AURORA: ATTACK-TRAJECTORY UNCERTAINTY-AWARE RISK ALLOCATION FOR LONG-TAILED ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing long-tailed adversarial training methods commonly allocate robust supervision using class-frequency priors or empirical endpoint failures. However, class frequency is not a direct measure of adversarial vulnerability, while endpoint failure rates are unreliable under limited class support and discard how failures develop along the attack trajectory. We propose AURORA, an uncertainty-aware framework for estimating class-wise adversarial risk from endpoint outcomes and intermediate attack predictions. Specifically, uncertainty-Aware Risk Anchoring (URA) applies a Wilson-score upper estimate to endpoint failures, allowing class support to affect allocation through estimation uncertainty rather than a predefined frequency-to-weight mapping. First-Passage Trajectory Refinement (FTR) then measures the normalized first-failure iteration among initially correct samples and uses this path information to refine the endpoint anchor. A bounded correction converts the resulting endpoint-to-path profile into class weights for the native robust objective of existing AT methods. Experiments across multiple long-tailed benchmarks and architectures show that AURORA consistently improves tail robustness while retaining competitive overall performance.

1 INTRODUCTION

Adversarial training (AT) is the standard empirical defense against norm-bounded perturbations (Madry et al., 2018; Zhang et al., 2019; Wu et al., 2020; Zhao et al., 2024a). However, its practical efficacy is closely tied to the distribution of training data. In long-tailed scenarios, the severe class imbalance leads to highly uneven robust risk estimation: head classes with abundant samples provide reliable vulnerability estimates, whereas tail classes with scarce observations yield highly uncertain ones. When the training objective treats class-wise predictions uniformly, optimization is biased toward sample-rich head classes, thereby compromising the robustness guarantee for tail classes.

Existing long-tailed AT methods allocate robust supervision using either static class priors or model-dependent feedback derived from adversarial accuracy, margins, feature geometry, and teacher predictions (Wu et al., 2021; Bui et al., 2022; Wei et al., 2023; Lee et al., 2024; Cho et al., 2025; Zhang et al., 2026). Frequency-based methods prescribe vulnerability from class scarcity, although rare classes are not necessarily the most attack-sensitive. Feedback-based methods estimate vulnerability more directly, but commonly treat empirical endpoint failure rates as equally reliable across classes and summarize an iterative attack by its final outcome.

While these methods achieve well long-tailed prediction performance, they overlook two critical issues. First, identical failure rates may be supported by substantially different numbers of observations, so the endpoint feedback does not reflect the uncertainty of the estimated risk. Second, even under matched support, two classes with the same endpoint failure rate may fail at very different attack stages, meaning that endpoint feedback also fails to capture the failure progression along the attack trajectory. Notably, the inner maximization of AT already produces a sequence of attack iterates whose intermediate predictions provide information beyond the final outcome. For an initially correct sample, the first iterate that changes the prediction indicates how quickly the current attack procedure exposes a local failure.

These observations motivate AURORA (*Attack-Trajectory Uncertainty-Aware Risk Allocation*), which estimates class-wise risk in two consecutive stages. URA (*Uncertainty-Aware Risk Anchoring*) replaces the raw endpoint failure rate with a Wilson upper confidence bound, producing a con-

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.

servative risk anchor when class support is limited. FTR (*First-Passage Trajectory Refinement*) then refines this anchor using the clean-conditional first-failure process, recovering temporal vulnerability hidden by the endpoint outcome. Finally, a bounded log-risk correction converts the endpoint-to-trajectory profile into normalized class weights for the next epoch.

Our contributions are summarized as follows:

- We identify an overlooked risk-estimation problem in long-tailed AT: class frequency does not directly measure adversarial vulnerability, while empirical endpoint failures are affected by unequal class support and omit failure dynamics along the attack trajectory.
- We propose URA, which constructs a support-sensitive endpoint risk estimate, and FTR, which refines this estimate using the clean-conditional first-failure process. Together, they form a unified endpoint-to-path failure profile.
- We derive a bounded, unit-mean allocation rule that reweights the native robust objective of AT methods without introducing an auxiliary training loss.
- Experiments across multiple datasets, architectures, base learners, and attacks demonstrate consistent improvements in tail robustness while preserving competitive overall performance.

2 Related Work

Robust optimization and iterative attacks. AT approximates worst-case empirical risk by optimizing inputs within a constrained neighborhood (Goodfellow et al., 2015; Madry et al., 2018). TRADES separates natural prediction from local distributional stability (Zhang et al., 2019), while AWP perturbs model parameters to favor flatter robust solutions (Wu et al., 2020). Subsequent work improves the inner maximization, robust generalization, efficiency, and data or representation regularization (Pang et al., 2021; Rebuffi et al., 2021; Gowal et al., 2021; Rade & Moosavi-Dezfooli, 2021; Li et al., 2022b; Herrmann et al., 2022; Wang et al., 2023). These methods primarily use intermediate attack iterates to optimize a final adversarial example. AURORA instead treats their predictions as detached observations of path-wise failure, without changing the inner maximization.

Learning from unequal class evidence. Long-tailed recognition studies how class frequency distorts both feature learning and the final decision rule. Margin- and logit-based methods incorporate class priors into the training objective (Cao et al., 2019; Ren et al., 2020; Menon et al., 2021; Li et al., 2022a), while decoupled training, classifier retraining, and weight balancing separate representation learning from classifier correction (Kang et al., 2020; Alshammari et al., 2022; Hasegawa & Sato, 2024; Sun et al., 2025; Lin & Yuan, 2026). Complementary approaches increase tail coverage through resampling, augmentation, or generated data (Park et al., 2022; Qin et al., 2023; Wang et al., 2024; Zhao et al., 2024b; Shao et al., 2024), and improve tail representations through contrastive learning, expert models, or conflict-aware optimization (Zhu et al., 2022; Li et al., 2022c; Hou et al., 2023; Wang et al., 2021; Sanchez Aimar et al., 2023; Zhou et al., 2024; Yi et al., 2025).

These methods correct the optimization bias caused by unequal class frequency, but they do not model the uncertainty of attack-derived class statistics.

Adaptive robustness under long tails. Long-tailed AT compounds natural classifier bias with unequal robust optimization. RoBal and BSL correct frequency-induced decision bias through scale-invariant classifiers, margins, or balanced logits (Wu et al., 2021; Yue et al., 2024). UDR introduces distributional robustness, whereas REAT and TAET seek to equalize class-wise adversarial learning (Bui et al., 2022; Li et al., 2023; Yu-Hang et al., 2025). A complementary line modifies robust supervision itself. CFA calibrates class-wise perturbation radii and robust coefficients using robust accuracy (Wei et al., 2023); DAFA uses feature-space distance to guide class-wise adaptation (Lee et al., 2024); RLT imposes long-tailed perturbation priors and schedules (Zhang et al., 2026); and self-distillation transfers robust knowledge from a balanced teacher (Cho et al., 2025). Together, these methods establish that long-tailed AT benefits from class-aware robust supervision.

Frequency-based methods treat class scarcity as a direct proxy for vulnerability, whereas endpoint-feedback methods typically reweight classes using empirical adversarial error. For tail classes, however, limited observations result in high-variance failure estimates. Direct weighting by frequency or endpoint error may therefore overemphase classes with spuriously high failure rates while underprotect those with spuriously low rates. URA instead uses a Wilson-score upper estimate to incorporate

finite-support uncertainty into endpoint risk. FTR complements this estimate with the normalized first-failure behavior of initially correct samples, distinguishing classes that have similar endpoint outcomes but fail at different stages of the attack.

3 METHOD

OVERVIEW AND LEARNING OBJECTIVE

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ contain C classes with unequal sample counts, and let $\mathcal{B}$ denote a base adversarial learner. We write AURORA as a common term ℓ_{com} plus a robust term $\ell_{\text{rob}}^{\mathcal{B}}$. It applies a fixed prior-aware logit correction motivated by long-tailed methods like RoBal (Wu et al., 2021) and reweights only the robust term using class statistics collected in the preceding epoch:

$$\mathcal{L}_{\text{AURORA}}^{(t)} = \frac{1}{B}\sum_{i=1}^{B}\left[\ell_{\text{com},i}^{(t)} + \widetilde{w}_i^{(t)}\ell_{\text{rob},i}^{\mathcal{B},(t)}\right]. \tag{1}$$

For TRADES, ℓ_{com} denotes the cross-entropy loss computed on clean samples and $\ell_{\text{rob}}^{\mathcal{B}}$ is the weighted clean–adversarial KL divergence. For PGD-AT and AWP, $\ell_{\text{com}} = 0$ and $\ell_{\text{rob}}^{\mathcal{B}}$ is the adversarial loss calculated on adversarial examples. The weight in Eq. (1) is expressed as

$$\widetilde{w}_i^{(t)} = \frac{w_{y_i}^{(t)}}{B^{-1}\sum_{j=1}^{B} w_{y_j}^{(t)}}, \tag{2}$$

and therefore satisfies $B^{-1}\sum_i \widetilde{w}_i^{(t)} = 1$. AURORA consequently redistributes the robust-loss multiplier. The class weight $w_c^{(t+1)}$ is derived from one failure profile observed in epoch t. URA establishes its endpoint anchor by combining observed failure with uncertainty from class support. FTR then uses clean-eligible attack trajectories to determine whether that anchored risk emerges early or late. We next derive this endpoint-to-path profile and map it to the weight $w_c^{(t+1)}$.

ATTACK TRAJECTORIES AS FAILURE RECORDS

For each training example, the base learner generates a K-step attack path

$$x_i^{(0)}, x_i^{(1)}, \ldots, x_i^{(K)}, \tag{3}$$

where $x_i^{(0)} = x_i$ and $x_i^{(K)}$ is the final adversarial example. Standard AT uses the intermediate iterates only to obtain $x_i^{(K)}$. AURORA additionally records their already-computed predictions. We use *attack time* to describe the normalized iteration index k/K, not physical time.

The final iterate defines the endpoint event

$$e_i^{(t)} = \mathbb{I}\left[\arg\max_j f_{\theta_t}(x_i^{(K)})_j \neq y_i\right]. \tag{4}$$

where $\mathbb{I}[\cdot]$ means the indicator. This endpoint event records failure at maximum attack time but does not locate its onset. A first-passage time is meaningful only when the path begins from a correct decision, so we define clean eligibility as

$$a_i^{(t)} = \mathbb{I}\left[\arg\max_j f_{\theta_t}(x_i^{(0)})_j = y_i\right]. \tag{5}$$

For each eligible example, the first-passage time is

$$T_i^{(t)} = \min\left\{k \in \{1, \ldots, K\} : \arg\max_j f_{\theta_t}(x_i^{(k)})_j \neq y_i\right\}. \tag{6}$$

If no iterate changes the prediction, we set $T_i^{(t)} = K + 1$ and regard the path as right-censored. Once a crossing occurs, later recovery does not erase it. Cleanly misclassified examples ($a_i^{(t)} = 0$) are excluded only from the first-passage statistic: they measure recognition error, not failure of an initially valid decision. They still contribute to the endpoint event and the training loss. Thus, $e_i^{(t)}$ measures the final failure magnitude, whereas $T_i^{(t)}$ builds its attack-time structure.

A UNIFIED ENDPOINT-TO-PATH FAILURE PROFILE

The proposed URA and FTR share the same confidence operator so that endpoint magnitude and attack-time structure are calibrated under the same notion of finite evidence. Given s events in $n > 0$ Bernoulli observations, let $\widehat{p} = s/n$ and define the Wilson upper endpoint

$$\mathcal{U}(s,n) = \frac{\widehat{p} + \frac{z^2}{2n} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}. \tag{7}$$

We use the upper endpoint of the Wilson score interval (Wilson, 1927) as a finite-support correction. We set $z = 1.96$, corresponding to the conventional nominal 95% score interval. Since the observations are collected online while the model evolves, we use $\mathcal{U}(s,n)$ as a score-based upper risk estimator. It approaches the empirical rate as support increases, while preventing a small number of observations from certifying a class as low risk. We refer to $\mathcal{U}$ as an *upper-confidence risk estimate*.

Uncertainty-Aware Risk Anchoring. For the class c, let

$$N_c^{(t)} = \sum_i \mathbb{I}[y_i = c], \qquad E_c^{(t)} = \sum_i \mathbb{I}[y_i = c] e_i^{(t)}. \tag{8}$$

URA defines the endpoint anchor

$$U_c^{\mathrm{end},(t)} = \mathcal{U}\Big(E_c^{(t)}, N_c^{(t)}\Big). \tag{9}$$

$U_c^{\mathrm{end},(t)}$ increases with observed endpoint failures and remains conservative when support is limited. Hence, URA captures not only the observed failure rate but also the uncertainty induced by unequal class evidence. If class c is absent from an epoch, we set $U_c^{\mathrm{end},(t)} = 1$, the maximal admissible risk.

First-Passage Trajectory Refinement. URA cannot distinguish classes with similar endpoint risks but different failure times. FTR resolves this ambiguity using attack trajectories. Let

$$M_c^{(t)} = \sum_i \mathbb{I}[y_i = c]\mathbb{I}[a_i^{(t)} = 1] \tag{10}$$

denote the number of clean-eligible samples in class c. Their cumulative first-passage count at step k is formulated as

$$H_{c,k}^{(t)} = \sum_{\substack{i: y_i = c \\ a_i^{(t)} = 1}} \mathbb{I}[T_i^{(t)} \le k]. \tag{11}$$

Applying the same operator to each cumulative count yields the upper first-passage curve

$$F_{c,k}^{+,(t)} = \mathcal{U}\Big(H_{c,k}^{(t)}, M_c^{(t)}\Big), \qquad k = 1, \dots, K, \quad M_c^{(t)} > 0. \tag{12}$$

The cumulative construction makes $F_{c,k}^{+,(t)}$ non-decreasing in k. We summarize the curve by its normalized area

$$D_c^{(t)} = \frac{1}{K} \sum_{k=1}^{K} F_{c,k}^{+,(t)}. \tag{13}$$

We call $D_c^{(t)}$ the *upper path risk*. Before applying the finite-support correction, a failure at step one contributes 1 to the area, a failure at step K contributes $1/K$, and a censored trajectory contributes zero. Consequently, a larger $D_c^{(t)}$ indicates earlier observed failures, greater uncertainty due to limited eligible support, or both. When $M_c^{(t)} = 0$, no valid first-passage evidence is available and FTR produces a neutral temporal refinement.

Together, $U_c^{\mathrm{end},(t)}$ and $D_c^{(t)}$ form an endpoint-to-path failure profile.

From Failure Profiles to Robust Allocation

The URA endpoint risk provides the primary allocation signal; FTR refines it using attack time. To place temporal risk on a comparable scale, define the observed set

$$\mathcal{C}_{\text{obs}}^{(t)} = \{c : M_c^{(t)} > 0\}, \tag{14}$$

and its mean value

$$\overline{D}^{(t)} = \frac{1}{|\mathcal{C}_{\text{obs}}^{(t)}|} \sum_{c \in \mathcal{C}_{\text{obs}}^{(t)}} D_c^{(t)}. \tag{15}$$

We then center and clip relative path risk:

$$d_c^{(t)} = \text{clip}\left(\frac{D_c^{(t)}}{\overline{D}^{(t)}} - 1, -1, 1 \right), \tag{16}$$

For classes with $M_c^{(t)} = 0$, we set $d_c^{(t)} = 0$. Positive values identify above-average upper path risk, caused by earlier observed failures, limited eligible support, or both; clipping prevents this refinement from overwhelming the endpoint anchor. If $\mathcal{C}_{\text{obs}}^{(t)}$ is empty, every temporal adjustment is neutral and the update uses endpoint evidence alone.

The complete class risk is

$$\begin{aligned} R_c^{(t)} &= U_c^{\text{end},(t)} \exp\left(\lambda \omega_t d_c^{(t)} \right), \\ \log R_c^{(t)} &= \log U_c^{\text{end},(t)} + \lambda \omega_t d_c^{(t)}. \end{aligned} \tag{17}$$

The second line makes the integration explicit: URA contributes the log-risk anchor $\log U_c^{\text{end},(t)}$, while FTR contributes the bounded residual $\lambda \omega_t d_c^{(t)}$. Here, $\lambda \geq 0$ is the sole AURORA-specific coefficient. The exponential map keeps risk positive. $\omega_t \in (0, 1]$ denotes the base perturbation warm-up used by the base recipe.

$$\omega_t = \min\left(\frac{t-1}{\rho T}, 1 \right), \tag{18}$$

where T is the number of epochs and $\rho \in (0, 1]$.

The next-epoch class weights are obtained by mean normalization:

$$w_c^{(t+1)} = \frac{R_c^{(t)}}{C^{-1} \sum_{j=1}^{C} R_j^{(t)}}. \tag{19}$$

Eq. (19) has class mean one, while Eq. (2) restores this property over the classes present in each mini-batch. Statistics from epoch t are detached before producing $w_c^{(t+1)}$; the model therefore cannot alter its current weight by differentiating through the estimator.

Stability of the Allocation Rule

The allocation is designed to refine endpoint risk. The following propositions formalize these guarantees; Appendix provides complete proofs and the Wilson-operator derivation.

Proposition 1 (Exact reduction to URA). *At $\lambda = 0$, the FTR residual vanishes and AURORA reduces exactly to URA: $w_c^{(t+1)} = U_c^{\text{end},(t)} / (C^{-1} \sum_j U_j^{\text{end},(t)})$.*

Proposition 2 (Bounded temporal refinement). *For $\lambda \geq 0$, $d_c^{(t)} \in [-1, 1]$, and $\omega_t \in [0, 1]$,*

$$e^{-\lambda} \leq \frac{R_c^{(t)}}{U_c^{\text{end},(t)}} \leq e^{\lambda}. \tag{20}$$

For any two classes c and j, the temporal refinement changes the ratio $R_c^{(t)}/R_j^{(t)}$ relative to the endpoint ratio $U_c^{\text{end},(t)}/U_j^{\text{end},(t)}$ by at most a factor of $e^{2\lambda}$.

Algorithm 1 AURORA: Uncertainty-Aware Risk Allocation

Input: training set $\mathcal{D}$, base learner $\mathcal{B}$, attack length K, coefficient λ, and epochs T.
Initialize class weights $w_c^{(1)} = 1$ for all classes.
for $t = 1, \ldots, T$ **do**
  Reset the endpoint and trajectory statistics for all classes.
  for each mini-batch (x, y) **do**
    Normalize the current class weights within the mini-batch using Eq. (2).
    Generate the original K-step attack trajectories using the inner maximization of $\mathcal{B}$.
    Record endpoint failures, clean eligibility, and first-passage events using Eqs. (4)–(6).
    Update the model using the weighted robust objective in Eq. (1).
  end for
  Compute the URA endpoint anchors $U_c^{\mathrm{end},(t)}$ using Eq. (9).
  Compute the FTR path statistics $D_c^{(t)}$ using Eqs. (12) and (13).
  Combine URA and FTR into class risks $R_c^{(t)}$ using Eq. (17).
  Normalize $R_c^{(t)}$ to obtain the next-epoch weights $w_c^{(t+1)}$ using Eq. (19).
end for
Return: trained model f_θ.

Proposition 3 (Two-level conservation)**.** *The class and mini-batch normalizations satisfy*

$$\frac{1}{C}\sum_{c=1}^{C} w_c^{(t+1)} = 1, \qquad \frac{1}{B}\sum_{i=1}^{B} \widetilde{w}_i^{(t)} = 1. \tag{21}$$

They preserve all pairwise weight ratios within their respective normalization domains.

Proposition 4 (Attack-time ordering)**.** *Consider classes c and j with equal endpoint anchors and positive eligible support. If $F_{c,k}^{+,(t)} \geq F_{j,k}^{+,(t)}$ at every attack step, with strict inequality at some step, then $D_c^{(t)} > D_j^{(t)}$. If the pre-clipping residuals of both classes lie strictly within $[-1, 1]$ and $\lambda\omega_t > 0$, then $w_c^{(t+1)} > w_j^{(t+1)}$.*

Proposition 5 (Well-posedness)**.** *For $z > 0$, $\lambda \geq 0$, and every nonempty mini-batch, all endpoint risks, class weights, and batch weights are finite and strictly positive. If no class supplies a clean-eligible trajectory in an epoch, the FTR adjustment becomes the identity and falls back to URA.*

4 EXPERIMENTS

Table 1: Comparative results on ResNet. Dedicated LT-AT methods are typically built upon TRADES. Light gray cells indicate the best result within the same base learner, and dark gray bold cells indicate the best result among all methods.

Dataset	Metric	Dedicated LT-AT				PGD-AT						TRADES						AWP					
		REAT	TAET	RoBal	BSL	Base	+UDR	+CFA	+DAFA	+RLT	+Ours	Base	+UDR	+CFA	+DAFA	+RLT	+Ours	Base	+UDR	+CFA	+DAFA	+RLT	+Ours
C10-LT	Nat.(all)	64.8%	66.0%	**72.9%**	64.9%	55.7%	57.1%	55.6%	64.1%	55.9%	71.5%	53.8%	54.8%	53.3%	62.8%	55.1%	68.2%	49.4%	42.9%	53.0%	54.1%	57.7%	64.8%
	Nat.(tail)	57.1%	58.7%	**68.7%**	57.1%	45.6%	47.2%	45.4%	56.9%	45.8%	68.3%	43.1%	44.2%	42.5%	55.7%	44.8%	63.2%	37.6%	29.6%	42.0%	44.5%	44.4%	65.2%
	PGD(all)	33.0%	33.1%	35.4%	33.5%	26.4%	29.5%	29.3%	30.5%	30.7%	36.7%	30.2%	29.7%	30.9%	31.6%	31.9%	37.5%	29.3%	28.5%	32.4%	32.6%	36.7%	**39.9%**
	PGD(tail)	21.0%	21.4%	27.9%	21.6%	13.1%	16.7%	15.9%	21.3%	18.5%	32.5%	16.5%	16.2%	17.0%	21.9%	19.2%	31.1%	15.3%	14.7%	18.7%	22.6%	27.1%	**40.4%**
	C&W(all)	31.9%	31.7%	34.5%	32.6%	29.6%	29.3%	29.0%	31.4%	30.7%	**37.1%**	29.4%	29.4%	29.9%	30.4%	30.8%	35.4%	28.0%	27.7%	31.3%	31.5%	30.1%	36.9%
	C&W(tail)	19.9%	19.8%	26.9%	20.4%	16.1%	16.4%	15.3%	22.6%	18.5%	34.8%	15.7%	15.7%	15.8%	20.7%	18.1%	28.4%	13.8%	13.8%	17.3%	21.2%	16.3%	**36.9%**
	AA(all)	29.9%	29.9%	32.1%	30.5%	27.5%	27.2%	27.3%	28.5%	27.9%	**35.1%**	28.2%	27.7%	29.0%	28.7%	29.5%	33.4%	27.2%	26.4%	29.8%	29.6%	28.5%	34.7%
	AA(tail)	17.8%	17.9%	24.3%	18.3%	13.8%	14.2%	13.5%	19.6%	15.5%	32.7%	14.5%	14.1%	14.8%	19.1%	16.8%	26.5%	12.9%	12.4%	15.7%	19.5%	14.5%	**34.5%**
C100-LT	Nat.(all)	46.2%	45.9%	45.6%	46.5%	43.1%	42.4%	43.0%	44.2%	45.7%	48.2%	43.7%	41.3%	43.4%	44.7%	45.2%	48.6%	44.8%	42.5%	46.8%	44.2%	45.7%	**50.3%**
	Nat.(tail)	43.4%	42.6%	44.9%	43.8%	38.6%	38.1%	38.4%	40.4%	41.7%	46.8%	38.7%	36.4%	37.9%	40.8%	40.6%	46.9%	39.4%	37.2%	41.8%	38.4%	39.7%	**50.1%**
	PGD(all)	20.8%	19.1%	19.9%	20.7%	15.4%	16.3%	17.7%	16.8%	17.9%	18.0%	20.1%	18.8%	20.8%	19.8%	20.8%	21.3%	21.4%	20.8%	22.0%	22.9%	23.0%	**23.4%**
	PGD(tail)	19.6%	17.6%	19.8%	19.4%	13.3%	14.5%	15.6%	15.6%	16.3%	17.6%	17.6%	16.4%	17.7%	18.3%	18.6%	21.0%	18.2%	18.1%	19.2%	20.4%	19.9%	**23.7%**
	C&W(all)	18.9%	17.4%	17.8%	19.0%	17.7%	16.3%	17.2%	16.5%	17.6%	17.3%	18.4%	17.1%	18.8%	17.8%	18.9%	19.1%	18.9%	18.9%	20.2%	20.7%	**21.0%**	**21.0%**
	C&W(tail)	17.8%	16.0%	17.8%	17.6%	15.4%	14.5%	15.1%	15.5%	15.9%	17.4%	15.9%	15.1%	16.2%	16.7%	17.1%	18.9%	16.2%	16.2%	17.4%	18.3%	18.8%	**21.2%**
	AA(all)	17.3%	15.9%	16.6%	17.3%	16.0%	14.7%	15.5%	14.7%	15.5%	15.4%	17.3%	15.8%	17.5%	16.5%	17.4%	17.5%	17.5%	17.0%	18.2%	18.9%	**19.3%**	19.1%
	AA(tail)	16.2%	14.7%	16.6%	16.2%	13.8%	13.2%	13.7%	14.0%	14.1%	15.5%	15.0%	13.9%	15.1%	15.4%	15.6%	17.3%	15.0%	14.7%	15.7%	16.8%	17.0%	**19.3%**
T-LT	Nat.(all)	37.8%	37.7%	37.2%	38.9%	34.5%	34.6%	36.9%	36.4%	37.3%	39.3%	35.9%	34.0%	37.6%	36.1%	36.3%	38.7%	33.9%	30.1%	38.4%	34.9%	34.4%	**39.6%**
	Nat.(tail)	36.0%	34.8%	36.1%	37.0%	29.8%	29.6%	32.1%	32.2%	32.8%	37.5%	30.8%	28.9%	32.4%	32.0%	31.1%	37.0%	28.0%	24.6%	32.7%	29.4%	28.1%	**38.8%**
	PGD(all)	13.6%	12.1%	13.5%	14.2%	9.0%	12.6%	12.2%	12.2%	12.6%	12.9%	13.8%	12.5%	14.9%	13.4%	13.7%	14.1%	16.7%	13.5%	16.0%	16.9%	16.5%	**18.1%**
	PGD(tail)	12.8%	10.7%	12.8%	13.2%	7.2%	9.9%	9.8%	10.4%	10.4%	11.6%	10.9%	9.8%	11.7%	11.4%	11.2%	13.3%	12.9%	10.3%	12.6%	13.8%	12.6%	**17.6%**
	C&W(all)	11.2%	10.2%	10.7%	11.7%	12.3%	11.5%	11.6%	11.8%	12.2%	12.3%	11.7%	10.1%	12.0%	10.7%	11.2%	11.4%	14.6%	11.7%	14.1%	14.5%	14.8%	**15.1%**
	C&W(tail)	10.3%	9.0%	10.0%	10.7%	9.6%	8.8%	9.4%	9.9%	9.8%	11.2%	9.2%	8.0%	9.4%	9.1%	9.0%	10.6%	11.3%	8.9%	11.0%	11.7%	11.3%	**14.4%**
	AA(all)	10.3%	9.3%	9.7%	10.6%	10.9%	10.2%	10.4%	10.4%	11.0%	10.8%	10.8%	9.4%	11.3%	9.8%	10.3%	10.3%	13.0%	10.4%	12.5%	13.2%	13.2%	**13.3%**
	AA(tail)	9.5%	8.3%	9.0%	9.7%	8.6%	7.8%	8.3%	8.8%	9.0%	10.3%	8.4%	7.5%	8.8%	8.3%	8.3%	9.6%	10.0%	7.7%	9.8%	10.7%	10.0%	**12.7%**

Table 2: Comparative results on PreActResNet-18.

Dataset	Metric	Dedicated LT-AT				PGD-AT						TRADES						AWP					
		REAT	TAET	RoBal	BSL	Base	+UDR	+CFA	+DAFA	+RLT	+Ours	Base	+UDR	+CFA	+DAFA	+RLT	+Ours	Base	+UDR	+CFA	+DAFA	+RLT	+Ours
C10-LT	Nat.(all)	65.5%	65.7%	**72.6%**	65.2%	57.9%	60.0%	57.5%	65.8%	57.4%	72.2%	56.5%	54.9%	54.4%	63.9%	57.2%	68.2%	50.8%	50.3%	55.3%	57.2%	45.3%	68.0%
	Nat.(tail)	57.9%	58.3%	68.6%	57.5%	48.3%	50.9%	47.6%	59.0%	47.6%	**69.1%**	46.5%	44.4%	43.8%	57.2%	47.5%	63.2%	39.3%	38.5%	44.9%	48.1%	32.3%	67.9%
	PGD(all)	33.0%	32.6%	36.1%	33.2%	26.4%	29.9%	29.8%	30.5%	31.0%	35.7%	31.3%	30.4%	30.7%	31.9%	31.9%	37.5%	29.9%	31.2%	33.0%	33.4%	31.4%	**40.5%**
	PGD(tail)	20.9%	20.8%	29.1%	21.1%	13.1%	17.0%	16.4%	21.5%	19.0%	31.4%	18.1%	16.8%	16.7%	22.5%	19.5%	30.7%	15.9%	17.6%	19.4%	23.6%	17.7%	**40.6%**
	C&W(all)	31.8%	31.8%	35.0%	32.2%	30.6%	30.0%	30.1%	31.2%	31.0%	37.7%	30.3%	29.7%	29.8%	30.7%	31.0%	35.3%	28.5%	30.5%	32.1%	32.6%	30.7%	**38.0%**
	C&W(tail)	19.7%	19.7%	27.8%	20.1%	17.4%	16.9%	16.4%	22.2%	18.4%	35.0%	17.0%	16.1%	15.7%	21.2%	18.6%	28.3%	14.5%	16.9%	18.4%	22.5%	17.0%	**37.8%**
	AA(all)	29.7%	29.8%	32.7%	29.8%	28.0%	27.6%	27.9%	28.5%	28.5%	35.2%	29.1%	28.4%	28.6%	29.1%	29.6%	33.1%	27.6%	28.7%	30.5%	30.3%	29.2%	**35.4%**
	AA(tail)	17.8%	17.7%	25.3%	17.7%	14.5%	14.4%	14.1%	19.4%	16.0%	32.4%	15.8%	14.9%	14.4%	19.6%	17.2%	26.1%	13.4%	14.9%	16.6%	20.3%	15.4%	**35.0%**
C100-LT	Nat.(all)	46.5%	45.9%	45.1%	46.8%	43.1%	42.2%	43.2%	44.3%	46.7%	48.1%	43.3%	41.0%	43.9%	44.4%	45.5%	48.7%	45.2%	46.0%	46.9%	44.3%	45.8%	**50.4%**
	Nat.(tail)	43.7%	42.6%	44.3%	43.7%	38.8%	37.9%	38.6%	40.7%	42.7%	46.4%	38.5%	35.9%	38.5%	40.6%	40.7%	47.0%	39.8%	41.1%	41.6%	39.1%	39.9%	**49.7%**
	PGD(all)	21.0%	19.4%	20.2%	20.8%	15.5%	15.6%	17.8%	16.9%	18.1%	18.5%	20.4%	18.7%	20.9%	19.9%	20.9%	21.7%	21.8%	21.6%	21.8%	23.1%	23.2%	**24.1%**
	PGD(tail)	19.7%	18.0%	20.1%	19.4%	13.8%	13.7%	15.8%	15.7%	16.5%	18.3%	17.8%	16.2%	17.8%	18.5%	18.9%	21.4%	18.6%	18.7%	19.0%	20.6%	20.3%	**23.7%**
	C&W(all)	19.0%	17.7%	17.6%	19.1%	17.7%	15.7%	17.2%	16.9%	18.0%	18.2%	18.5%	17.1%	18.8%	17.7%	19.3%	19.6%	19.3%	20.0%	18.4%	21.1%	21.3%	**22.3%**
	C&W(tail)	17.8%	16.5%	17.5%	17.9%	15.6%	13.8%	15.1%	15.9%	16.3%	17.8%	16.4%	15.0%	16.4%	16.8%	17.6%	19.1%	16.7%	17.4%	17.5%	19.0%	18.8%	**21.8%**
	AA(all)	17.3%	16.4%	16.4%	17.6%	16.3%	14.2%	15.7%	15.3%	16.2%	16.0%	17.1%	15.8%	17.5%	16.4%	17.8%	17.8%	17.9%	18.1%	17.7%	19.2%	19.2%	**19.8%**
	AA(tail)	16.2%	15.3%	16.3%	16.5%	14.5%	12.6%	13.8%	14.4%	14.7%	15.9%	15.1%	14.0%	15.3%	15.5%	16.2%	17.6%	15.5%	15.8%	16.0%	17.4%	16.9%	**19.5%**
T-LT	Nat.(all)	38.3%	37.5%	36.4%	38.6%	35.1%	34.5%	37.0%	36.7%	37.3%	39.5%	35.9%	34.2%	37.6%	36.5%	36.8%	39.6%	34.5%	34.0%	38.8%	35.3%	34.6%	**40.2%**
	Nat.(tail)	36.6%	34.7%	35.3%	36.8%	30.8%	29.9%	32.2%	32.6%	32.5%	37.9%	30.7%	29.0%	32.2%	32.4%	31.7%	38.1%	28.6%	28.4%	33.1%	29.9%	28.2%	**39.3%**
	PGD(all)	13.9%	12.4%	13.1%	14.3%	9.1%	10.5%	12.2%	12.1%	12.8%	12.7%	13.8%	12.3%	15.0%	13.2%	14.2%	14.2%	17.0%	15.1%	16.3%	17.1%	16.2%	**18.3%**
	PGD(tail)	13.1%	10.9%	12.3%	13.4%	7.3%	8.4%	9.8%	10.3%	10.6%	12.0%	10.9%	9.5%	11.6%	11.3%	11.4%	13.6%	13.4%	11.8%	12.8%	14.1%	12.8%	**17.8%**
	C&W(all)	11.4%	10.6%	10.7%	11.8%	12.4%	11.4%	11.9%	12.5%	12.9%	12.5%	11.7%	10.1%	11.8%	10.9%	11.7%	13.2%	14.9%	13.5%	14.4%	14.6%	14.4%	**15.4%**
	C&W(tail)	10.6%	9.3%	10.1%	10.8%	9.6%	9.0%	9.9%	10.6%	10.1%	11.7%	9.4%	7.9%	9.2%	9.4%	9.4%	12.1%	11.7%	10.4%	11.4%	11.7%	10.8%	**14.8%**
	AA(all)	10.3%	9.6%	9.9%	10.7%	11.2%	10.1%	10.6%	11.2%	11.4%	11.2%	10.9%	9.2%	11.3%	10.1%	10.6%	11.5%	13.5%	12.1%	12.6%	13.0%	12.9%	**13.7%**
	AA(tail)	9.5%	8.5%	9.3%	9.8%	8.6%	8.1%	8.3%	9.5%	9.0%	10.3%	8.8%	7.2%	8.8%	8.6%	8.4%	10.4%	10.5%	9.3%	9.9%	10.5%	9.7%	**13.1%**

Experimental settings

Benchmarks. Following the standard exponential construction (Cao et al., 2019), we derive C10-LT, C100-LT, and T-LT from CIFAR10, CIFAR100 (Krizhevsky et al., 2009), and TinyImageNet (Le et al., 2015). The imbalance ratio is 50 for C10-LT and 10 for C100-LT and T-LT. Test sets retain their original balanced class distributions. We report results with ResNet18, PreActResNet (He et al., 2016), and WideResNet (Zagoruyko & Komodakis, 2016).

Training and competitors. We attach the same AURORA implementation to PGD-AT (Madry et al., 2018), TRADES (Zhang et al., 2019), and AWP (Wu et al., 2020). We compare against dedicated LT-AT methods, including RoBal, REAT, BSL, and TAET, and plug-in methods, including UDR, CFA, DAFA, and RLT. Plug-in results are compared under matched base learners, isolating the allocation strategy from the underlying adversarial objective.

Metrics and attacks. We report natural and robust accuracy over all classes and the 80% least frequent classes, denoted by Nat.(all), Rob.(all), Nat.(tail), and Rob.(tail). The PGD evaluation uses 20-step ℓ_∞ with radius $8/255$ and step size $2/255$. We also report C&W (Carlini & Wagner, 2017) and AutoAttack (Croce & Hein, 2020) to evaluate robustness.

Reproducibility details. All models are trained for 110 epochs using SGD with momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.1. Training attacks use $K = 10$ steps. The shared base recipe uses prior correction with $\tau = 1.0$ and the class-dependent perturbation schedule of RLT (Zhang et al., 2026) with coefficient 0.3; these settings remain unchanged in all estimator ablations. We set $z = 1.96$, reuse the base warm-up fraction $\rho = 0.8$, and fix the only AURORA-specific coefficient to $\lambda = 0.8$ across all settings. The clipping interval in Eq. (16) is fixed to $[-1, 1]$. Statistics are computed exclusively from training mini-batches, detached, and applied in the following epoch; neither validation nor test data enter the allocation rule.

Comparative Results

Tables 1–3 compare AURORA with dedicated LT-AT methods and plug-in baselines across different datasets, architectures, and base learners.

Tail-class performance. AURORA achieves the highest base-matched Nat.(tail), PGD(tail), C&W(tail), and AA(tail) across all 27 combinations of datasets, architectures, and base learners. The improvements are particularly clear on C10-LT. With ResNet, AURORA achieves PGD(tail) accuracies of 32.5%, 31.1%, and 40.4% with PGD-AT, TRADES, and AWP, respectively, compared with 21.3%, 21.9%, and 27.1% for the strongest competing plug-ins. The corresponding improvements are 11.2, 9.2, and 13.3 percentage points. A similar advantage is observed under AutoAttack, where AURORA improves AA(tail) from 19.6%, 19.1%, and 19.5% to 32.7%, 26.5%, and 34.5%, respectively. Meanwhile, its Nat.(tail) accuracies reach 68.3%, 63.2%, and 65.2%, suggesting that the robust-tail gains are not obtained by substantially weakening clean tail recognition.

Table 3: Comparative results on WideResNet (WRN-28-10). Light gray cells indicate the best result within the same base learner, while dark gray bold cells indicate the best result among all methods.

Dataset	Metric	Dedicated LT-AT		PGD-AT					TRADES					AWP				
		RoBal	BSL	Base	+UDR	+DAFA	+RLT	+Ours	Base	+UDR	+DAFA	+RLT	+Ours	Base	+UDR	+DAFA	+RLT	+Ours
C10-LT	Nat.(all)	72.5%	67.7%	58.9%	59.4%	65.1%	60.7%	**74.1%**	59.0%	58.7%	65.6%	60.5%	70.1%	53.4%	57.9%	63.3%	68.9%	72.1%
	Nat.(tail)	67.7%	60.6%	49.3%	50.1%	57.4%	51.6%	70.0%	49.5%	49.1%	58.5%	51.4%	64.7%	42.4%	47.9%	55.3%	56.6%	**71.7%**
	PGD(all)	34.6%	30.6%	26.9%	28.1%	29.6%	31.5%	36.1%	32.4%	31.4%	33.5%	33.6%	39.0%	31.2%	32.7%	35.9%	38.7%	**43.3%**
	PGD(tail)	25.8%	19.1%	13.1%	15.0%	18.7%	19.1%	29.5%	19.1%	18.1%	22.9%	21.0%	31.1%	17.0%	19.5%	25.9%	27.1%	**43.1%**
	C&W(all)	35.5%	32.6%	31.2%	28.9%	31.1%	32.7%	36.2%	31.8%	30.9%	33.0%	32.7%	37.6%	30.2%	32.4%	35.4%	37.1%	**40.9%**
	C&W(tail)	27.8%	25.8%	18.0%	15.9%	21.0%	20.5%	31.7%	18.4%	17.5%	22.5%	20.0%	29.2%	15.8%	19.1%	25.1%	28.4%	**40.4%**
	AA(all)	32.3%	29.6%	28.9%	27.0%	28.7%	30.0%	33.3%	30.8%	29.5%	31.2%	31.2%	35.1%	29.2%	29.9%	33.0%	35.0%	**38.2%**
	AA(tail)	24.5%	22.6%	15.5%	13.8%	18.6%	17.8%	28.7%	17.3%	16.2%	20.7%	18.4%	26.8%	14.8%	16.4%	22.8%	24.7%	**37.3%**
C100-LT	Nat.(all)	48.8%	47.9%	46.2%	45.5%	46.9%	49.1%	50.8%	47.6%	45.0%	48.5%	50.0%	52.9%	50.0%	50.7%	49.8%	50.8%	**54.6%**
	Nat.(tail)	47.4%	45.1%	41.7%	40.9%	43.2%	44.6%	48.9%	42.2%	39.9%	44.8%	45.3%	50.5%	44.6%	45.7%	45.1%	45.2%	**54.2%**
	PGD(all)	21.9%	20.1%	17.2%	17.3%	18.1%	18.6%	19.6%	22.1%	20.3%	22.0%	23.2%	23.6%	24.8%	22.9%	25.1%	**25.3%**	24.9%
	PGD(tail)	21.2%	18.8%	15.0%	15.2%	16.9%	16.7%	19.3%	19.6%	17.9%	20.5%	21.1%	22.9%	21.3%	20.3%	22.8%	22.6%	**25.5%**
	C&W(all)	19.4%	19.2%	19.5%	17.5%	18.5%	18.9%	19.4%	20.8%	19.0%	20.4%	21.6%	22.1%	22.1%	21.8%	23.5%	23.8%	**24.1%**
	C&W(tail)	19.1%	18.3%	17.4%	15.6%	17.7%	17.3%	19.2%	18.4%	16.8%	19.2%	19.6%	21.3%	19.1%	19.3%	21.3%	21.2%	**24.5%**
	AA(all)	18.0%	17.3%	17.4%	15.8%	16.8%	16.9%	17.4%	19.4%	17.5%	19.0%	20.2%	20.4%	20.7%	20.0%	21.4%	21.5%	**21.8%**
	AA(tail)	17.7%	16.6%	15.5%	14.1%	16.1%	15.5%	17.2%	17.1%	15.5%	18.0%	18.3%	19.6%	18.0%	17.8%	19.5%	19.2%	**22.2%**
T-LT	Nat.(all)	41.3%	41.7%	37.7%	39.2%	39.8%	41.7%	43.4%	38.8%	40.9%	39.2%	39.8%	42.9%	39.1%	41.3%	35.1%	38.0%	**43.5%**
	Nat.(tail)	40.1%	36.9%	33.5%	34.5%	35.9%	36.9%	41.7%	33.6%	35.9%	34.9%	34.8%	41.4%	32.9%	35.7%	29.7%	31.3%	**42.7%**
	PGD(all)	16.1%	14.7%	10.5%	12.8%	13.8%	14.7%	13.8%	16.1%	13.7%	15.9%	16.8%	16.7%	14.9%	12.7%	17.0%	19.3%	**20.5%**
	PGD(tail)	15.1%	12.2%	8.4%	10.3%	11.8%	12.2%	13.1%	12.7%	11.2%	13.5%	13.7%	16.1%	11.3%	10.7%	14.0%	15.3%	**20.2%**
	C&W(all)	13.0%	15.1%	14.4%	13.3%	14.1%	15.0%	14.2%	13.8%	12.2%	13.3%	14.2%	14.0%	12.5%	10.5%	16.5%	17.1%	**17.2%**
	C&W(tail)	12.1%	13.5%	11.6%	10.6%	12.1%	12.1%	13.4%	10.9%	10.0%	11.2%	11.5%	13.1%	9.4%	8.4%	13.1%	13.4%	**16.4%**
	AA(all)	12.1%	13.3%	12.7%	11.9%	12.9%	13.2%	12.5%	12.9%	11.1%	12.3%	13.1%	12.7%	11.2%	9.8%	11.7%	15.5%	**15.7%**
	AA(tail)	11.3%	12.0%	10.2%	9.5%	11.1%	11.0%	11.8%	10.2%	9.1%	10.3%	10.5%	11.7%	8.3%	7.6%	8.9%	12.2%	**15.0%**

The tail-class improvements remain visible on C100-LT and T-LT. On C100-LT with WideResNet and AWP, AURORA achieves 25.5%, 24.5%, and 22.2% tail accuracy under PGD, C&W, and AutoAttack, exceeding the strongest base-matched alternatives by 2.7, 3.2, and 2.7 points. On T-LT, its PGD(tail) improvements over the strongest AWP-based competitors are 3.8, 3.7, and 4.9 points with ResNet, PreActResNet, and WideResNet, respectively. These results suggest that AURORA consistently benefits tail classes across datasets and attack types.

Overall performance. Beyond the tail-class gains, AURORA generally maintains a favorable trade-off over the full test distribution. On C10-LT with ResNet, it improves PGD(all) over the strongest base-matched competitor by 6.0, 5.6, and 3.2 points under PGD-AT, TRADES, and AWP, respectively, while also increasing Nat.(all) by 7.4, 5.4, and 7.1 points. The improvements on C100-LT and T-LT follow a similar trend, suggesting that the benefit is not confined to a single dataset.

Performance consistency. These gains are consistent across training objectives, architectures, and evaluation attacks (PGD, C&W, AutoAttack). In particular, AWP+AURORA attains the best reported result on every T-LT metric.

Ablation Studies

Comparison of robust-weight estimators. Table 4 examines whether the gains arise from class-wise reweighting in general or from the risk information used to construct the weights. All variants share the same settings and differ only in the class statistic used to determine w_c. Let

$$\widehat{p}_c^{\,\mathrm{end}} = \frac{E_c}{N_c}, \qquad a_c^{\mathrm{rob}} = 1 - \widehat{p}_c^{\,\mathrm{end}}, \tag{22}$$

where E_c and N_c denote the endpoint failures and observations of class c, respectively. The compared weighting rules are defined as

$$\begin{aligned} &s_c^{\mathrm{uni}} = 1, \qquad s_c^{\mathrm{end}} = \widehat{p}_c^{\,\mathrm{end}}, \\ &b_c^{\mathrm{CFA}} = \beta\left(\lambda_2 + a_c^{\mathrm{rob}}\right), \qquad s_c^{\mathrm{CFA}} = \frac{b_c^{\mathrm{CFA}}}{1 + b_c^{\mathrm{CFA}}}, \\ &s_c^{\mathrm{URA}} = U_c^{\mathrm{end}}, \qquad s_c^{\mathrm{AURORA}} = U_c^{\mathrm{end}} \exp(\lambda \omega_t d_c)\,. \end{aligned} \tag{23}$$

Each score is converted into a unit-mean class weight.

Table 4 evaluates how different class-wise weighting strategies affect long-tailed adversarial training. Directly weighting the robust loss by endpoint failure yields modest robustness gains over uniform allocation, improving PGD(all) and PGD(tail) by 0.6 and 1.3 points, respectively, while slightly

Table 4: Comparison of class-wise robust-weight estimators on C10-LT with ResNet. All variants use the same training configuration and differ only in the statistic used to construct w_c.

Method	URA	FTR	Nat.(all)	PGD(all)	Nat.(tail)	PGD(tail)
Uniform Allocation	×	×	63.7	31.9	57.9	22.8
Endpoint Failure	×	×	63.2	32.5	58.3	24.1
CFA-style Weight Feedback	×	×	64.1	33.1	58.8	25.0
URA	✓	×	67.4	36.7	62.0	28.4
AURORA	✓	✓	**68.2**	**37.5**	**63.2**	**31.1**

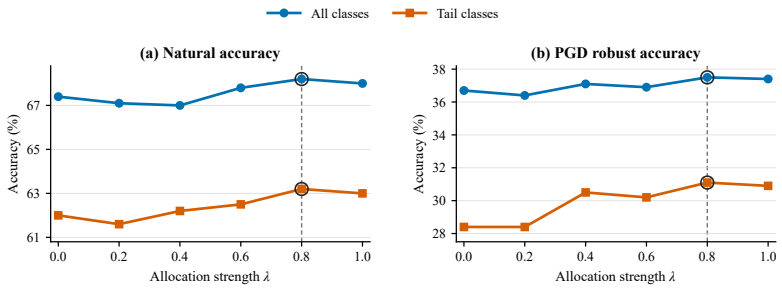

Figure 1: Ablation study of the AURORA-specific coefficient λ in Eq. (17). λ controls the contribution of FTR relative to the URA endpoint anchor.

reducing Nat.(all) by 0.5 points. CFA-style feedback performs more favorably, improving the four metrics by 0.9, 1.3, 0.8, and 2.1 points over uniform allocation. URA provides a clearer improvement. Compared with CFA-style feedback, it raises Nat.(all), PGD(all), Nat.(tail), and PGD(tail) by 3.3, 3.6, 3.2, and 3.4 points, suggesting that accounting for unequal class support is important when estimating class-wise risk. FTR further improves PGD(tail) from 28.4% to 31.1%, while PGD(all) increases from 36.7% to 37.5%. This tail gain is consistent with first-passage information providing a useful refinement for tail classes. Overall, the results support the complementary roles of support-aware endpoint estimation and trajectory-based refinement in constructing robust-loss weights.

Effect of trajectory-refinement strength. Figure 1 studies the AURORA-specific coefficient λ. At $\lambda = 0$, AURORA reduces to URA and achieves 28.4% PGD(tail). Increasing λ strengthens the first-passage correction, raising PGD(tail) to 31.1% at $\lambda = 0.8$. The relatively stable all-class performance indicates that the bounded correction primarily improves tail-risk allocation without disrupting the endpoint anchor. Increasing λ further to 1.0 produces a slight decline across all metrics, suggesting that excessive trajectory refinement can overemphasize failure timing.

5 Conclusion

We introduced AURORA, a risk-allocation framework that uses information already produced by iterative adversarial training. URA converts endpoint failures and class support into an upper-confidence risk estimate, while FTR refines this estimate using the first-failure behavior of initially correct samples. Their bounded composition reweights the native robust objective without introducing an auxiliary loss. Across the evaluated long-tailed benchmarks, AURORA provides consistent gains in tail robustness and remains competitive on overall natural and robust accuracy, showing that attack trajectories contain useful statistical evidence beyond their final adversarial examples.

AI Use Statement. Generative AI assisted with code development and language editing. All AI-assisted content was verified and revised by the authors.

Ethics Statement. This work uses public benchmarks and involves no human subjects or personal data. The authors adhere to the ICLR Code of Ethics.

Reproducibility Statement. The method, experimental settings, and evaluation protocols are documented in the paper, with code provided in the supplementary material.

REFERENCES

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6897–6907, 2022.

Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Phung. A unified wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in neural information processing systems*, volume 32, 2019.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Seungju Cho, Hongsin Lee, and Changick Kim. Long-tailed adversarial training with self-distillation. In *International Conference on Learning Representations*, volume 2025, pp. 35954–35972, 2025.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Calian, and Timothy Mann. Improving robustness using generated data. In *Advances in Neural Information Processing Systems*, 2021.

Naoya Hasegawa and Issei Sato. Exploring weight balancing on long-tailed recognition problem. In *International Conference on Learning Representations*, volume 2024, pp. 40705–40732, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645, 2016.

Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13419–13429, 2022.

Chengkai Hou, Jieyu Zhang, Haonan Wang, and Tianyi Zhou. Subclass-balancing contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5395–5407, 2023.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Hyungyu Lee, Saehyung Lee, Hyemi Jang, Junsung Park, Ho Bae, and Sungroh Yoon. Dafa: Distance-aware fair adversarial training. In *International Conference on Learning Representations*, volume 2024, pp. 28780–28809, 2024.

Guanlin Li, Guowen Xu, and Tianwei Zhang. Alleviating the effect of data imbalance on adversarial training, 2023.

Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6929–6938, 2022a.

Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. Subspace adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13409–13418, 2022b.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S. Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6918–6928, 2022c.

Fudong Lin and Xu Yuan. Long-tailed recognition via information-preservable two-stage learning. *Advances in Neural Information Processing Systems*, 38:47035–47064, 2026.

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

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6887–6896, 2022.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18434–18443, 2023.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. In *Advances in Neural Information Processing Systems*, 2021.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, pp. 4175–4186, 2020.

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. Balanced product of calibrated experts for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19967–19977, 2023.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. In *Advances in Neural Information Processing Systems*, 2024.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, volume 2025, pp. 26681–26697, 2025.

Binwu Wang, Pengkun Wang, Wei Xu, Xu Wang, Yudong Zhang, Kun Wang, and Yang Wang. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *International Conference on Learning Representations*, volume 2024, pp. 19402–19425, 2024.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021.

Zeyu Wang, Tianyu Pang, Chao Du, Min Lin, and Weiwei Liu. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 2023.

Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8193–8201, 2023.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, pp. 2958–2969, 2020.

Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8655–8664. IEEE, 2021.

Lingjie Yi, Michael Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *International Conference on Learning Representations*, 2025.

Wang Yu-Hang, Junkang Guo, Aolei Liu, Kaihao Wang, Zaitong Wu, Zhenyu Liu, Wenfei Yin, and Jian Liu. Taet: Two-stage adversarial equalization training on long-tailed distributions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15476–15485. IEEE, 2025.

Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao. Revisiting adversarial training under long-tailed distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24492–24501, 2024.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

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

Lilin Zhang, Yimo Guo, Yue Li, Jiancheng Shi, and Xianggen Liu. Taming the long tail: Rebalancing adversarial training via adaptive perturbation, 2026.

Mengnan Zhao, Lihe Zhang, Jingwen Ye, Huchuan Lu, Baocai Yin, and Xinchao Wang. Adversarial training: A survey, 2024a.

Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19510–19520. IEEE, 2024b.

Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Pareto deep long-tailed recognition: A conflict-averse solution. In *International Conference on Learning Representations*, volume 2024, pp. 54500–54526, 2024.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6908–6917, 2022.

A Proofs and Additional Derivations

This appendix proves every deterministic property invoked by the method. These results concern the estimator and its optimization interface; they do not imply that AURORA must improve test accuracy. Generalization and robustness gains remain empirical questions addressed by the experiments.

NOTATION AND PROOF ROADMAP

We first make the dependency chain explicit. Fix an epoch t and temporarily suppress its superscript. For each class c, endpoint failures (E_c, N_c) determine the URA anchor U_c^{end}. Clean-eligible first-passage counts $(H_{c,k}, M_c)$ determine the upper cumulative curve $F_{c,k}^{+}$ and its area D_c.

$$\mathcal{U}(s,n) = \frac{\widehat{p} + \frac{z^2}{2n} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}. \tag{24}$$

$$M_c^{(t)} = \sum_i \mathbb{I}[y_i = c] a_i^{(t)}. \tag{25}$$

$$H_{c,k}^{(t)} = \sum_i \mathbb{I}[y_i = c] a_i^{(t)} \mathbb{I}[T_i^{(t)} \le k]. \tag{26}$$

$$F_{c,k}^{+,(t)} = \mathcal{U}\Big(H_{c,k}^{(t)}, M_c^{(t)}\Big), \qquad k = 1, \dots, K, \quad M_c^{(t)} > 0. \tag{27}$$

$$D_c^{(t)} = \frac{1}{K} \sum_{k=1}^{K} F_{c,k}^{+,(t)}. \tag{28}$$

The centered statistic d_c then refines the anchor in log space, producing R_c, and two positive normalizations map R_c to class and mini-batch weights:

$$\begin{aligned} (E_c, N_c) &\longmapsto U_c^{\mathrm{end}}, \\ (H_{c,1:K}, M_c) &\longmapsto F_{c,1:K}^{+} \longmapsto D_c \longmapsto d_c, \\ (U_c^{\mathrm{end}}, d_c) &\longmapsto R_c \longmapsto w_c \longmapsto \widetilde{w}_i. \end{aligned} \tag{29}$$

The proofs follow this order. We use the Wilson operator only when its denominator is positive. The method handles the two zero-support cases separately: $N_c = 0$ gives the conservative endpoint value $U_c^{\mathrm{end}} = 1$, while $M_c = 0$ gives the neutral temporal residual $d_c = 0$. These conventions are therefore not hidden divisions by zero.

DERIVATION OF THE WILSON UPPER OPERATOR

Lemma 1 (Score-interval inversion)**.** *Let s be the number of successes in $n > 0$ Bernoulli observations, $\widehat{p} = s/n$, and $z > 0$. Inverting the normal score acceptance region*

$$\frac{|\widehat{p} - p|}{\sqrt{p(1-p)/n}} \le z \tag{30}$$

over $p \in (0,1)$, and taking its closure in $[0,1]$, gives an interval $[p_-, p_+]$ whose upper endpoint is $\mathcal{U}(s,n)$ in Eq. (7) of the main text.

Proof. For $p \in (0,1)$, the denominator in Eq. (30) is positive. We may therefore square both sides without changing the feasible set. Multiplying by $p(1-p)/n$ and then by n gives

$$n(\widehat{p} - p)^2 \le z^2 p(1-p). \tag{31}$$

Expanding both sides,

$$n\widehat{p}^2 - 2n\widehat{p}p + np^2 \le z^2 p - z^2 p^2. \tag{32}$$

Collecting terms of equal degree in p produces

$$(n + z^2)p^2 - (2n\widehat{p} + z^2)p + n\widehat{p}^2 \le 0. \tag{33}$$

Let $a = n + z^2$, $b = -(2n\widehat{p} + z^2)$, and $c = n\widehat{p}^2$. Since $a > 0$, the quadratic is convex, so the inequality holds precisely between its two real roots. Its discriminant is non-negative because

$$\begin{aligned} \Delta &= b^2 - 4ac && (34) \\ &= (2n\widehat{p} + z^2)^2 - 4(n + z^2)n\widehat{p}^2 && (35) \\ &= 4n^2\widehat{p}^2 + 4nz^2\widehat{p} + z^4 - 4n^2\widehat{p}^2 - 4nz^2\widehat{p}^2 && (36) \\ &= 4nz^2\widehat{p}(1-\widehat{p}) + z^4 && (37) \\ &= 4n^2z^2\left(\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}\right) \ge 0. && (38) \end{aligned}$$

The quadratic formula now yields

$$p_{\pm} = \frac{2n\widehat{p} + z^2 \pm \sqrt{(2n\widehat{p} + z^2)^2 - 4(n + z^2)n\widehat{p}^2}}{2(n + z^2)}. \tag{39}$$

By Eq. (38),

$$\sqrt{\Delta} = 2nz\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}}, \tag{40}$$

where $z > 0$. Substituting this expression into the larger root p_+ and dividing numerator and denominator by $2n$ gives

$$p_+ = \frac{\widehat{p} + \frac{z^2}{2n} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \tag{41}$$

which is exactly $\mathcal{U}(s, n)$. The derivation was carried out on $p \in (0, 1)$. Taking the closure adds a boundary value only when a root equals 0 or 1, so the same formula remains valid for the boundary observations $s = 0$ and $s = n$.

The value $z = 1.96$ corresponds to the conventional two-sided normal score interval with nominal coverage 95%. We use only its upper endpoint because the allocation asks how large the class failure probability may plausibly be. This is a score-based confidence construction, not a claim of exact finite-sample coverage.

Lemma 2 (Range, optimism, and consistency)**.** *For $z > 0$, $n > 0$, and $0 \le s \le n$, the Wilson upper endpoint satisfies*

$$0 \le \widehat{p} \le \mathcal{U}(s, n) \le 1, \tag{42}$$

where $\mathcal{U}(s, n) > 0$, and the optimism inequality is strict unless $s = n$. For any sequence with $n \to \infty$ and $s/n \to p \in [0, 1]$, we further have $\mathcal{U}(s, n) \to p$ and $\mathcal{U}(s, n) - \widehat{p} \to 0$.

Proof. Write $\widehat{p} = s/n \in [0, 1]$ and

$$q = \frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}. \tag{43}$$

To show that the upper endpoint is at most one, it is sufficient to prove

$$z\sqrt{q} \le 1 - \widehat{p} + \frac{z^2}{2n}, \tag{44}$$

because adding $\widehat{p} + z^2/(2n)$ to the left side then makes the numerator of Eq. (24) no larger than $1 + z^2/n$. Both sides of Eq. (44) are non-negative. After division by $z > 0$, the difference between the squared right- and left-hand sides is

$$\left(\frac{1-\widehat{p}}{z} + \frac{z}{2n}\right)^2 - q = \frac{(1-\widehat{p})^2}{z^2} + \frac{1-\widehat{p}}{n} - \frac{\widehat{p}(1-\widehat{p})}{n} \tag{45}$$

$$= (1-\widehat{p})^2\left(\frac{1}{z^2} + \frac{1}{n}\right) \ge 0. \tag{46}$$

Thus, $\mathcal{U}(s, n) \le 1$. Its numerator is non-negative, so $\mathcal{U}(s, n) \ge 0$. Strict positivity follows directly from Eq. (24): if $s > 0$, then its numerator contains the positive term $\widehat{p}$; if $s = 0$, then

$$\mathcal{U}(0, n) = \frac{z^2/n}{1 + z^2/n} = \frac{z^2}{n + z^2} > 0. \tag{47}$$

It remains to prove optimism. Subtracting $\widehat{p}$ from Eq. (24) gives

$$\mathcal{U}(s, n) - \widehat{p} = \frac{z\sqrt{q} + \frac{z^2}{n}(\frac{1}{2} - \widehat{p})}{1 + z^2/n}. \tag{48}$$

For $\widehat{p} \le 1/2$, the numerator is non-negative. For $\widehat{p} > 1/2$, both sides of

$$\sqrt{q} \ge \frac{z}{n}\left(\widehat{p} - \frac{1}{2}\right) \tag{49}$$

are non-negative, and the difference of their squares is

$$q - \frac{z^2}{n^2}\left(\widehat{p} - \frac{1}{2}\right)^2 = \widehat{p}(1-\widehat{p})\left(\frac{1}{n} + \frac{z^2}{n^2}\right) \geq 0. \tag{50}$$

Hence Eq. (48) is non-negative and $\mathcal{U}(s,n) \geq \widehat{p}$. If $\widehat{p} \leq 1/2$, then $z\sqrt{q} > 0$, so the premium is strict. If $1/2 < \widehat{p} < 1$, the last squared difference is strictly positive because $\widehat{p}(1-\widehat{p}) > 0$. At $\widehat{p} = 1$, direct substitution gives $\mathcal{U}(n,n) = 1$. This proves that equality occurs only at $s = n$.

For consistency, let $\widehat{p}_n = s/n \to p$. Since $0 \leq \widehat{p}_n(1-\widehat{p}_n) \leq 1/4$,

$$0 \leq z\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n} + \frac{z^2}{4n^2}} \leq \frac{z}{2}\sqrt{\frac{1}{n} + \frac{z^2}{n^2}} \longrightarrow 0. \tag{51}$$

The remaining correction $z^2/(2n)$ also tends to zero and the denominator tends to one. Therefore $\mathcal{U}(s,n) \to p$. Combining this with $\widehat{p}_n \to p$ yields $\mathcal{U}(s,n) - \widehat{p}_n \to 0$.

Lemma 3 (Support-dependent uncertainty premium)**.** *Fix an empirical failure rate $\widehat{p} \in [0,1]$ and $z > 0$, and regard the support $n > 0$ as continuous in the Wilson formula. If $n_2 > n_1$, then*

$$\mathcal{U}_{n_2}(\widehat{p}) \leq \mathcal{U}_{n_1}(\widehat{p}). \tag{52}$$

The inequality is strict for $\widehat{p} < 1$ and becomes equality for $\widehat{p} = 1$.

Proof. For fixed $\widehat{p}$, define the score-accepted set

$$\mathcal{A}_n = \left\{p \in [0,1] : n(\widehat{p} - p)^2 \leq z^2 p(1-p)\right\}. \tag{53}$$

If $p \in \mathcal{A}_{n_2}$, then

$$n_1(\widehat{p} - p)^2 < n_2(\widehat{p} - p)^2 \leq z^2 p(1-p) \tag{54}$$

whenever $p \neq \widehat{p}$, while $p = \widehat{p}$ belongs to both sets. Hence $\mathcal{A}_{n_2} \subseteq \mathcal{A}_{n_1}$. By Lemma 1, the upper endpoint of $\mathcal{A}_n$ is $\mathcal{U}_n(\widehat{p})$, so set inclusion gives $\mathcal{U}_{n_2}(\widehat{p}) \leq \mathcal{U}_{n_1}(\widehat{p})$.

For $\widehat{p} < 1$, Lemma 2 gives $u_1 := \mathcal{U}_{n_1}(\widehat{p}) > \widehat{p}$. As an endpoint of the score interval, u_1 satisfies

$$n_1(\widehat{p} - u_1)^2 = z^2 u_1(1-u_1). \tag{55}$$

Because $n_2 > n_1$ and $u_1 \neq \widehat{p}$,

$$n_2(\widehat{p} - u_1)^2 > z^2 u_1(1-u_1), \tag{56}$$

so $u_1 \notin \mathcal{A}_{n_2}$ and the upper endpoint under n_2 is strictly smaller. For $\widehat{p} = 1$, Lemma 2 gives $\mathcal{U}_n(1) = 1$ for every n.

Lemma 3 formalizes how URA uses class support. For two classes with the same observed failure proportion, the class with fewer observations receives the larger upper risk anchor. Scarcity therefore changes the uncertainty premium rather than being inserted directly as a frequency-derived vulnerability score.

Lemma 4 (Strict monotonicity in the event count)**.** *For fixed $n > 0$ and $z > 0$, $\mathcal{U}(s,n)$ is strictly increasing in $s \in \{0,\ldots,n\}$.*

Proof. It is sufficient to treat $\widehat{p} \in [0,1]$ as continuous. Define

$$q(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z^2}{4n^2}. \tag{57}$$

Differentiating Eq. (24) with respect to $\widehat{p}$ gives

$$\frac{\partial \mathcal{U}}{\partial \widehat{p}} = \frac{1 + \frac{z(1-2\widehat{p})}{2n\sqrt{q(\widehat{p})}}}{1 + z^2/n}. \tag{58}$$

For $\widehat{p} \leq 1/2$, its numerator is strictly positive. For $1/2 < \widehat{p} < 1$, we need $2n\sqrt{q(\widehat{p})} \geq z(2\widehat{p} - 1)$. Both sides are non-negative, and their squared difference is

$$4n^2 q(\widehat{p}) - z^2(2\widehat{p}-1)^2 = 4n\widehat{p}(1-\widehat{p}) + z^2 - z^2(1 - 4\widehat{p}(1-\widehat{p})) \tag{59}$$

$$= 4\widehat{p}(1-\widehat{p})(n + z^2) \geq 0. \tag{60}$$

The final expression is strictly positive for $\widehat{p} \in (1/2, 1)$, so Eq. (58) is strictly positive for every $\widehat{p} \in [0, 1)$. At $\widehat{p} = 1$, the derivative equals zero, but this does not destroy strict monotonicity. Indeed, for any $0 \le p_1 < p_2 \le 1$, the open interval (p_1, p_2) contains points below one on which the derivative is positive. The fundamental theorem of calculus gives

$$\mathcal{U}(p_2, n) - \mathcal{U}(p_1, n) = \int_{p_1}^{p_2} \frac{\partial \mathcal{U}(u, n)}{\partial u}\, du > 0. \tag{61}$$

Therefore $\mathcal{U}$ is strictly increasing on $[0, 1]$ and, because $s \mapsto s/n$ is strictly increasing, also in the discrete event count s.

Corollary 1 (Monotonic first-passage envelope)**.** *For fixed eligible support $M_c^{(t)} > 0$, the sequence $F_{c,1}^{+,(t)}, \ldots, F_{c,K}^{+,(t)}$ is non-decreasing.*

Proof. The cumulative counts obey $H_{c,k+1}^{(t)} \ge H_{c,k}^{(t)}$ by definition. Applying Lemma 4 with the fixed denominator $M_c^{(t)}$ gives $F_{c,k+1}^{+,(t)} \ge F_{c,k}^{+,(t)}$.

First-Passage Area and Boundary Depth

We first remove the confidence correction to expose the quantity represented by the curve area. Define the empirical cumulative curve

$$\widehat{F}_{c,k}^{(t)} = \frac{H_{c,k}^{(t)}}{M_c^{(t)}} \tag{62}$$

and its area $\widehat{D}_c^{(t)} = K^{-1} \sum_{k=1}^{K} \widehat{F}_{c,k}^{(t)}$.

Lemma 5 (Exact first-passage identity)**.** *For $M_c^{(t)} > 0$,*

$$\widehat{D}_c^{(t)} = \frac{1}{M_c^{(t)}} \sum_{i: y_i = c,\, a_i^{(t)} = 1} \frac{(K - T_i^{(t)} + 1)_+}{K}, \tag{63}$$

where $(u)_+ = \max(u, 0)$.

Proof. Let

$$\mathcal{E}_c^{(t)} = \{i : y_i = c,\ a_i^{(t)} = 1\} \tag{64}$$

be the clean-eligible index set, so $|\mathcal{E}_c^{(t)}| = M_c^{(t)}$. From Eq. (26),

$$H_{c,k}^{(t)} = \sum_{i \in \mathcal{E}_c^{(t)}} \mathbb{I}[T_i^{(t)} \le k]. \tag{65}$$

Substituting this identity into the empirical area gives

$$\widehat{D}_c^{(t)} = \frac{1}{K} \sum_{k=1}^{K} \frac{H_{c,k}^{(t)}}{M_c^{(t)}} \tag{66}$$

$$= \frac{1}{K M_c^{(t)}} \sum_{k=1}^{K} \sum_{i \in \mathcal{E}_c^{(t)}} \mathbb{I}[T_i^{(t)} \le k]. \tag{67}$$

Both sums are finite, so their order can be exchanged without an integrability assumption:

$$\widehat{D}_c^{(t)} = \frac{1}{M_c^{(t)}} \sum_{i \in \mathcal{E}_c^{(t)}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[T_i^{(t)} \le k]. \tag{68}$$

For a fixed eligible sample, the inner sum can be evaluated exactly:

$$\sum_{k=1}^{K} \mathbb{I}[T_i^{(t)} \le k] = \begin{cases} K - T_i^{(t)} + 1, & 1 \le T_i^{(t)} \le K, \\ 0, & T_i^{(t)} = K + 1. \end{cases} \tag{69}$$

The two cases are jointly written as $(K - T_i^{(t)} + 1)_+$. Substitution proves Eq. (63). Notice that the censored value $T_i^{(t)} = K+1$ contributes zero by definition; it is retained in the eligible denominator $M_c^{(t)}$ rather than discarded as a missing observation.

Lemma 5 makes the boundary cases explicit: failure at step one contributes 1, failure only at step K contributes $1/K$, and survival through step K contributes 0. For example, if $K = 4$ and a sample first fails at $T_i = 3$, its cumulative indicators are $(0, 0, 1, 1)$ and its contribution is $(0+0+1+1)/4 = 1/2$, equal to $(4 - 3 + 1)/4$. Thus, the empirical area is the mean normalized early-failure score on the discrete attack scale. Eq. (28) replaces each empirical cumulative rate with its upper confidence endpoint, adding an uncertainty premium while preserving monotonicity in every cumulative event count by Lemma 4.

Corollary 2 (Optimistic path area)**.** *For every class with $M_c^{(t)} > 0$,*

$$D_c^{(t)} \geq \widehat{D}_c^{(t)}. \tag{70}$$

The inequality is strict unless $H_{c,k}^{(t)} = M_c^{(t)}$ for every k.

Proof. For every k, Lemma 2 applied with $s = H_{c,k}^{(t)}$ and $n = M_c^{(t)}$ gives

$$F_{c,k}^{+,(t)} = \mathcal{U}(H_{c,k}^{(t)}, M_c^{(t)}) \geq \frac{H_{c,k}^{(t)}}{M_c^{(t)}} = \widehat{F}_{c,k}^{(t)}. \tag{71}$$

Summing the K inequalities and dividing by $K > 0$ yields

$$D_c^{(t)} - \widehat{D}_c^{(t)} = \frac{1}{K} \sum_{k=1}^{K} \left(F_{c,k}^{+,(t)} - \widehat{F}_{c,k}^{(t)} \right) \geq 0. \tag{72}$$

Every summand is non-negative, and Lemma 2 states that a summand is zero exactly when $H_{c,k}^{(t)} = M_c^{(t)}$. The average is therefore zero exactly when this equality holds at every attack step.

Proposition 6 (First-order attack-time dominance)**.** *Suppose two classes c and j have equal positive eligible support. If $H_{c,k}^{(t)} \geq H_{j,k}^{(t)}$ for every k, then $D_c^{(t)} \geq D_j^{(t)}$. The inequality is strict if at least one cumulative count is strict.*

Proof. Let the common eligible support be $M > 0$. At every step k, the assumption and Lemma 4 imply

$$F_{c,k}^{+,(t)} = \mathcal{U}(H_{c,k}^{(t)}, M) \geq \mathcal{U}(H_{j,k}^{(t)}, M) = F_{j,k}^{+,(t)}. \tag{73}$$

Consequently,

$$D_c^{(t)} - D_j^{(t)} = \frac{1}{K} \sum_{k=1}^{K} \left(F_{c,k}^{+,(t)} - F_{j,k}^{+,(t)} \right) \geq 0. \tag{74}$$

If at least one cumulative count is strict, strict monotonicity of $\mathcal{U}$ makes the corresponding summand strictly positive. All other summands remain non-negative, so the entire sum is strictly positive.

Properties of the Unified Allocation

We isolate the effect of centering and clipping. This step explains which order relations are preserved when $D_c^{(t)}$ is converted into $d_c^{(t)}$.

Lemma 6 (Centering and clipping)**.** *Assume $\mathcal{C}_{\text{obs}}^{(t)} \neq \varnothing$, so $\overline{D}^{(t)} > 0$. For every observed class define*

$$u_c^{(t)} = \frac{D_c^{(t)}}{\overline{D}^{(t)}} - 1, \qquad d_c^{(t)} = \text{clip}(u_c^{(t)}, -1, 1). \tag{75}$$

Then: (i) $d_c^{(t)} \in [-1, 1]$; (ii) $d_c^{(t)}$ has the same sign as $D_c^{(t)} - \overline{D}^{(t)}$; and (iii) $D_c^{(t)} \geq D_j^{(t)}$ implies $d_c^{(t)} \geq d_j^{(t)}$. The final inequality is strict whenever $D_c^{(t)} > D_j^{(t)}$ and the two un-clipped values are not mapped to the same saturation boundary.

Proof. Property (i) follows directly from the codomain of $\mathrm{clip}(\cdot, -1, 1)$. Because $\overline{D}^{(t)} > 0$,

$$u_c^{(t)} = \frac{D_c^{(t)} - \overline{D}^{(t)}}{\overline{D}^{(t)}}, \tag{76}$$

so $u_c^{(t)}$ has the same sign as $D_c^{(t)} - \overline{D}^{(t)}$. Clipping to a symmetric interval containing zero preserves that sign, proving (ii).

For (iii), division by the common positive value $\overline{D}^{(t)}$ and subtraction of one are strictly increasing transformations. Hence $D_c^{(t)} \geq D_j^{(t)}$ implies $u_c^{(t)} \geq u_j^{(t)}$. The clipping map is non-decreasing, so

$$d_c^{(t)} = \mathrm{clip}(u_c^{(t)}, -1, 1) \geq \mathrm{clip}(u_j^{(t)}, -1, 1) = d_j^{(t)}. \tag{77}$$

If $u_c^{(t)} > u_j^{(t)}$, equality after clipping can occur only when both are clipped to 1 or both to -1. Excluding this shared saturation gives strict inequality. Before clipping, the residuals have zero mean over the observed classes because

$$\frac{1}{|\mathcal{C}_{\mathrm{obs}}^{(t)}|} \sum_{c \in \mathcal{C}_{\mathrm{obs}}^{(t)}} u_c^{(t)} = \frac{|\mathcal{C}_{\mathrm{obs}}^{(t)}|^{-1} \sum_{c \in \mathcal{C}_{\mathrm{obs}}^{(t)}} D_c^{(t)}}{\overline{D}^{(t)}} - 1 = 0. \tag{78}$$

Proposition 7. *At $\lambda = 0$, the FTR residual vanishes and AURORA reduces exactly to URA: $w_c^{(t+1)} = U_c^{\mathrm{end},(t)} / (C^{-1} \sum_j U_j^{\mathrm{end},(t)})$. (***Proposition 1 of the main text.****)*

In the main text, the complete class risk is

$$\begin{aligned} R_c^{(t)} &= U_c^{\mathrm{end},(t)} \exp\Big(\lambda \omega_t d_c^{(t)}\Big), \\ \log R_c^{(t)} &= \log U_c^{\mathrm{end},(t)} + \lambda \omega_t d_c^{(t)}, \end{aligned} \tag{79}$$

and the next-epoch class weights are obtained by mean normalization:

$$w_c^{(t+1)} = \frac{R_c^{(t)}}{C^{-1} \sum_{j=1}^{C} R_j^{(t)}}. \tag{80}$$

Proof. Setting $\lambda = 0$ makes the temporal exponent zero for every class, independently of $d_c^{(t)}$ and ω_t. Eq. (79) therefore gives

$$R_c^{(t)} = U_c^{\mathrm{end},(t)} \exp(0) = U_c^{\mathrm{end},(t)}. \tag{81}$$

Let

$$\overline{U}^{\mathrm{end},(t)} = \frac{1}{C} \sum_{j=1}^{C} U_j^{\mathrm{end},(t)}. \tag{82}$$

Substitution into Eq. (80) gives

$$w_c^{(t+1)} = \frac{U_c^{\mathrm{end},(t)}}{\overline{U}^{\mathrm{end},(t)}}, \tag{83}$$

which is exactly the normalized URA allocation. In particular, the path statistic $D_c^{(t)}$ has no influence at $\lambda = 0$; this establishes an exact reduction rather than an approximation.

Proposition 8 (Bounded temporal refinement)**.** *For $\lambda \geq 0$, $d_c^{(t)} \in [-1, 1]$, and $\omega_t \in [0, 1]$,*

$$e^{-\lambda} \leq \frac{R_c^{(t)}}{U_c^{\mathrm{end},(t)}} \leq e^{\lambda}. \tag{84}$$

(**Proposition 2 of the main text.**)

Proof. From Eq. (79),

$$\frac{R_c^{(t)}}{U_c^{\mathrm{end},(t)}} = \exp(\lambda \omega_t d_c^{(t)}). \tag{85}$$

Because $\lambda \geq 0$, $0 \leq \omega_t \leq 1$, and $-1 \leq d_c^{(t)} \leq 1$,

$$-\lambda \leq -\lambda\omega_t \leq \lambda\omega_t d_c^{(t)} \leq \lambda\omega_t \leq \lambda. \tag{86}$$

The exponential function is strictly increasing, so exponentiating the middle expression gives

$$e^{-\lambda\omega_t} \leq \frac{R_c^{(t)}}{U_c^{\mathrm{end},(t)}} \leq e^{\lambda\omega_t}. \tag{87}$$

Because $\omega_t \leq 1$, this implies the epoch-independent bound in Eq. (84). For two classes, direct cancellation of the endpoint anchors gives

$$\frac{R_c^{(t)}/R_j^{(t)}}{U_c^{\mathrm{end},(t)}/U_j^{\mathrm{end},(t)}} = \exp\Big(\lambda\omega_t(d_c^{(t)} - d_j^{(t)})\Big). \tag{88}$$

Since $d_c^{(t)} - d_j^{(t)} \in [-2, 2]$, the ratio in Eq. (88) lies in the tighter epoch-wise interval

$$[e^{-2\lambda\omega_t}, e^{2\lambda\omega_t}] \subseteq [e^{-2\lambda}, e^{2\lambda}]. \tag{89}$$

This proves both the single-class and pairwise bounds.

Corollary 3 (Sufficient condition for endpoint-order preservation)**.** *For any classes c and j, if*

$$\frac{U_c^{\mathrm{end},(t)}}{U_j^{\mathrm{end},(t)}} > e^{2\lambda\omega_t}, \tag{90}$$

then $R_c^{(t)} > R_j^{(t)}$ and $w_c^{(t+1)} > w_j^{(t+1)}$, regardless of their clipped temporal depths.

Proof. Because $d_c^{(t)} - d_j^{(t)} \geq -2$, Eq. (88) gives

$$\frac{R_c^{(t)}}{R_j^{(t)}} \geq \frac{U_c^{\mathrm{end},(t)}}{U_j^{\mathrm{end},(t)}} e^{-2\lambda\omega_t} > 1. \tag{91}$$

Hence $R_c^{(t)} > R_j^{(t)}$. Class normalization divides both risks by the same positive constant, so it preserves this ordering. This result makes the role of clipping explicit: temporal evidence can refine a close endpoint comparison, but cannot reverse an endpoint ratio separated by more than the largest permitted temporal factor.

Proposition 9 (Two-level conservation)**.** *The class and mini-batch normalizations satisfy*

$$\frac{1}{C}\sum_{c=1}^{C} w_c^{(t+1)} = 1, \qquad \frac{1}{B}\sum_{i=1}^{B} \widetilde{w}_i^{(t)} = 1. \tag{92}$$

(**Proposition 3 of the main text.**)

Proof. Let $\overline{R} = C^{-1}\sum_j R_j^{(t)}$. Eq. (80) gives

$$\frac{1}{C}\sum_{c=1}^{C} w_c^{(t+1)} = \frac{1}{C}\sum_{c=1}^{C} \frac{R_c^{(t)}}{\overline{R}} = \frac{C^{-1}\sum_c R_c^{(t)}}{\overline{R}} = 1. \tag{93}$$

For a mini-batch, let $S_B = \sum_{j=1}^{B} w_{y_j}^{(t)}$.

$$\widetilde{w}_i^{(t)} = \frac{w_{y_i}^{(t)}}{B^{-1}\sum_{j=1}^{B} w_{y_j}^{(t)}}, \tag{94}$$

Then Eq. (94) gives

$$\frac{1}{B}\sum_{i=1}^{B} \widetilde{w}_i^{(t)} = \frac{1}{B}\sum_{i=1}^{B} \frac{w_{y_i}^{(t)}}{S_B/B} = \frac{S_B}{S_B} = 1. \tag{95}$$

Both normalizations divide every element in their domain by the same positive constant. Consequently,

$$\frac{w_c^{(t+1)}}{w_j^{(t+1)}} = \frac{R_c^{(t)}}{R_j^{(t)}}, \qquad \frac{\widetilde{w}_i^{(t)}}{\widetilde{w}_j^{(t)}} = \frac{w_{y_i}^{(t)}}{w_{y_j}^{(t)}}, \tag{96}$$

which proves preservation of pairwise ratios. Notice that this result preserves the *average multiplier*, not the numerical value of the weighted robust loss; changing the allocation is intended to change the latter.

Proposition 10. *Consider classes c and j with equal endpoint anchors and positive eligible support. If $F_{c,k}^{+,(t)} \geq F_{j,k}^{+,(t)}$ at every attack step, with strict inequality at some step, then $D_c^{(t)} > D_j^{(t)}$. If neither centered depth is clipped and $\lambda\omega_t > 0$, then $w_c^{(t+1)} > w_j^{(t+1)}$.*

(Proposition 4 of the main text.)

In the main text, we center and clip relative path risk:

$$d_c^{(t)} = \mathrm{clip}\left(\frac{D_c^{(t)}}{\overline{D}^{(t)}} - 1, -1, 1\right), \tag{97}$$

Proof. The assumed pointwise dominance gives

$$D_c^{(t)} - D_j^{(t)} = \frac{1}{K}\sum_{k=1}^{K}\left(F_{c,k}^{+,(t)} - F_{j,k}^{+,(t)}\right) > 0, \tag{98}$$

because every summand is non-negative and at least one is positive. Lemma 6 then gives $d_c^{(t)} > d_j^{(t)}$ under the stated non-saturation condition. Equal endpoint anchors and Eq. (79) yield

$$\frac{R_c^{(t)}}{R_j^{(t)}} = \exp\Big(\lambda\omega_t(d_c^{(t)} - d_j^{(t)})\Big) > 1. \tag{99}$$

The exponent is positive because $\lambda\omega_t > 0$, hence the ratio exceeds one. Proposition 3 shows that class normalization preserves this ratio, completing the chain

$$F_c^+ \succ F_j^+ \Longrightarrow D_c > D_j \Longrightarrow d_c > d_j \Longrightarrow R_c > R_j \Longrightarrow w_c > w_j. \tag{100}$$

WARM-UP AND DELAYED OPTIMIZATION

In the main text, to suppress unreliable paths before the attack reaches a meaningful strength, we reuse the base perturbation warm-up

$$\omega_t = \min\left(\frac{t-1}{\rho T}, 1\right), \tag{101}$$

Lemma 7 (Warm-up properties)**.** *For $t \geq 1$, $T > 0$, and $\rho \in (0, 1]$, the intensity in Eq.* (101) *satisfies: (i) $0 \leq \omega_t \leq 1$; (ii) ω_t is non-decreasing in t; (iii) $\omega_1 = 0$; and (iv) $\omega_t = 1$ whenever $t - 1 \geq \rho T$.*

Proof. Because $\rho T > 0$, Eq. (18) has the equivalent piecewise form

$$\omega_t = \begin{cases} \dfrac{t-1}{\rho T}, & t - 1 < \rho T, \\ 1, & t - 1 \geq \rho T. \end{cases} \tag{102}$$

For $t \geq 1$, the numerator $t - 1$ is non-negative, proving $\omega_t \geq 0$. The minimum with one proves $\omega_t \leq 1$, which establishes (i). The affine branch $(t - 1)/(\rho T)$ has positive slope $1/(\rho T)$, and the second branch is constant at the first branch's cap; therefore the full function is non-decreasing, proving (ii). Substituting $t = 1$ gives $\omega_1 = 0$, proving (iii). Finally, if $t - 1 \geq \rho T$, the second branch applies and gives $\omega_t = 1$, proving (iv).

Lemma 7 shows that the temporal refinement is inactive in the first epoch and follows the attack intensity already present in the base recipe. For $T = 110$ and $\rho = 0.8$, we have $\rho T = 88$. The smallest integer epoch satisfying $t - 1 \geq 88$ is $t = 89$, so the full temporal coefficient becomes active at epoch 89.

In the main text, AURORA reweights only the robust term using class statistics collected in the preceding epoch:

$$\mathcal{L}^{(t)}_{\text{AURORA}} = \frac{1}{B}\sum_{i=1}^{B}\left[\ell^{(t)}_{\text{com},i} + \widetilde{w}^{(t)}_i \ell^{\mathcal{B},(t)}_{\text{rob},i}\right]. \tag{103}$$

Proposition 11 (No within-step gradient through allocation)**.** *Suppose $w^{(t)}$ is computed from detached statistics collected in epoch $t-1$ and held fixed throughout epoch t. Then no learner-specific objective in Eq.* (103) *differentiates through the risk estimator. In particular, the outer-update gradients take the form*

$$\begin{aligned}
\nabla_\theta \mathcal{L}^{(t)}_{\textit{AURORA,AT}} &= \frac{1}{B}\sum_i \widetilde{w}^{(t)}_i \nabla_\theta \ell^{\text{adv}}_i(\theta_t), \\
\nabla_\theta \mathcal{L}^{(t)}_{\textit{AURORA,TRADES}} &= \frac{1}{B}\sum_i \left[\nabla_\theta \ell_\pi(f_{\theta_t}(x_i), y_i) + \beta \widetilde{w}^{(t)}_i \nabla_\theta d^{\text{KL}}_i(\theta_t)\right], \\
\nabla_\theta \mathcal{L}^{(t)}_{\textit{AURORA,AWP}} &= \frac{1}{B}\sum_i \widetilde{w}^{(t)}_i \nabla_\theta \ell^{\text{adv}}_i(\theta_t + \delta^*_t),
\end{aligned} \tag{104}$$

where the generated adversarial examples and the estimated AWP perturbation are held fixed during the corresponding outer update, as in the base implementations.

Proof. Write the mini-batch normalizer as

$$Z^{(t)}_B = \frac{1}{B}\sum_{j=1}^{B} w^{(t)}_{y_j}, \qquad \widetilde{w}^{(t)}_i = \frac{w^{(t)}_{y_i}}{Z^{(t)}_B}. \tag{105}$$

Both the numerator and denominator depend only on labels in the current mini-batch and on weights computed from detached statistics in epoch $t-1$. They are therefore constant with respect to the current model parameters:

$$\nabla_\theta w^{(t)}_{y_i} = 0, \qquad \nabla_\theta Z^{(t)}_B = 0, \qquad \nabla_\theta \widetilde{w}^{(t)}_i = 0. \tag{106}$$

For the AT branch, the product rule gives

$$\nabla_\theta\left(\widetilde{w}^{(t)}_i \ell^{\text{adv}}_i(\theta_t)\right) = \underbrace{\ell^{\text{adv}}_i(\theta_t)\nabla_\theta \widetilde{w}^{(t)}_i}_{=0} + \widetilde{w}^{(t)}_i \nabla_\theta \ell^{\text{adv}}_i(\theta_t). \tag{107}$$

Summing over i gives the displayed AT gradient. For TRADES, the clean classification term contains no AURORA weight, while the same product-rule calculation applies to $\beta\widetilde{w}^{(t)}_i d^{\text{KL}}_i(\theta_t)$; β is fixed, so this gives the displayed TRADES gradient. For AWP, the parameter perturbation δ^*_t is first estimated by the proxy step and then treated as fixed during the outer update. Applying the AT calculation at the perturbed parameter $\theta_t + \delta^*_t$ gives the displayed AWP gradient. The adversarial inputs are likewise detached before the outer update, following the corresponding base learner.

Thus, AURORA changes the magnitude of the native robust gradient through $\widetilde{w}^{(t)}_i$, but no derivative path leads from the current loss back through E_c, $H_{c,k}$, the Wilson operator, or the preceding epoch's counting process.