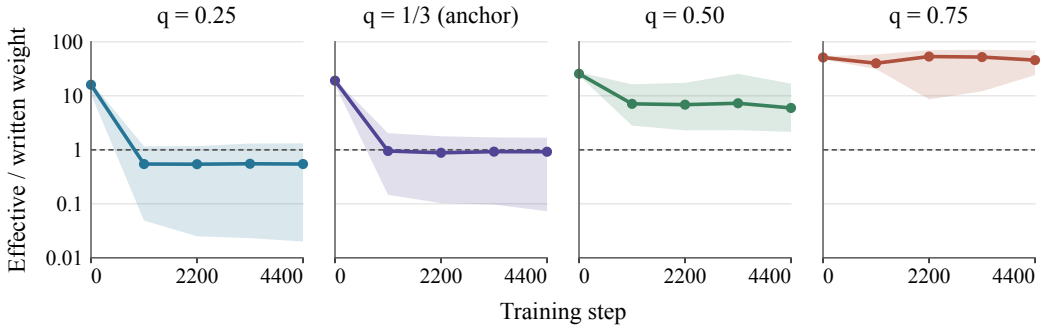

Figure 2: **Implicit weights evolve and remain heterogeneous.** CIFAR Componentwise public-probe measurements, hard $C = 1$, LR .025, $\beta = 1$, seed 42. Lines are medians; bands are 10th–90th record percentiles, not confidence intervals. The dashed unit ratio is the local joint-query reference. Segments connect five snapshots; intermediate steps were not measured.

This dispersion is not itself a model-selection score. Early geometry identifies collapsed models but adds no predictive information beyond q, and a composite diagnostic fails to rank competitive configurations (Appendix D.2).

Averaged over three seeds within each setting/map, off-axis energy at the frozen allocation is 2.70–5.34% of mean squared Joint norm on the same records. Analytic floors retain 2.24–4.67% on this scale after allocation tuning. Hard values use a fixed common-saturated subset; these are contribution-energy fractions, not batch errors or accuracy gaps. The probes also quantify cancellation and radial retention (Appendix D.3).

Correct local directions can still leave a different batch query. At final Joint checkpoints, seeds 45–47, all rules are evaluated on the same public examples and parameters. Figure 3 fits a nonnegative scalar multiple of each alternative mean to the Joint mean. For the direction-restored query D, mean residuals are about 29%/20% in CIFAR and 27%/19% in GTSRB under hard/AUTO-S. These noiseless public queries therefore differ beyond a global step-size change; the residual is not an accuracy prediction.

4.4 Fresh-seed direction, norm, and radial interventions

We train Joint, Componentwise, direction-restored, magnitude-restored and radial rules at frozen anchors, seeds 45–47 (Figure 4). M retains U's directions; switching to J would change both direction and magnitude. Its purpose is controlled intervention, not cheaper computation. Runs share initialization and random streams but follow their own trajectories (anchors and audits: Appendix C.4.3).

Direction and magnitude have distinct training effects. In CIFAR, direction-restored and magnitude-restored training both improve on Componentwise without reaching Joint. M exceeds

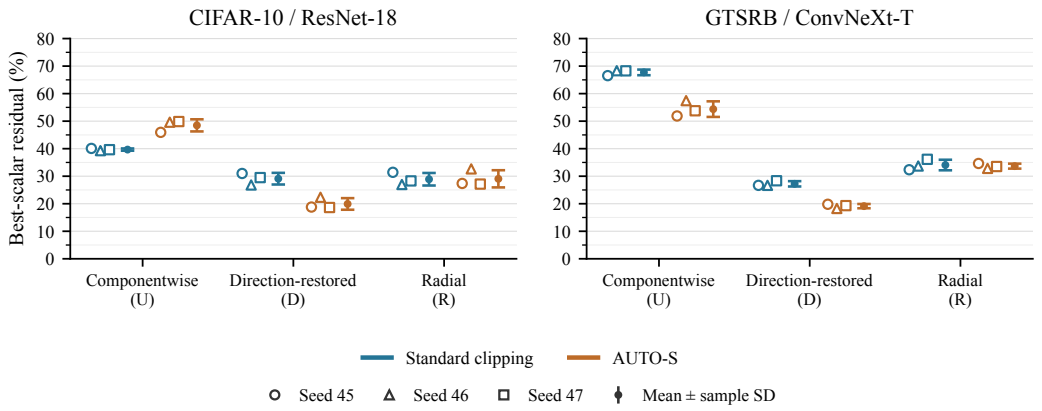

Figure 3: **Preserving local directions does not preserve the aggregate.** Componentwise (U), direction-restored (D) and radial (R) queries at final Joint checkpoints, 512 public examples. Even D/R retain nonzero scalar-fit residuals. Hollow circles/triangles/squares: seeds 45/46/47; separate filled marks and error bars: mean ± sample SD. Axis definition: Appendix B.4; all values: Table 10.

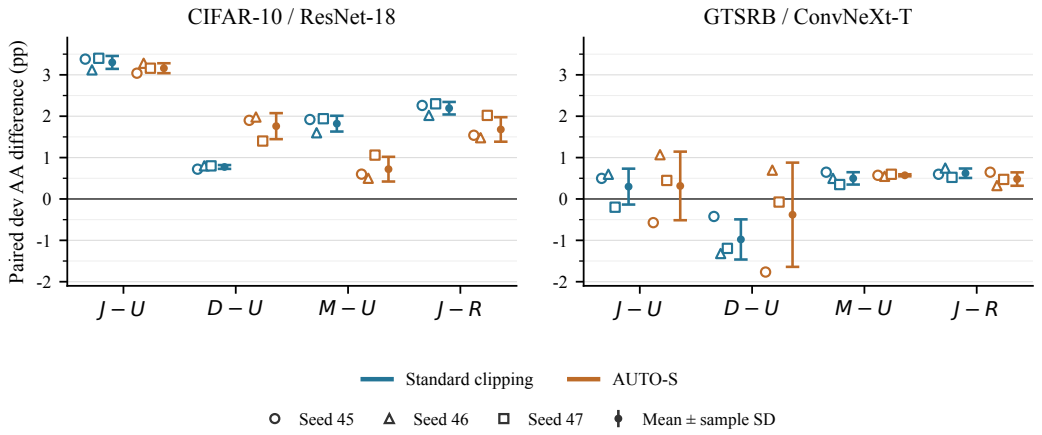

Figure 4: **Magnitude restoration improves robustness more consistently than direction restoration.** Paired dev AA differences (percentage points). Hollow circles/triangles/squares: seeds 45/46/47; separate filled marks and error bars: mean ± sample SD. J is Joint; U Componentwise. D restores Joint directions, M restores Joint norms, and R uses norm-proportional allocation. M−U and J−R are positive in all twelve pairs; D−U changes sign between settings under hard clipping. Each setting/map has its own frozen anchor; cross-setting contrasts do not isolate architecture effects.

D under hard clipping, while D exceeds M under AUTO-S, across all three seeds (Tables 13 and 9). The anchors also differ in allocation and LR, so this reversal does not isolate the map alone.

On GTSRB development data, magnitude restoration improves Componentwise by .50 ± .15 and .57 ± .02 points under hard/AUTO-S. Direction restoration instead gives −.98 ± .49 under hard clipping, negative in all seeds, and −.38±1.26 under AUTO-S. Joint–Componentwise gaps are .30± .43 and .32 ± .83, with track-bootstrap intervals including zero. The interventions thus distinguish consequential magnitude and direction changes even where the direct Joint–Componentwise contrast is small.

Restoring component balance still leaves a measurable effect. Joint exceeds the radial control in all twelve paired development evaluations: 2.19 ± .15/1.68 ± .30 points in CIFAR and .62 ±

.11/.48 ± .16 in GTSRB, for hard/AUTO-S. Both preserve composite-gradient directions but weight examples differently. This demonstrates training consequences of the radial distinction in Lemma 3.

4.5 Selection and held-out robustness

Separate selection retains CIFAR seed-42 dev AA gaps of 3.90 points under hard clipping at the original clean floor and 2.94 under shared-γ AUTO-S without a clean constraint. AA covers every clean-eligible hard cell and the full shared-γ soft grid; these are within-search comparisons (Appendix C.4.1).

GTSRB surfaces distinguish collapsed from competitive configurations (Figure 5). Selected hard Joint uses LR .1, versus the matched .025. Relative to selected Componentwise, its AA difference is $-.17 \pm 1.19$ on dev and $-.27 \pm .24$ on test, with test clean $+.44 \pm .26$ points higher.

Radial and magnitude effects persist on test. CIFAR retains Joint–radial gaps of $2.13 \pm .19$/$2.19 \pm .20$ hard/AUTO-S points. In GTSRB, Joint–radial is $+.48 \pm .09$/$+.34 \pm .14$, and magnitude restoration improves Componentwise by $+.56 \pm .10$/$+.69 \pm .02$, positive in every seed (Table 15). Joint–Componentwise is $-.52 \pm .46$/$-.28 \pm .68$; M also exceeds Joint in every GTSRB seed at $8/255$.

The hard GTSRB test Joint–Componentwise gap persists without noise ($-.34 \pm .13$ points at $8/255$). At $4/255$, Joint instead leads by $1.41 \pm .19$ privately and $1.53 \pm .05$ without noise. Ordering depends on radius and evaluation split; selection remains fixed (Appendix D.5).

5 Discussion

From component balance to example weighting. At the hard GTSRB anchor, uniform attenuation cJ nearly matches Joint's test AA at $8/255$ ($+.008 \pm .049$ points) but exceeds radial R ($+.491 \pm .049$). Constant attenuation alone thus does not reproduce R's robustness at this anchor (Appendix D.5), strengthening the evidence for consequential example-dependent weighting.

Choosing and reporting the bounding rule. Separate branch caps are optimizer choices, not equivalent implementations of a loss. When using them, inspect contribution magnitudes as well as directions and select by robustness at the intended radius. Report differentiation, bounding, aggregation and noise. Matched controls explain changes at fixed settings; separate tuning assesses achievable utility.

Limitations. Experiments concern partial fine-tuning at $\varepsilon = 4$; one zero-noise anchor does not establish a privacy-budget response curve. KL differentiation is fixed within each setting but differs between them, so cross-setting contrasts do not isolate architecture effects (Appendix C.3). Robustness claims use standard AA; enhanced diagnostics cover the tuning cohort and 19 key fresh checkpoints (Appendix C.5).

Our accounting certifies each fixed fine-tuning mechanism. End-to-end certification additionally specifies upstream data dependencies, private model selection and jointly released outputs (Appendix A.2).

6 Conclusion

Preserving component balance need not preserve example weighting or the aggregate update. Our analysis characterizes this distinction; controlled training demonstrates radial gaps and benefits from magnitude correction. Both levels of weighting matter when constructing bounded updates: neither the loss nor the privacy budget specifies their optimization behavior.

AI USE STATEMENT

AI was used solely to obtain feedback on stylistic improvements to the manuscript and to polish the writing. It was not used to conduct experiments, obtain results, or formulate hypotheses.

REPRODUCIBILITY STATEMENT

Proofs are provided in Appendix B; data preparation, training configurations and evaluation protocols are detailed in Appendix C. The supplementary archive contains scientific code, pinned environments, frozen configurations, data-split identifiers, the public CIFAR initialization checkpoint and a SHA-256 inventory. Code and model weights are released at `https://github.com/quiet-gradient/clipping-shapes-robustness`. Independent GPU reruns require external datasets and a CUDA host; GTSRB also requires its public initialization checkpoint and a platform artifact adapter. The release excludes private training examples and internal sampling and noise-generation states.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Zhiqi Bu and Yuan Zhang. Differentially private optimizers can learn adversarially robust models. *Transactions on Machine Learning Research*, 2023.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding gradient clipping in private SGD: A geometric perspective. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13773–13782. Curran Associates, Inc., 2020.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018. URL `https://proceedings.mlr.press/v80/chen18a.html`.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216, 2020.

Maria S. Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C. Cresswell. Disparate impact in differential privacy from gradient misalignment. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=qLOaeRvteqbx`.

Jamie Hayes, Borja Balle, and M. Pawan Kumar. Learning to be adversarially robust and differentially private. *arXiv preprint arXiv:2201.02265*, 2022.

Jiyan He, Xuechen Li, Da Yu, Huishuai Zhang, Janardhan Kulkarni, Yin Tat Lee, Arturs Backurs, Nenghai Yu, and Jiang Bian. Exploring the limits of differentially private deep learning with group-wise clipping. In *International Conference on Learning Representations*, 2023.

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.

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

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium*, pp. 263–275, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with Renyi differential privacy. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-70L8lpp9DF.

Jiang Qian, Yuren Wu, Bojin Zhuang, Shaojun Wang, and Jing Xiao. Understanding gradient clipping in incremental gradient methods. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1504–1512, 2021.

Anne Sielemann, Lena Loercher, Max-Lion Schumacher, Stefan Wolf, Masoud Roschani, Jens Ziehn, and Juergen Beyerer. Synset Signset Germany: A Synthetic Dataset for German Traffic Sign Recognition. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024. doi: 10.1109/ITSC58415.2024.10920175. URL https://huggingface.co/datasets/FraunhoferIOSB/Synset-Signset-Germany-GTSRB-Subset.

Liwei Song, Reza Shokri, and Prateek Mittal. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops*, pp. 50–56, 2019. doi: 10.1109/SPW.2019.00021.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. doi: 10.1016/j.neunet.2012.02.016.

Tsubasa Takahashi, Shun Takagi, Hajime Ono, and Tatsuya Komatsu. Differentially private variational autoencoders with term-wise gradient aggregation. *arXiv preprint arXiv:2006.11204*, 2020.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58244–58282, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, 2019.

A GRADIENT CONVENTIONS AND PRIVACY GUARANTEE

This appendix supports Section 2. It specifies the gradient conventions and the scope of the per-run privacy statement. Their experimental implementation is detailed in Appendix C.

A.1 OUTER-KL DIFFERENTIATION CONVENTIONS

For fixed x^{adv}, write $p = p_\theta(x)$ and $v = p_\theta(x^{\mathrm{adv}})$ for the predicted class-probability vectors. Define $\mathrm{sg}(\cdot)$ as stop-gradient: it preserves the forward value but its derivative is zero. The KL direction is $\mathrm{KL}(p\|v)$. Differentiating both predictions or stopping the original-input target gives

$$r_{\mathrm{full}} = \nabla_\theta \sum_k p_k \log(p_k/v_k), \tag{17}$$

$$r_{\mathrm{stop}} = \nabla_\theta \sum_k \mathrm{sg}(p_k) \log(\mathrm{sg}(p_k)/v_k). \tag{18}$$

Since $\sum_k \nabla p_k = 0$,

$$r_{\mathrm{full}} - r_{\mathrm{stop}} = \sum_k \nabla_\theta p_k \, \log(p_k/v_k). \tag{19}$$

Equal forward KL values therefore need not imply equal training gradients. Our CIFAR-10 implementation uses r_{stop}; our GTSRB implementation uses r_{full}. Neither differentiates through x^{adv}.

The geometric statements apply to either resulting gradient pair, subject to their stated nondegeneracy and saturation conditions. This is invariance of the mathematical characterization, not equality of the gradients or of the models trained with them. Each empirical comparison holds the differentiation convention fixed; we do not identify the cause of differences between experimental settings by changing this convention alone.

A.2 PRIVACY GUARANTEE AND ACCOUNTING PARAMETERS

Consider one run with initialization and hyperparameters held fixed across adjacent private datasets. Conditional on the preceding privatized history, each record contributes a vector of norm at most B, independently of other records in that step. The add/remove sensitivity of the sum is therefore at most B; replacement adjacency would give a $2B$ bound. For hard clipping $B = C$; for the specified AUTO-S maps $B = 1$. The triangle inequality bounds U, the radial factors bound R, and the direction/magnitude substitutions give $\|D\| \le \|U\| \le B$ and $\|M\| \le \|J\| \le B$. Their internal branch calculations are not separately released.

Independent Poisson sampling followed by one $\mathcal{N}(0, \sigma^2 B^2 I)$ perturbation gives a sampled Gaussian mechanism at each step. We compose its Rényi privacy bounds over the prescribed steps and convert at $\delta = 10^{-5}$ (Mironov, 2017; Mironov et al., 2019). Dividing by the fixed denominator 512, applying momentum, Nesterov acceleration and weight decay, and returning the final iterate are post-processing. Physical microbatches only partition the sum: they do not create extra releases or change the accountant's sampling probability.

Setting	Sampling probability	Steps	σ	RDP-certified ε
CIFAR-10	$1/88$	4,400	1.123046875	3.997635
GTSRB	$512/35189$	3,436	1.22344970703125	3.998383

Sampling probability, step count, noise multiplier and averaging denominator are protocol constants held fixed across adjacent datasets. Accounting uses Opacus 1.5.4 with orders $\{1.1, 1.2, \ldots, 10.9\} \cup \{12, \ldots, 63\}$; GTSRB also includes $\{96, 128, 256, 512, 1024\}$. Both minima occur at order 5.8. These certify $(4, 10^{-5})$-DP for each fixed-run mechanism under its stated assumptions; a shared upper bound does not assert equality of exact privacy loss. Executable reports retain the accounting details.

The stated guarantee concerns fine-tuning with the pretrained checkpoint and hyperparameters fixed across adjacent private datasets. End-to-end certification follows the data dependencies of the full pipeline: upstream access to protected records and private model selection require their own privacy

analysis. Evaluating candidates on public inputs does not remove the privacy cost of training those candidates (Papernot & Steinke, 2022).

Fresh Gaussian draws across steps give each intervention run the stated marginal guarantee. Common draws across methods support paired comparisons, but certification of their joint release requires analysis of that correlated mechanism rather than reuse of the marginal bound. Our reported budgets describe the fixed-run mechanisms, not an end-to-end budget for releasing the experimental collection.

B PROOFS AND EDGE CASES

Table 2: **Notation for gradient construction and geometric analysis.** Record indices are omitted when considering one fixed example and model state.

Symbol	Meaning
g_i, r_i, h_i, s_i	Task gradient; unweighted robustness gradient; weighted robustness gradient $h_i = \beta r_i$; sum $s_i = g_i + h_i$.
β	Robustness coefficient in the loss, before bounding.
C, B	Hard-clipping bound; total bound used for noise calibration ($B = C$ for hard clipping, $B = 1$ for AUTO-S).
$q, 1 - q$	Fractions of the bound assigned to robustness and task contributions.
γ	AUTO-S smoothing parameter.
J, U	Joint and Componentwise bounded contributions.
R, D, M	Balance-preserving radial, direction-restored and magnitude-restored controls.
σ, L	Gaussian noise multiplier and fixed averaging denominator.
a_i, b_i	Component norms $\|g_i\|$ and $\|h_i\|$.
α_g, α_h	Scalars applied separately to the two gradient components.
β^{eff}/β	Relative robustness-weight change α_h/α_g.
q_i^*	Record-specific balance-preserving allocation $b_i/(a_i + b_i)$ in Lemma 1.
$P_i, E(q)$	Projection perpendicular to s_i; mean squared norm of projected Componentwise contributions.
w_i, q_w, V_w	Allocation-sensitivity weights and their weighted mean/variance of q_i^*; these are not training-example weights.
t_i, κ_i	Radial retention $R_i = t_i J_i$ and cancellation ratio $\|s_i\|/(a_i + b_i)$.
$\bar{J}, \bar{R}, v$	Mean contributions and their scale–vector covariance in Equation (15).

B.1 PROOF OF LEMMA 1: COEFFICIENT PRESERVATION

We use the notation of Section 3: $a = \|g\| > 0$, $b = \|h\| > 0$ and $s = g + h$. When both hard branches reach their bounds, equality of their coefficients is $(1 - q)C/a = qC/b$, equivalent to $q = b/(a + b)$. For AUTO-S, coefficient equality gives

$$(1 - q)[b + q\gamma] = q[a + (1 - q)\gamma],$$

so the γ terms cancel and the same condition follows. For a fixed q, coefficient equality across all such records therefore requires their norm ratios $b/a = q/(1 - q)$ to agree. This necessity does not hold for hard records with both branches below their caps: both coefficients are then one regardless of q. If exactly one coefficient is strictly below one, coefficient equality is impossible.

For linearly independent g, h, the equation $U = \lambda(g + h)$ implies $\alpha_g = \alpha_h = \lambda$ by uniqueness of the two coefficients. For collinear vectors the representation is not unique: unequal coefficients may preserve, reverse or cancel the summed direction. These statements concern coefficient equality, not an unconditional equivalence between allocation and direction.

Under double hard saturation and $\beta > 0$, substitution of $h = \beta r$ also gives

$$U = C\left[(1 - q)\frac{g}{\|g\|} + q\frac{r}{\|r\|}\right].$$

The explicit β cancels at fixed g, r while both branches remain saturated. Changing β can change which records saturate and the later model parameters, so this is not independence of the whole training process from β.

Taking logarithms of $\beta^{\text{eff}}/\beta = qa/[(1-q)b]$, with $\text{logit}(z) = \log[z/(1-z)]$, gives

$$\log \frac{\beta_i^{\text{eff}}}{\beta} = \text{logit}(q) - \text{logit}(q_i^*). \tag{20}$$

For two allocations that keep the same records doubly saturated, each log ratio changes by the same constant, $\text{logit}(q') - \text{logit}(q)$. Centered moments and interquantile widths are invariant to this translation. In particular, the empirical variance equals $\text{Var}_i \log(b_i/a_i)$. This dispersion identity does not extend unchanged to finite-γ AUTO-S.

B.2 Proof of Theorem 2: the allocation floor

All gradients below are evaluated at the same fixed model state, with fixed adversarial examples. Write $e_i = g_i/a_i$, $f_i = h_i/b_i$ and $P_i = I - s_i s_i^\top/\|s_i\|^2$, requiring $a_i, b_i > 0$ and $s_i \neq 0$. Since $(1-q_i^*)e_i + q_i^* f_i = s_i/(a_i + b_i)$,

$$P_i[(1-q)e_i + qf_i] = (q - q_i^*)P_i(f_i - e_i). \tag{21}$$

Consequently $w_i = \|P_i(f_i - e_i)\|^2 \in [0, 4]$. For $\rho_i = \langle e_i, f_i \rangle$, the equivalent scalar expression is

$$w_i = \frac{(a_i + b_i)^2(1 - \rho_i^2)}{a_i^2 + b_i^2 + 2a_i b_i \rho_i}. \tag{22}$$

The projection definition is preferable near cancellation. Collinear nonzero pairs with $s_i \neq 0$ have $w_i = 0$: coefficient changes need not create a perpendicular component.

Hard clipping: exact constrained minimum. Choose one nonempty closed interval $I \subset (0, 1)$ satisfying $a_i \geq (1-q)C$ and $b_i \geq qC$ for every retained record and every $q \in I$. For example, restricting q to $[.25, .75]$ permits the fixed subset $a_i, b_i \geq .75C$. This subset must not change as q varies. On it $U_i(q) = C[(1-q)e_i + qf_i]$, so Equation (21) implies

$$E(q) = \frac{C^2}{n} \sum_i w_i (q - q_i^*)^2.$$

If $W = \sum_i w_i > 0$, define $q_w = W^{-1} \sum_i w_i q_i^*$ and $V_w = W^{-1} \sum_i w_i (q_i^* - q_w)^2$. Expanding about q_w cancels the cross term and proves Equation (12). The minimizer on I is the Euclidean projection of q_w onto I; the minimum is

$$\min_{q \in I} E(q) = C^2 \bar{w} \{V_w + \text{dist}(q_w, I)^2\}. \tag{23}$$

The distance term matters even if $V_w = 0$: zero-weight collinear records can restrict the common saturation interval. If $W = 0$, the energy is identically zero and weighted means are undefined. If the common interval is empty, the saturated-cohort claim is inapplicable, not evidence of a positive floor. The weighted-variance decomposition itself is elementary; its role here is to separate allocation error from geometric heterogeneity in the specified query.

AUTO-S: a uniform bound without saturation. The componentwise map in Equation (6) gives

$$P_i U_i(q) = k_i(q)(q - q_i^*)P_i(f_i - e_i), \tag{24}$$

$$k_i(q) = \frac{a_i b_i}{[a_i + (1-q)\gamma][b_i + q\gamma]}. \tag{25}$$

This follows either from Equation (21) or from $P_i g_i = -P_i h_i$ and subtraction of the two branch coefficients. Since both denominator factors are positive and bounded above by $a_i + \gamma$ and $b_i + \gamma$, respectively, $k_i(q) \geq m_i := a_i b_i/[(a_i + \gamma)(b_i + \gamma)] > 0$. With $\widetilde{w}_i = w_i m_i^2$,

$$E(q) = \frac{1}{n} \sum_i w_i k_i(q)^2 (q - q_i^*)^2 \geq \overline{\widetilde{w}} \big[V_{\widetilde{w}} + (q - q_{\widetilde{w}})^2\big] \geq \overline{\widetilde{w}} V_{\widetilde{w}}. \tag{26}$$

The last bound is positive whenever q_i^* vary among positive-weight records. Conversely, a common q_i^* on all positive-weight records makes $E(q) = 0$ at that allocation. The bound need not be tight, and $q_{\widetilde{w}}$ need not minimize the exact soft energy because $k_i(q)$ depends on q. As $\gamma \to \infty$ the bound can vanish; this agrees with the increasingly linear map and is not a uniform-in-γ impossibility result.

Scope and degeneracies. An exactly zero composite gradient has no reference direction, so $s_i = 0$ is excluded rather than assigned an artificial angle. One-zero-branch records already produce a collinear contribution but do not have the unit-vector representation used above. Empirical applications report these exclusions. Zero off-axis energy can also accompany a reversed collinear vector. Moreover, per-record perpendicular components can cancel in a batch. Thus neither a positive energy floor nor a geometric oracle allocation proves a batch-direction gap, an accuracy ordering, or an optimal training hyperparameter. Public-probe calculations in Appendix D.3 hold checkpoints fixed and do not alter selected configurations.

B.3 Proof of Lemma 3: radial reweighting

Substituting $1 - q^* = a/(a+b)$ and $q^* = b/(a+b)$, the two hard branch coefficients both become $\min(1, C/(a+b))$. The first AUTO-S coefficient becomes

$$\frac{a/(a+b)}{a + a\gamma/(a+b)} = \frac{1}{a+b+\gamma},$$

as does the second. Summation proves Lemma 3; triangle inequality establishes the bounds. The radial identity extends continuously to one zero branch by assigning the whole bound to the other ($q^* = 0$ or 1). This defines the endpoint by a limit, not by evaluating a $0/0$ expression; endpoints are not part of the fixed-q sweep. If both branches vanish, all queries and R are zero and q^* is unspecified. If $s = 0$ with nonzero branches, $J = R = 0$, but U need not vanish.

For $s \neq 0$, write $A = a + b$ and $c = \|s\|$. The relative scale $R = tJ$ is

$$t_{\text{hard}} = \begin{cases} 1, & A \leq C, \\ C/A, & c \leq C < A, \\ c/A, & C < c, \end{cases} \qquad t_{\text{AUTO-S}} = \frac{c+\gamma}{A+\gamma}. \tag{27}$$

These formulas follow by dividing R's coefficient by J's. The triangle inequality gives $0 < c \leq A$ and hence $0 < t \leq 1$. With cancellation ratio $\kappa = c/A$, soft retention is $\kappa + (1-\kappa)\gamma/(A+\gamma)$. In the jointly saturated hard regime it is exactly κ. When $c \leq C < A$, however, hard retention stays C/A; increasing cancellation does not strictly reduce retention in every regime.

B.4 Aggregation and scalar corrections

Expanding the centered product in Equation (15) gives $v = \overline{tJ} - \bar{t}\bar{J}$, proving $\bar{R} = \bar{t}\bar{J} + v$. For $\bar{J} \neq 0$, let $P = I - \bar{J}\bar{J}^\top/\|\bar{J}\|^2$. Then $P\bar{R} = Pv$. If this perpendicular component vanishes, $\bar{R} = \lambda\bar{J}$ with

$$\lambda = \bar{t} + \frac{\langle \bar{J}, v \rangle}{\|\bar{J}\|^2}.$$

The means have the same orientation only for $\lambda > 0$; they have opposite orientations for $\lambda < 0$, and the reweighted mean vanishes for $\lambda = 0$. For zero contributions, any finite assignment of t_i consistent with $R_i = t_i J_i = 0$ preserves the identity. The same identity applies to D with $t_i = \|U_i\|/\|J_i\|$ when J is nonzero; unlike R's scales, these values may exceed one.

To quantify whether a scalar correction can reproduce the Joint mean, let $\bar{Q}$ be an alternative mean query computed at the same state. For nonzero $\bar{J}, \bar{Q}$, set $\rho = \langle \bar{J}, \bar{Q} \rangle / (\|\bar{J}\|\|\bar{Q}\|)$. Then

$$\min_{\eta \geq 0} \frac{\|\bar{J} - \eta\bar{Q}\|}{\|\bar{J}\|} = \sqrt{1 - \max(0, \rho)^2}, \qquad \eta^* = \max\left(0, \frac{\langle \bar{J}, \bar{Q} \rangle}{\|\bar{Q}\|^2}\right). \tag{28}$$

Expanding the squared norm gives a quadratic in η; minimizing it on the nonnegative half-line yields the formula. If $\bar{Q} = 0$ and $\bar{J} \neq 0$, the residual is one. If $\bar{J} = 0$, the relative residual is undefined. Figure 3 computes this quantity within each seed before summarizing seeds. It does not fit one scalar to already averaged cosines, nor test whether retuning a noisy training trajectory could recover accuracy.

B.5 Learning-rate correction with fixed Gaussian noise

Condition on one model state, batch and all query-construction randomness, including augmentation and the inner attack. Let $\mu_J, \mu_Q \in \mathbb{R}^d$ denote two deterministic aggregate contributions after the

common averaging denominator, and $\tau > 0$ the shared per-coordinate Gaussian standard deviation at this stage. Absorb the reference LR into the units. For a relative LR $a \geq 0$ and a common standard Gaussian Z, compare

$$X_J = \mu_J + \tau Z, \qquad X_Q(a) = a(\mu_Q + \tau Z).$$

Using $\mathbb{E}[Z] = 0$ and $\mathbb{E}[ZZ^\top] = I_d$ gives

$$\mathbb{E}\|X_J - X_Q(a)\|^2 = \|\mu_J - a\mu_Q\|^2 + d\tau^2(1-a)^2. \tag{29}$$

The first term measures signal mismatch; the second is the cost of rescaling the same noise. The minimizing nonnegative multiplier is

$$a^* = \max\left(0, \frac{\langle \mu_J, \mu_Q \rangle + d\tau^2}{\|\mu_Q\|^2 + d\tau^2}\right).$$

Comparing the marginal Gaussian means and covariances proves that X_J and $X_Q(a)$ have identical distributions if and only if $a = 1$ and $\mu_J = \mu_Q$. This conclusion concerns one conditional query update before momentum, not already diverged optimizer states or the unconditioned Poisson mixture. With independent Gaussian draws, the coupling's noise term would be $d\tau^2(1 + a^2)$; the marginal-equality condition is unchanged.

The pre-LR noise scale is held fixed here. This does not contradict joint clipping-scale/LR reparameterizations that rescale both signal and noise, as in automatic clipping (Bu et al., 2023). Nor does unequal one-step distribution establish unequal final accuracy or preclude successful LR tuning. Equation (29) is a Gaussian second-moment identity applied to the matched-noise comparison, not a new transport or convergence theorem.

B.6 Projection inequality

Let $\mathcal{B}_C = \{v : \|v\| \leq C\}$ and let $\Pi_{\mathcal{B}_C}$ denote Euclidean projection onto this ball. For $J = \Pi_{\mathcal{B}_C}(s)$, its variational inequality is $\langle s - J, V - J \rangle \leq 0$ for $V \in \mathcal{B}_C$. Thus

$$\|s - V\|^2 = \|s - J\|^2 + \|V - J\|^2 - 2\langle s - J, V - J \rangle \geq \|s - J\|^2 + \|V - J\|^2. \tag{30}$$

Taking $V = U$ gives the stated distance comparison. This is a standard convex projection fact, not a new convergence result or a bound on batch-sum error.

B.7 Radial response and bounded interventions

The radial response is the change caused by increasing a vector's norm while holding its direction fixed. To distinguish it from changes perpendicular to the vector, let $F(x) = f(\rho)x$, $\rho = \|x\| > 0$ and $v = x/\rho$. Its Jacobian is

$$DF_x = f(\rho)(I - vv^\top) + [f(\rho) + \rho f'(\rho)]vv^\top.$$

The matrices $vv^\top$ and $I - vv^\top$ project onto the radial and perpendicular directions. Hard clipping at $\rho > C$ has perpendicular/radial gains C/ρ and 0; AUTO-S has $1/(\rho + \gamma)$ and $\gamma/(\rho + \gamma)^2$. At zero, AUTO-S has derivative I/γ; the hard boundary is nondifferentiable.

Choice of smooth extension. For a branch with share p, our rule is $px/(\|x\| + p\gamma)$. It has norm below p and derivative I/γ at zero, irrespective of the share. Using $px/(\|x\| + \gamma)$ would also enforce the cap but would instead have derivative pI/γ: allocation would then change the small-gradient objective. The share-scaled denominator avoids this additional reweighting in the linear regime. Our comparisons concern this specified extension, not all possible smooth componentwise maps.

With $\mathrm{unit}(0) = 0$, $\|D\| \leq \|U\| \leq B$ and $\|M\| \leq \|J\| \leq B$, including cancellations. These bounds require no extra noisy release: only the final sum of D or M contributions is perturbed. Direction substitution for each record need not preserve the direction of the sum, momentum or the parameter step. Magnitude restoration preserves the total norm bound, but does not in general preserve Componentwise's separate branch caps; it is not a drop-in replacement when those caps are a design requirement.

Magnitude substitution and normalization. Direction–norm substitutions have a methodological precedent in Esipova et al. (2023), whose diagnostic controls exchange the norm and direction of clipped and unclipped *batch* gradients. Here both source vectors are already bounded *per-record* queries for a composite loss. Substitution occurs before summation, so restoring individual directions need not restore the aggregate direction. The resulting D/M contributions also retain the common per-record sensitivity bound.

The magnitude intervention can be written explicitly as

$$M_{\mathrm{hard}} = \min\{C, \|s\|\}\,\mathrm{unit}(U), \qquad M_{\mathrm{AUTO-S}} = \frac{\|s\|}{\|s\| + \gamma}\,\mathrm{unit}(U).$$

For saturated hard Joint contributions, it is exactly normalization to C when $U \neq 0$; for unsaturated contributions it retains $\|s\|$. AUTO-S instead assigns a smoothly varying norm below one. Thus the trained M–U contrast tests this specified magnitude change, not whether record-specific Joint norms outperform every simpler normalization. A fixed-norm rule $B\,\mathrm{unit}(U)$ or publicly calibrated $\mathrm{clip}_B(kU)$ would test that further distinction. The cJ experiment in Appendix D.5 addresses constant attenuation versus radial R, not these alternatives to M.

For nonzero J and U, write $j = \|J\|$, $u = \|U\|$, $e_J = \mathrm{unit}(J)$ and $e_U = \mathrm{unit}(U)$. The four direction–magnitude combinations satisfy

$$J - D - M + U = (j - u)(e_J - e_U). \tag{31}$$

This follows by substituting $J = je_J$, $U = ue_U$, $D = ue_J$ and $M = je_U$. Even at a fixed state the factors can interact. Their trained-accuracy contrasts therefore should not be interpreted as additive mediation percentages.

Smooth bounding need not imply a strong response to the written robustness weight. Holding nonzero r, the model state and attack inputs fixed, the soft robustness branch is $u_h(\beta) = q\beta r/(\beta\|r\| + q\gamma)$, for $\beta > 0$. Its logarithmic norm elasticity is

$$\frac{\partial \log\|u_h(\beta)\|}{\partial \log \beta} = \frac{q\gamma}{\beta\|r\| + q\gamma}. \tag{32}$$

Logarithmically differentiating its norm gives the identity. The response is positive but can be arbitrarily small; this concerns one weighted branch, not the norm of the summed query or invariance of the trained model to β. The logarithm is undefined when the branch vanishes.

C Detailed Experimental Protocol

C.1 CIFAR initialization and fine-tuning

The public CIFAR-100 source is a seed 42 robust ResNet-18 checkpoint at epoch 170, selected using source validation. It uses a 3×3 stride 1 first convolution, no initial max pooling and group normalization with 32 groups (Wu & He, 2018). Private CIFAR-10 fine-tuning replaces the classifier with a 10-class head: zero bias, $\mathcal{N}(0, .01^2)$ weights. Layer4 and classifier have 8,398,858 trainable parameters. Split seed 20260820 partitions 45,000 private/5,000 public-dev images. The 10,000 official test images are held out for frozen-model evaluation.

Training uses four-pixel padding, random crop and horizontal flip, and one normalization with means $(.4914, .4822, .4465)$ and standard deviations $(.2023, .1994, .2010)$. Inner attacks use 10 steps, $2/255$ step size and $8/255$ radius after coordinate conversion; the .001 Gaussian initialization is in normalized coordinates. SGD uses momentum .9, Nesterov acceleration, weight decay 5×10^{-4} and 50-epoch cosine decay to .01 times the initial rate. There are 4,400 Poisson steps with probability $1/88$, multiplier 1.123046875 and denominator 512.

C.2 GTSRB initialization and fine-tuning

ConvNeXt-T with a convolutional stem processes 96×96 inputs and 43 classes. An ImageNet-initialized network is robustly pretrained on the GTSRB subset of Synset Signset Germany (SSG) (Sielemann et al., 2024), using images from the Cycles and Ogre renderers: 43,000 images from 21,500 scenes, split into 34,400/8,600 public train/validation images without scene-group crossing. Private runs load the fixed epoch 50 checkpoint, retain its classifier and train only the last stage and head normalization/classifier.

The real split has 35,189 private images and 4,020 public-dev images from 134 tracks. Privacy protects one image, not a track; track bootstrap does not change the privacy unit. Fine-tuning has 3,436 Poisson steps at $512/35189$, denominator 512 and multiplier 1.22344970703125. Physical microbatch 128 controls accumulation, not privacy sampling. Nesterov SGD uses the same momentum/decay family, with 10% linear warmup then cosine decay. Inner attacks use 10 steps, $2/255$ step size, $8/255$ radius in raw pixels.

The provenance audit found no exact overlap between the synthetic pretraining images and the real-data images. The source does not fully document the texture generator's training photographs, so this audit establishes image-level non-duplication, not independence of the upstream training data. We accordingly scope the privacy certificate to fine-tuning from the fixed initialization.

C.3 KL differentiation

In the CIFAR-10 experiments, we stop the gradient through the original-input prediction when differentiating the KL regularizer. In the GTSRB experiments, we differentiate both the original-input and perturbed-input predictions. The adversarial input is held fixed in both implementations. Appendix A.1 gives the derivatives.

Our analysis concerns how a given pair of task and robustness gradients is bounded; it does not require a particular choice of KL derivative. Both definitions therefore fall within its scope. In every matched comparison, the gradient definition is held fixed and only the bounding construction changes. This isolates the effect of query construction within each setting. The difference in KL differentiation does, however, preclude attributing differences between settings to the dataset or architecture alone.

C.4 Hyperparameter selection and LR calibration

CIFAR's 36-cell hard attribution grid crosses $\beta \in \{0, 1, 2, 3\}$, $C \in \{.1, .3, 1\}$ and LR $\in \{.025, .05, .1\}$. Two additional LR .2 checks sit outside this rectangular grid and are retained in the search record. Candidates for development AA are first screened using clean accuracy and a 20-step PGD attack. A candidate is Pareto-undominated if no model in the same comparison group has at least as high clean and PGD accuracies, with one strictly higher. Screening is performed

separately for natural-DP and Joint; best-PGD representatives of otherwise omitted β values are added.

For each evaluated Joint model, the natural-DP comparator is the highest-AA model with clean accuracy at least as high as that Joint's. Their AA difference is the clean-matched uplift. A one-standard-error rule retains candidates whose deficit from the largest uplift is no larger than the bootstrap standard error of that *paired deficit*. The standard error uses common resampled examples for the candidate and the best model; it is not the uncertainty of either uplift in isolation. Ties are resolved by higher clean accuracy, smaller β, then run identifier. This selects Joint $\beta = 1, C = .3$, LR $.05$ and its natural-DP comparator at $C = .3$, LR $.05$.

The hard Componentwise grid contains every combination of $q \in \{.25, 1/3, .5, .75\}$, $C \in \{.1, .3, 1\}$ and LR $\in \{.025, .05, .1\}$ (Table 4). Its AA shortlist is formed without a clean constraint: four global clean/PGD-frontier points plus the best-clean and best-PGD representatives at each of $q = .25$ and $q = 1/3$. Deduplication leaves five candidates; three are below the subsequent clean floor. Table 3 summarizes coverage by screening decision.

For the balanced selection, we require clean accuracy of at least 68.60%, the earlier selected Joint reference's measured clean accuracy ($3{,}430/5{,}000$ examples). This inherited constraint asks whether Componentwise can improve robustness without an additional clean-accuracy cost relative to that fixed reference; the cutoff was not fitted to the Componentwise results. It is a development-derived operating-point criterion, not a test of statistical equivalence. Exactly two Componentwise configurations meet it, both with AA evaluations: $(q, C, \mathrm{LR}) = (.25, 1, .025)$ and $(1/3, 1, .025)$. Their AA accuracies are 25.82% and 26.36%, respectively, selecting the latter. Its exact Joint control uses the same $C = 1$ and LR $.025$; it is distinct from the earlier reference at $C = .3$, LR $.05$.

The closest clean-ineligible configuration, $(.25, .3, .05)$, has 68.54% clean accuracy, only three dev examples below the floor, and 27.90% PGD accuracy. It is also dominated in clean/PGD by the selected $(1/3, 1, .025)$ point and was not attacked with AA. Among the five AA-evaluated Componentwise candidates, relaxing the constraint yields a highest observed AA of 27.90% at 66.00% clean accuracy. This is an observed alternative operating point, not an exhaustive unconstrained AA optimum over all 36 cells.

Table 3: **AA coverage of the complete hard Componentwise search.** CIFAR-10, seed 42. Rows are disjoint; clean eligibility is recorded after the frontier/safety screen and does not define that screen. All 36 cells have clean and PGD-20 measurements; both cells meeting the inherited 68.60% clean floor have AA.

Screening group	Trained	AA evaluated	Clean-eligible
Global clean/PGD frontier	4	4	2
Additional $q = .25$ safety point	1	1	0
Other $q \in \{.25, 1/3\}$ cells	13	0	0
All $q \in \{.5, .75\}$ cells	18	0	0

The 31 cells not shortlisted are clean/PGD-dominated and are not additional safety representatives. They are also below the later clean floor, so missing AA does not affect the constrained maximization within this grid. PGD dominance is not an AA certificate: an omitted cell could still improve the unconstrained AA frontier. The complete cell-level table below separates these two claims. The $q = .75$ cells have near-chance clean accuracy (9.88–10.74%); $q = .5$ is degraded but above chance (19.30–38.84%). Neither group is used to inflate the matched comparisons at the selected allocations.

AUTO-S Natural and Joint each explore three γ values and three LRs. For each objective and γ, let $\bar{u}_H$ and $\bar{u}_S$ be mean hard and AUTO-S contributions on the same 512 public probe examples, evaluated without DP noise at a hard reference checkpoint. With reference LR η_H, the center is

$$\eta_S^0 = \mathrm{clamp}_{[\eta_H/4,\, 4\eta_H]}\left(\eta_H \frac{\|\bar{u}_H\|}{\|\bar{u}_S\| + 10^{-12}}\right).$$

The LR grid is $\{.5, 1, 2\}\eta_S^0$. This matches a noiseless mean-update scale heuristically, not training dynamics or optimal LR. All AUTO-S training starts from the common public pretrain, not the

Table 4: **Every hard Componentwise configuration in the CIFAR-10 search.** Training seed 42, $\beta = 1$, fixed final checkpoints and 5,000 public-dev examples; PGD-20 and AutoAttack (AA) use radius $8/255$. Accuracies are percentages. F marks the clean/PGD frontier; S marks the additional best-PGD safety representative for $q = .25$. The other safety representatives already belong to F. A dash means AA was not evaluated, not zero accuracy. Bold clean values meet the inherited 68.60% floor; both eligible cells have AA. Thus the table reports all 36 trained cells, not just the shortlisted or clean-eligible configurations.

q	C	LR	Clean	PGD-20	AA	AA screen
.25	.1	.025	61.28	28.60	—	—
.25	.1	.050	64.86	29.04	—	—
.25	.1	.100	66.78	27.76	—	—
.25	.3	.025	66.90	29.08	27.82	S
.25	.3	.050	68.54	27.90	—	—
.25	.3	.100	66.60	24.90	—	—
.25	1	.025	**69.70**	27.24	25.82	F
.25	1	.050	67.38	24.20	—	—
.25	1	.100	65.10	22.76	—	—
$1/3$	.1	.025	59.48	27.72	—	—
$1/3$	.1	.050	63.54	29.22	—	—
$1/3$	.1	.100	66.28	28.92	—	—
$1/3$	.3	.025	66.00	29.88	27.90	F
$1/3$	.3	.050	68.24	29.10	27.48	F
$1/3$	.3	.100	65.78	25.92	—	—
$1/3$	1	.025	**68.60**	28.10	26.36	F
$1/3$	1	.050	65.58	25.58	—	—
$1/3$	1	.100	63.40	23.72	—	—
.50	.1	.025	20.44	14.16	—	—
.50	.1	.050	29.32	16.72	—	—
.50	.1	.100	27.92	17.90	—	—
.50	.3	.025	32.24	20.44	—	—
.50	.3	.050	33.06	21.08	—	—
.50	.3	.100	20.60	14.88	—	—
.50	1	.025	38.84	21.02	—	—
.50	1	.050	20.96	14.88	—	—
.50	1	.100	19.30	14.18	—	—
.75	.1	.025	9.88	9.88	—	—
.75	.1	.050	9.90	9.90	—	—
.75	.1	.100	9.88	9.88	—	—
.75	.3	.025	10.74	10.74	—	—
.75	.3	.050	9.88	9.88	—	—
.75	.3	.100	9.88	9.88	—	—
.75	1	.025	10.68	10.12	—	—
.75	1	.050	10.74	10.74	—	—
.75	1	.100	10.74	10.74	—	—

private calibration checkpoint. CIFAR references use the selected hard Natural/Joint checkpoints at $C = .3$, LR $.05$.

CIFAR U fixes $\gamma = .1$ and crosses four q values with $\{.0075513543, .0151027086, .0302054171\}$. All 12 receive AA; the clean floor selects $q = .25$ at the center. Independently selected Joint has $\gamma = .01$, LR $.0149960855$; it is not Table 1's exact control.

C.4.1 Separate hyperparameter selection

The CIFAR search also tests whether separately selecting hyperparameters closes the J/U gap. At $\beta = 1$, hard J varies three bounds and three LRs; U additionally varies four allocations. Exactly two configurations per method meet the original 68.60% dev clean floor, and all four have AA evaluations. For AUTO-S at shared $\gamma = .1$, all three J LRs and all twelve U allocation–LR combinations have AA evaluations. Table 5 summarizes these selections without combining different γ search budgets.

Table 5: **Independent selection retains a robustness gap.** CIFAR-10, training seed 42, 5,000 public-dev examples, AA at $8/255$, $\beta = 1$. Each method maximizes AA within its grid and the stated clean constraint. Accuracies are percentages; differences are percentage points. AUTO-S fixes shared $\gamma = .1$. These are selection results, not fresh-seed retuning or unrestricted optima. J denotes Joint and U Componentwise; J−U is the AA difference.

Map	Clean constraint	Clean J	Clean U	AA J	AA U	J−U
Hard	$\geq 68.60\%$	68.60	68.60	30.26	26.36	+3.90
AUTO-S	$\geq 68.60\%$	69.14	69.60	30.98	27.40	+3.58
AUTO-S	None	69.14	68.24	30.98	28.04	+2.94

The hard selections use $(C, \mathrm{LR}) = (.3, .05)$ for J and $(q, C, \mathrm{LR}) = (1/3, 1, .025)$ for U. Both AUTO-S selections use LR $.0151027086$, with U allocation $.25$ under the clean constraint and $1/3$ without it. Halving the LR of the clean-eligible soft U increases AA from 27.40% to 27.78%, but reduces clean accuracy from 69.60% to 67.06%. The unrestricted hard AA optimum is not established: only five of the 36 U configurations received AA. D/M/R have no separate tuning surface and remain fixed-anchor mechanism controls. The Joint-only β sweep does not test matched J/U sensitivity to β; those comparisons fix $\beta = 1$ and $\varepsilon = 4$.

C.4.2 GTSRB selections

GTSRB fixes $\beta = 1$ for the robustness-regularized models and evaluates all 57 hard models: 9 Natural-DP (labeled CleanDP in the GTSRB reports), 9 Joint, 36 U and 3 non-private cross-entropy baselines. Its practical pair is the highest-AA Joint and the highest-AA U with clean accuracy at least as high as that Joint's. This is a development-derived rule, not a preregistered choice or the CIFAR uplift rule. The parameter-matched control instead uses U's bound and LR. Configuration choices are frozen before further seed and held-out evaluations.

The 30 AUTO-S models comprise 9 CleanDP, 9 Joint and 12 U. Shared-$\gamma = .1$ selection gives J/U at LR $.04304104621706269$, U $q = .25$. All cells are retained, including the unconstrained U point with AA 14.9254% but clean 54.5771%. GTSRB calibration uses preliminary hard references at $C = 1$, LR $.05$, rather than the later selected hard Joint at LR $.1$. Shared $\gamma = .1$ for U was inherited from the CIFAR study. In a separate sensitivity analysis, evaluating both Joint references on the same CPU changes centers by factors of 1.92–2.38 across γ. This comparison changes both reference parameters and reference LR; it does not isolate the checkpoint alone or measure the accuracy of a different grid. The original training configurations were retained.

C.4.3 Frozen anchors for the direction, magnitude and radial controls

Anchors were selected on seed-42 dev results and frozen before fresh seeds 45–47. Each five-rule comparison retains its initialization, differentiation convention, $\beta = 1$, bounding map, noise calibration, sampling, steps and optimizer schedule. Table 6 specifies the shared LR and reference Componentwise allocation; the radial rule instead uses its analytic record-dependent allocation. Every rule recomputes its queries at its own parameters.

The experiment in Section 4.4 asks what changing query construction does at a selected operating point. Independent control-specific tuning is valid and could improve utility, but combines query changes with hyperparameter changes. It answers a different question, rather than being required for this matched comparison. These controls are not presented as independently optimized competitors.

Table 6: **Exact anchors for all five-rule comparisons.** The common settings apply to Joint (J), Componentwise (U), direction-restored (D), magnitude-restored (M) and radial (R) training, seeds 45–47, target $(\varepsilon, \delta) = (4, 10^{-5})$. The last column specifies Componentwise's fixed allocation, used to construct D/M; R uses q_i^*. Printed LRs are rounded; configurations retain full precision.

Setting	Map	Bound / smoothing	LR	Reference q
CIFAR	Hard	$C = 1$	.025	$1/3$
	AUTO-S	$B = 1,\ \gamma = .1$	.0151027086	.25
GTSRB	Hard	$C = 1$	.025	$1/3$
	AUTO-S	$B = 1,\ \gamma = .1$	.0430410462	.25

Paired runs share streams for sampling, augmentation, inner attacks and Gaussian noise. Stream audits verify commitments and saved generator positions; final augmentation positions were not recorded for GTSRB and are outside that comparison. Coupling reduces one source of comparison variability, but neither forces equal trajectories nor increases the number of independent training seeds.

Uniform-attenuation and zero-noise controls use the frozen hard GTSRB anchor, seeds 45–47. The uniform control applies $Q_i = cJ_i$ with $c = .8140694935$, calibrated once on the public initialization; it retains the original unit sensitivity declaration and absolute Gaussian noise. The zero-noise controls retain the bounded J/U queries, sampler, optimizer and schedule but set $\sigma = 0$ and therefore make no DP claim. All nine trainings, public-dev AA evaluations and official-test suites pass their integrity checks. Table 7 reports development results; Appendix D.5 reports the full test suite. These fixed-anchor controls are not a new tuning surface.

Table 7: **Uniform-attenuation and zero-noise GTSRB controls on public dev.** Mean $\pm$ sample SD across seeds 45–47, in percent. Private J/R/U are the already frozen hard controls. Uniform cJ tests one constant attenuation; $\sigma = 0$ J/U tests whether the J–U contrast requires Gaussian noise at this anchor.

Query	Clean	PGD-20	AA $8/255$
Uniform cJ	$62.49 \pm .65$	$21.19 \pm .18$	$14.05 \pm .19$
Private J	$63.01 \pm .69$	$21.34 \pm .21$	$14.05 \pm .22$
Private R	$62.37 \pm .25$	$20.65 \pm .32$	$13.43 \pm .13$
Private U	$61.12 \pm .07$	$19.55 \pm .16$	$13.76 \pm .22$
$\sigma = 0$ J	$63.47 \pm .04$	$21.26 \pm .04$	$14.10 \pm .17$
$\sigma = 0$ U	$61.31 \pm .06$	$19.51 \pm .05$	$13.83 \pm .04$

C.5 Attack and uncertainty protocol

Standard AA uses APGD-CE, targeted APGD-DLR (APGD-T), targeted FAB and Square. APGD is adaptive projected gradient descent; CE and DLR identify its cross-entropy and difference-of-logits-ratio attack losses. FAB seeks small boundary-crossing perturbations, while Square is a score-based attack. Historical fields named `apgd_dlr` refer to the targeted APGD stage, not a separate untargeted attack. We retain correctness after each stage, counts of examples still correctly classified, ordered sample IDs, and checks of floating-point precision, evaluation mode and perturbation bounds. CIFAR test additionally has PGD-20 with five restarts and AA $4/255$. AA batches are 128 for CIFAR and 64 for GTSRB.

Enhanced diagnostics cover all 46 triggered GTSRB tuning checkpoints: 45 with PGD–AA gaps above five points and one non-private checkpoint with a zero-gradient warning. The largest additional reduction is .1741 points (mean .0600); three higher-budget Square seeds find no further failures. We separately evaluate 19 key fresh checkpoints: hard J/U/M and practical J, AUTO-S

J/U (all three seeds), and AUTO-S D seed 45. The enhanced union adds APGD-CE and APGD-DLR (100 steps, five restarts) and Square (10,000 queries, seeds 42–44) to each standard-AA mask. Its largest additional reduction is $.0995$ points (mean $.0524$); Square finds no failure beyond the white-box stages. The hard J–U, hard M–U and AUTO-S J–U gaps change from $.299/.498/.315$ to $.315/.531/.357$ points, respectively. Thus the stress test slightly tightens AA without changing these paired conclusions; it is not coverage of every fresh checkpoint or a proof that masking is impossible.

Historical selection and paired-example bootstrap use 2,000 resamples. Fresh GTSRB track intervals use 10,000 resamples of the 134 public-dev tracks. These conditional intervals neither cover training-seed variation nor adjust for selection. The privacy unit remains one image.

One original CIFAR Joint seed-42 evaluation at radius $4/255$ has a one-example discrepancy between predictions recorded during the attack and subsequent reevaluation. We retain the conservative attack-time value, 47.64%, rather than the posthoc 47.65%; the discrepancy is $.01$ percentage points. It does not concern the parameter-matched Joint rows of Table 1 or the $8/255$ headline comparisons.

D ADDITIONAL RESULTS AND PUBLIC-PROBE DIAGNOSTICS

D.1 HELD-OUT CIFAR RESULTS

These tables expand the matched held-out comparisons in Section 4.2 and Table 1, and the fresh-seed test summary in Section 4.5. Table 8 retains the full accuracy statistics of the initial parameter-matched comparison. Its checkpoints are distinct from the fresh intervention cohort. The secondary attack radius is supplementary; model selection uses AutoAttack accuracy on the development set at $8/255$.

Table 8: Initial matched CIFAR test results, seeds 42–44. Mean $\pm$ sample SD in percent. J is U's exact parameter-matched control, not independently selected Joint.

Map	Query	Clean	AA $8/255$	AA $4/255$
Hard	Joint (J)	$67.50 \pm .36$	$29.52 \pm .33$	$47.35 \pm .15$
	Componentwise (U)	$67.16 \pm .15$	$26.22 \pm .30$	$45.35 \pm .18$
AUTO-S	Joint (J)	$67.62 \pm .28$	$30.83 \pm .21$	$48.21 \pm .12$
	Componentwise (U)	$68.26 \pm .13$	$27.05 \pm .30$	$46.41 \pm .05$

Table 9: Fresh CIFAR official-test interventions, seeds 45–47: mean $\pm$ sample SD in percent. All entries use three seeds. D combines Joint directions with Componentwise norms; M combines Componentwise directions with Joint norms; R uses norm-proportional allocation.

	Hard clean	Hard AA $8/255$	AUTO-S clean	AUTO-S AA $8/255$
Joint (J)	$67.58 \pm .28$	$29.78 \pm .17$	$67.56 \pm .04$	$31.19 \pm .11$
Componentwise (U)	$67.59 \pm .36$	$26.56 \pm .04$	$68.48 \pm .21$	$27.48 \pm .05$
Direction-restored (D)	$67.06 \pm .20$	$27.38 \pm .21$	$67.44 \pm .11$	$29.57 \pm .14$
Magnitude-restored (M)	$68.29 \pm .27$	$28.27 \pm .10$	$69.21 \pm .19$	$28.39 \pm .12$
Radial (R)	$67.10 \pm .23$	$27.64 \pm .14$	$67.42 \pm .05$	$29.00 \pm .16$

D.2 NUMERICAL CONVENTIONS AND DIAGNOSTIC RANKING

These conventions and fixed-state measurements support Section 4.3 and Figures 2 and 3; they are distinct from the trained-model outcomes in Appendix D.4. The hard-clipping trajectories use the same 512 public records at steps 0, 1100, 2200, 3300 and 4400. Logged coefficients use a stabilizer $\tau = 10^{-12}$:

$$\tilde{\alpha}_g = \min\left(1, \frac{(1-q)C}{\|g\| + \tau}\right), \qquad \tilde{\alpha}_h = \min\left(1, \frac{qC}{\|h\| + \tau}\right).$$

The recorded effective-weight ratio is $\tilde{\alpha}_h/(\tilde{\alpha}_g + \tau)$ and $\tilde{U} = \tilde{\alpha}_g g + \tilde{\alpha}_h h$. The formulas in Section 3 omit this numerical stabilizer. Recorded task norms bound the ratio stabilizer's relative effect below 5.11×10^{-9} before floating-point rounding; this does not bound all numerical error.

At a measured step t, define the median-mismatch score

$$d_t = \left|\log\left(\operatorname{median}_i \frac{\tilde{\alpha}_{h,i,t}}{\tilde{\alpha}_{g,i,t} + \tau}\right)\right|.$$

It measures deviation of the median ratio from one, not the spread across records. In the 36-configuration hard U grid, d_0 and d_{1100} perfectly rank the 18 models with final clean accuracy below 50% above the other 18. However, simply ranking by q also separates these groups perfectly. Thus the ranking does not establish added information beyond allocation. Step 0 contains only 12 distinct (q, C) settings repeated across LRs, not 36 independent observations.

A second diagnostic combines four quantities: d_t, mean angular rotation of $\tilde{U}$ from s, one minus mean bound utilization $\|\tilde{U}\|/C$, and mean $|\tilde{\alpha}_g - \tilde{\alpha}_h|$. Logged utilization is restricted to $[0, 1]$ and uses $C+\tau$ in the denominator. At each of steps 0 and 1100, each quantity ranks the 36 configurations increasingly, uses average ranks for ties, and rescales ranks as $(\mathrm{rank}-1)/35$. Their mean is the score, with larger values interpreted as greater distortion. None of the four lowest-scoring configurations

belongs to the four-model final clean/PGD Pareto set, at either step. This is a negative result for this particular proposed score, not evidence against every geometry-based selection rule.

Fixed-state aggregate comparisons. At each final CIFAR J checkpoint from seeds 45–47, we evaluate all query rules on the same public examples, holding model parameters fixed. Mean per-record cosine similarity between U and s is .9795 for hard clipping and .9602 for AUTO-S, averaged over the three seeds. Yet the cosine between their batch-mean queries, U and J, is only .9179 and .8744. Even D, which has J's direction for every nonzero record, yields batch-mean cosines .9566 and .9798. Record-wise alignment is therefore not enough to establish agreement of the batch update. These frozen-checkpoint measurements are separate from the trained-model accuracies below.

Table 10: **Numerical values behind Figure 3.** Best-scalar residual $100 \min_{a \geq 0} \|a\bar{Q} - \bar{J}\|/\|\bar{J}\|$, in percent, computed separately at each final Joint checkpoint on the same 512 public examples. Componentwise (U), direction-restored (D), and radial (R) queries are counterfactuals at that fixed state, not separately trained models. The final column is the arithmetic mean and sample standard deviation across the three checkpoint seeds; values are rounded to two decimals. These residuals measure disagreement between noiseless aggregate queries, not accuracy differences or a confidence interval.

Bounding map	Query construction	Seed 45	Seed 46	Seed 47	Mean $\pm$ SD
CIFAR-10 / ResNet-18					
Hard	Componentwise (U)	40.09	39.28	39.64	39.67 ± 0.41
Hard	Direction-restored (D)	30.99	26.81	29.50	29.10 ± 2.12
Hard	Radial (R)	31.42	26.98	28.29	28.90 ± 2.28
AUTO-S	Componentwise (U)	45.93	49.62	49.85	48.47 ± 2.20
AUTO-S	Direction-restored (D)	18.79	22.36	18.63	19.93 ± 2.11
AUTO-S	Radial (R)	27.37	32.65	27.14	29.05 ± 3.12
GTSRB / ConvNeXt-T					
Hard	Componentwise (U)	66.53	68.33	68.28	67.71 ± 1.02
Hard	Direction-restored (D)	26.67	26.68	28.35	27.23 ± 0.97
Hard	Radial (R)	32.35	33.70	36.14	34.06 ± 1.92
AUTO-S	Componentwise (U)	51.88	57.45	53.77	54.37 ± 2.83
AUTO-S	Direction-restored (D)	19.80	18.28	19.31	19.13 ± 0.78
AUTO-S	Radial (R)	34.62	32.83	33.52	33.66 ± 0.90

D.3 Public-probe allocation floors

This subsection supplies the allocation-floor and radial-retention measurements summarized in Section 4.3, using the definitions and proofs in Appendices B.2 and B.3. The diagnostic in Theorem 2 uses frozen final Joint checkpoints and public scalar geometry, not private gradients or a new training sweep. For hard clipping at $C = 1$, the common domain is $I = [.25, .75]$ and the fixed cohort contains records with $a_i, b_i \geq .75$. Its membership is held constant for all compared allocations. AUTO-S uses $\gamma = .1$ and all numerically nondegenerate probe records. The reference allocations are the frozen componentwise values, $q_0 = 1/3$ for hard clipping and .25 for AUTO-S. A geometric oracle minimizes off-axis contribution energy at the same checkpoint; it neither selects a new trained model nor estimates an AA-optimal allocation.

To show absolute scale, define $G_J = |S|^{-1} \sum_{i \in S} \|J_i\|^2$ on exactly the cohort S used to compute $E(q)$. Table 11 normalizes all energies by G_J, not by the possibly small reference distortion $E(q_0)$ or the squared norm of the mean query. Hard and soft cohorts differ, so their energy fractions do not isolate a change of bounding map on identical records.

The hard oracle allocations average .39913 and .35384 for CIFAR and GTSRB; the soft numerical solutions average .40348 and .30752. These are descriptive fixed-state optima, not proposed replacements for the frozen allocations. The soft search evaluates 4,001 grid points and refines every grid-local minimum with a bounded scalar search, including endpoint limits. Equation (26), not the numerical search, supplies the certified positive lower bound.

Ordered public IDs match across the six snapshots within each setting. Hard cohort sizes are 452/455/452 for CIFAR and 332/334/331 for GTSRB; soft uses all 512 records in every snapshot.

Table 11: **Allocation tuning leaves residual energy at a measurable absolute scale.** Final Joint checkpoints, seeds 45–47; 512 public records per original probe. Values are mean ± sample SD of per-seed percentages. Coverage is the retained fraction; the other columns report $100E(q_0)/G_J$, $100L_E/G_J$ for the analytic floor L_E, and $100E(q_{\text{opt}})/G_J$. Hard minima are exact on $[.25, .75]$; soft achieved energies come from numerical search over $[0, 1]$, not a certified global minimum. Energy fractions are not accuracy losses or batch-error bounds.

Setting	Map	Coverage	Frozen q_0	Analytic floor	Achieved energy
CIFAR	Hard	$88.48 \pm .34$	$2.70 \pm .12$	$2.24 \pm .10$	$2.24 \pm .10$
	AUTO-S	$100.00 \pm .00$	$4.80 \pm .07$	$2.25 \pm .04$	$2.36 \pm .03$
GTSRB	Hard	$64.91 \pm .30$	$4.74 \pm .10$	$4.67 \pm .10$	$4.67 \pm .10$
	AUTO-S	$100.00 \pm .00$	$5.34 \pm .19$	$4.24 \pm .15$	$4.95 \pm .15$

There are no zero-branch or near-cancellation exclusions: the minimum observed $\|s\|/(a+b)$ exceeds .152. Independent Gram-form energy checks agree within 7×10^{-17} on saved scalars. Stored CIFAR q^* values have FP32 rounding below 7.2×10^{-8}; the analysis recomputes them from norms and does not infer unavailable raw vectors. Source hashes, individual-seed values, exclusions and numerical tolerances are retained in the analysis outputs. These measurements establish local residual energy, not its causal contribution to learned robustness.

Cancellation and radial retention. Table 12 evaluates $\kappa_i = \|s_i\|/(a_i + b_i)$ and $t_i = \|R_i\|/\|J_i\|$ on the same cohorts. Smaller κ_i means stronger cancellation. Equation (27) determines t_i exactly, so the table measures the prevalence and heterogeneity of this mechanism rather than testing an independent correlation. An interquartile width of .329–.515 shows that R does not apply one common attenuation. Heterogeneity alone does not prove aggregate rotation; Figure 3 separately measures the resulting aggregate discrepancy on the full 512-record probes.

Table 12: **Balance-preserving allocation retains heterogeneous attenuation.** Final Joint checkpoints, seeds 45–47, with the same cohorts as Table 11. Entries average per-seed record quantiles or interquartile widths; they are not pooled quantiles, confidence intervals or training-seed SDs. Individual-seed statistics are retained in the analysis outputs. Here $\kappa = \|g + h\|/(\|g\| + \|h\|)$ measures cancellation, $t = \|R\|/\|J\|$ radial retention, and IQR is the interquartile range.

Setting	Map	κ: P5	Median	P95	t: median	IQR
CIFAR	Hard	.339	.838	.989	.838	.515
	AUTO-S	.333	.835	.990	.836	.499
GTSRB	Hard	.323	.731	.963	.731	.360
	AUTO-S	.357	.917	.997	.918	.329

D.4 Development and held-out intervention results

The development tables give the full accuracies underlying Figure 4 and Section 4.4; the complete three-seed mechanism test table and tuning surface support Section 4.5.

Table 13: CIFAR fresh-seed public-dev interventions: mean ± sample SD, seeds 45–47, not the historical official-test models. Accuracies are percentages; AA uses $8/255$. D combines Joint directions with Componentwise norms; M combines Componentwise directions with Joint norms; R uses norm-proportional allocation.

	Hard clean	Hard AA	AUTO-S clean	AUTO-S AA
Joint (J)	$68.48 \pm .32$	$30.62 \pm .23$	$69.04 \pm .18$	$31.15 \pm .20$
Componentwise (U)	$68.52 \pm .08$	$27.32 \pm .16$	$69.46 \pm .27$	$27.99 \pm .12$
Direction-restored (D)	$68.33 \pm .18$	$28.09 \pm .20$	$68.77 \pm .23$	$29.75 \pm .22$
Magnitude-restored (M)	$69.22 \pm .14$	$29.14 \pm .26$	$70.21 \pm .03$	$28.71 \pm .40$
Radial (R)	$68.39 \pm .31$	$28.43 \pm .15$	$68.70 \pm .17$	$29.47 \pm .27$

Table 14: GTSRB fresh-seed public-dev interventions: mean ± sample SD, seeds 45–47. J denotes the exact matched control in each clipping rule. Accuracies are percentages; AA uses $8/255$. D combines Joint directions with Componentwise norms; M combines Componentwise directions with Joint norms; R uses norm-proportional allocation.

	Hard clean	Hard AA	AUTO-S clean	AUTO-S AA
Joint (J)	$63.01 \pm .69$	$14.05 \pm .22$	$62.45 \pm .33$	$14.00 \pm .86$
Componentwise (U)	$61.12 \pm .07$	$13.76 \pm .22$	$62.71 \pm .09$	$13.68 \pm .04$
Direction-restored (D)	$62.34 \pm .20$	$12.78 \pm .45$	$62.16 \pm .07$	13.30 ± 1.29
Magnitude-restored (M)	$63.20 \pm .10$	$14.25 \pm .11$	$63.88 \pm .21$	$14.25 \pm .02$
Radial (R)	$62.37 \pm .25$	$13.43 \pm .13$	$61.87 \pm .24$	13.52 ± 1.02

For GTSRB, track-bootstrap 95% intervals for mean seed-paired dev differences are $[-1.00, 1.48]$/$[-1.42, 1.87]$ for J$-$U and $[.15, 1.29]$/$[.17, .86]$ for J$-$R, under hard clipping/AUTO-S. M$-$U intervals are $[.05, 1.03]$/$[.14, 1.04]$. These condition on the three trained pairs; they do not make the number of training seeds larger.

Table 15: **Complete GTSRB mechanism test results.** All 30 checkpoints: five rules, two maps, seeds 45–47; 12,630 official-test images per checkpoint. Accuracies are mean ± sample SD in percent. Joint and Componentwise are parameter-matched. Direction-restored (D) retains Componentwise norms and Joint directions, magnitude-restored (M) retains Componentwise directions and Joint norms, and radial (R) uses norm-proportional allocation.

Map	Query	Clean	AA $8/255$	AA $4/255$
Hard	Joint (J)	$59.33 \pm .27$	$13.18 \pm .25$	$28.10 \pm .05$
	Componentwise (U)	$57.12 \pm .02$	$13.70 \pm .22$	$26.69 \pm .15$
	Direction-restored (D)	$57.73 \pm .06$	$12.28 \pm .36$	$26.78 \pm .14$
	Magnitude-restored (M)	$59.48 \pm .05$	$14.26 \pm .12$	$27.58 \pm .15$
	Radial (R)	$58.38 \pm .15$	$12.70 \pm .26$	$27.28 \pm .10$
AUTO-S	Joint (J)	$59.21 \pm .24$	$13.74 \pm .78$	$28.92 \pm .19$
	Componentwise (U)	$58.35 \pm .25$	$14.02 \pm .12$	$27.37 \pm .34$
	Direction-restored (D)	$57.86 \pm .21$	$13.31 \pm .90$	$27.87 \pm .48$
	Magnitude-restored (M)	$60.13 \pm .18$	$14.71 \pm .13$	$28.16 \pm .22$
	Radial (R)	$58.00 \pm .07$	$13.41 \pm .75$	$27.95 \pm .28$

At the secondary radius $4/255$, Joint exceeds Componentwise by $1.41 \pm .19$/$1.55 \pm .17$ hard/AUTO-S points, unlike their ordering at $8/255$. This is a threat-radius-dependent comparison, not a change in the privacy budget.

The fresh dev cohort contains 42 trained models, all of which have official-test evaluations: 30 mechanism checkpoints and 12 baseline or separately selected checkpoints. The non-private Natural seed-45 run has dev/test clean accuracies $.7463\%$/$.4751\%$ and is retained as a failed outcome.

Table 16: **Frozen GTSRB baselines and separately selected Joint on test.** The 12 checkpoints outside the five-rule comparisons: three seeds (45–47) per row, 12,630 official-test images. Accuracies are mean ± sample SD in percent. Non-private Natural includes the collapsed seed-45 run rather than dropping it. Separately selected hard Joint uses LR .1; the matched Joint in Table 15 uses .025.

Model	Clean	AA $8/255$	AA $4/255$
Non-private Natural	51.59 ± 44.34	6.35 ± 5.19	16.28 ± 13.72
Hard CleanDP	$62.77 \pm .92$	$10.97 \pm .09$	$25.14 \pm .21$
AUTO-S CleanDP	$62.88 \pm .44$	$10.97 \pm .09$	$24.89 \pm .14$
Selected hard Joint	$57.55 \pm .24$	$13.43 \pm .03$	$28.60 \pm .80$

Across the other 41 checkpoints, test clean accuracy is 2.73–6.89 points below dev. Matched preprocessing and identity checks pass. A class-frequency decomposition attributes $+.14$ to $+.70$ points

to the changed class mix, which therefore does not explain the decrease; the within-class component is -7.32 to -3.09 points. This decomposition is descriptive, not an identified causal mechanism.

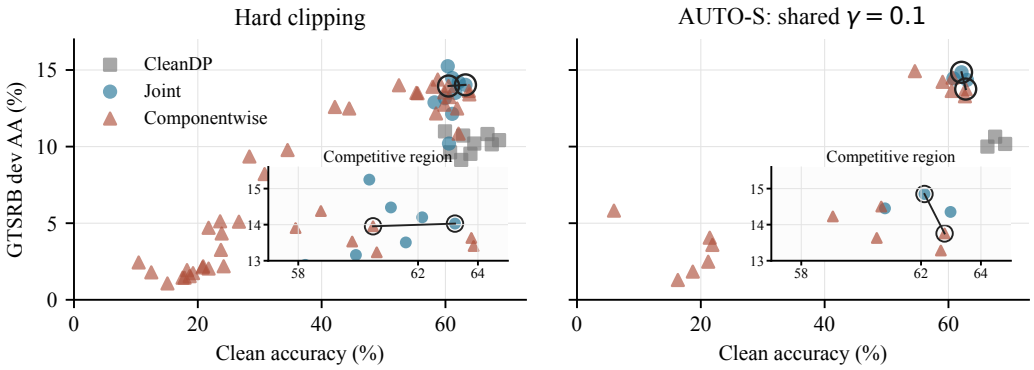

Figure 5: **Large collapsed-model gaps do not describe the competitive region.** GTSRB seed-42 dev surfaces. The hard panel shows the full private grid; the soft panel restricts both methods to shared $\gamma = .1$. Rings and lines identify the exact parameter-matched Joint and Componentwise selections, not interpolated performance. Insets show the competitive region. The three non-private baselines and other soft γ values are outside these panels.

D.5 Uniform attenuation and noise removal on official test

The nine additional controls evaluate constant attenuation and noise removal at the frozen hard GTSRB anchor ($C = 1$, LR .025, $\beta = 1$, reference $q = 1/3$). Table 17 reports all three seeds, reusing the private Joint, Componentwise and radial references without retraining or reselection. Each checkpoint is evaluated on all 12,630 official-test images with PGD-20 and five restarts, and standard AutoAttack (AA) at both radii. Archive and checkpoint hashes validate; all nine suites pass, with no NaNs, scientific warnings or attack-time/posthoc prediction discrepancies.

Table 17: **Constant attenuation and noise removal at a frozen GTSRB anchor.** Official-test accuracy (%), mean ± sample SD across seeds 45–47. J is Joint, U Componentwise, and R the norm-proportional radial rule. Uniform cJ uses $c = .8140694935$ and the original absolute Gaussian noise; the $\sigma = 0$ runs retain bounding but have no DP guarantee.

Query	Clean	PGD-20×5	AA 8/255	AA 4/255
Uniform cJ	$58.85 \pm .33$	$19.63 \pm .19$	$13.19 \pm .23$	$27.74 \pm .02$
Private J	$59.33 \pm .27$	$19.75 \pm .28$	$13.18 \pm .25$	$28.10 \pm .05$
Private R	$58.38 \pm .15$	$18.89 \pm .29$	$12.70 \pm .26$	$27.28 \pm .10$
Private U	$57.12 \pm .02$	$18.97 \pm .43$	$13.70 \pm .22$	$26.69 \pm .15$
$\sigma = 0$ J	$59.47 \pm .04$	$19.93 \pm .05$	$13.30 \pm .11$	$28.22 \pm .03$
$\sigma = 0$ U	$57.49 \pm .04$	$18.85 \pm .03$	$13.65 \pm .02$	$26.69 \pm .04$

Constant attenuation does not reproduce the radial control. At $8/255$, uniform cJ differs from Joint by $+.008 \pm .049$ percentage points, but exceeds R by $+.491 \pm .049$. Their paired-example 95% bootstrap intervals are $[-.069, +.082]$ and $[+.383, +.604]$, respectively. These intervals use 2,000 resamples of ordered test examples after averaging paired correctness differences over the three fixed seeds; they are conditional on those trained models, not intervals over training randomness. The near-equality to Joint concerns AA at $8/255$: clean accuracy is $.480 \pm .078$ points lower, and AA at $4/255$ is $.362 \pm .046$ points lower. The result distinguishes this calibrated constant attenuation from R's example-dependent scaling at one anchor, rather than establishing clean-matched superiority or ruling out every scalar choice.

The contrast persists without noise and reverses with attack radius. At $8/255$, the paired Joint–Componentwise difference is $-.523 \pm .461$ points privately and $-.343 \pm .126$ with $\sigma = 0$. At $4/255$,

the corresponding differences are $+1.409 \pm .192$ and $+1.528 \pm .049$. Thus Gaussian noise is not required for the separation at this anchor, and the radius-dependent ordering occurs both with and without it. The private-minus-zero-noise difference-in-differences at $8/255$ is $-.179 \pm .448$ points across training seeds. Its conditional paired-example interval, $[-.280, -.082]$, excludes zero for these fixed models, but does not establish a stable interaction across training seeds; neither a zero interaction nor noise-independence is inferred.

The hard private Joint–Componentwise difference also changes from $+.299 \pm .434$ points on development data to $-.523 \pm .461$ on test at $8/255$. This split-dependent ordering complements the radius comparison without changing the frozen selection. The geometric results establish different bounded updates, not a universal accuracy ordering.

D.6 Exploratory membership-inference diagnostics

An earlier, separate fine-tuning campaign used a nominal privacy target $\varepsilon_{\mathrm{DP}} = 8$ and $\delta = 10^{-5}$ to exercise a membership-inference pipeline: adversarial-loss scores (Song et al., 2019), online and offline LiRA (Carlini et al., 2022), and RMIA (Zarifzadeh et al., 2024). Across eleven bounded-update targets, the recorded AUC ranges were 0.508–0.512, 0.501–0.507, 0.506–0.521, and 0.511–0.516, respectively. LiRA and RMIA retain the stronger AUC for each target from two banks of sixteen shadow models; the second bank uses non-private training. MIA and RMIA score $45{,}000$ members and $10{,}000$ non-members; LiRA uses $5{,}000$ held-out non-members.

These are descriptive attack diagnostics, not an empirical certification of a privacy upper bound or an ordering of query constructions by leakage. The historical recipes, checkpoint-selection rules, and sampling protocol differ from the matched $\varepsilon_{\mathrm{DP}} = 4$ experiments in this paper. We therefore do not combine the two campaigns or use the earlier diagnostics to establish the present models' privacy or robustness conclusions.