PROBABILISTIC ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Building on a probabilistic perspective in which adversarial examples arise from the overlap between a distance-based distribution p_{dis} and a victim-classifier-induced distribution p_{vic}, we start from a simple intuition: adversarial examples become harder to generate when these two distributions are pushed apart, as their overlap becomes smaller, thereby increasing robustness. This intuition naturally motivates a KL-based robustness objective. We then prove that $\mathrm{KL}(p_{\mathrm{dis}} \| p_{\mathrm{vic}}) - \log Z_{\mathrm{vic}}$ is a lower bound on probabilistic robustness (PR), where Z_{vic} denotes the normalizing constant of p_{vic}. Since PR is generally intractable to compute directly, maximizing this KL-based lower bound provides a tractable surrogate objective for improving PR. We further show that this objective recovers a scaled form of adversarial training, offering a probabilistic interpretation of adversarial training and a principled route to robustness improvement. We call the resulting method **probabilistic adversarial training**. Experiments show that it consistently improves PR, and ablation studies demonstrate that the induced scaling factor can even enhance the PR of non-probabilistic adversarial training methods. Code is available at `https://anonymous.4open.science/r/PATICLR27`.

1 INTRODUCTION

Adversarial robustness refers to a model's ability to maintain correct predictions when its inputs are subjected to small perturbations. This property is important in the era of deep learning, as a substantial body of work (Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018; Moosavi-Dezfooli et al., 2016; 2017; Papernot et al., 2016; Szegedy et al., 2014) has shown that deep neural networks can be highly vulnerable to adversarial perturbations.

To better understand and evaluate this vulnerability, Zhang et al. (2024a) proposed probabilistic adversarial attack. By applying Langevin Dynamics, they demonstrated that adversarial examples can be viewed as samples drawn from an adversarial distribution. This distribution is essentially the product of a "victim" distribution p_{vic} that encourages misclassification, and a "distance" distribution p_{dis} that constrains the perturbation magnitude.

However, their framework is strictly confined to targeted attacks and focuses exclusively on the generation of adversarial examples, leaving a systematic approach to model defense unexplored. On the other side, to formally quantify model security, Webb et al. (2018) introduced the concept of probabilistic robustness (PR), defined as the probability that a classifier successfully resists perturbations sampled from a specific distribution.

In this work, we unify the probabilistic attack perspective (Zhang et al., 2024a) with the probabilistic robustness framework (Webb et al., 2018). We theoretically prove that $\mathrm{KL}(p_{\mathrm{dis}} \| p_{\mathrm{vic}}) - \log Z_{\mathrm{vic}}$ is a lower bound on probabilistic robustness, where Z_{vic} denotes the normalizing constant of p_{vic}. Since computing and optimizing PR directly is generally intractable, maximizing this KL-based lower bound provides a tractable surrogate objective to effectively enhance PR.

This theoretical result aligns with a simple yet profound intuition. In essence, adversarial examples naturally arise from the overlap between the distance distribution p_{dis} (which confines the perturbation to be close to the original image) and the victim distribution p_{vic} (which actively drives the model toward misclassification). By maximizing the KL divergence $\mathrm{KL}(p_{\mathrm{dis}} \| p_{\mathrm{vic}})$, we are pushing these two distributions apart. As their overlap shrinks, it becomes inherently more difficult for an attacker to sample an adversarial example that simultaneously satisfies the distance constraint and triggers a misclassification.

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.

We further show that maximizing this lower bound essentially recovers a scaled form of adversarial training. This provides a probabilistic interpretation of adversarial training itself, and provides a principled route to robustness improvement, which we term Probabilistic Adversarial Training (PAT). Empirically, PAT consistently improves probabilistic robustness. Moreover, our ablation studies reveal an intriguing byproduct: the theoretical scaling factor derived from our framework can plug into and enhance traditional, non-probabilistic defense methods, such as PGD-based adversarial training.

To summarize, our main contributions are as follows:

- We formalize the untargeted version of probabilistic adversarial attacks. (Section 3)
- We prove that $\mathrm{KL}(p_{\mathrm{dis}} \| p_{\mathrm{vic}}) - \log Z_{\mathrm{vic}}$ is a lower bound on probabilistic robustness (PR). (Section 4)
- By maximizing this KL-based lower bound, we propose Probabilistic Adversarial Training (PAT), which is theoretically guaranteed to improve PR. Moreover, its importance sampling formulation makes the estimation of the bound tractable (Section 5).
- We empirically demonstrate that PAT consistently improves PR, and its derived scaling factor can even enhance non-probabilistic methods like PGD-based adversarial training. (Section 6)

2 Preliminaries

2.1 Adversarial Attack

Let $C : [0,1]^d \to \mathcal{Y}$ be a classifier with input dimension d and label space $\mathcal{Y}$. Let $x_{\mathrm{ori}} \in [0,1]^d$ be a clean (original) image with its corresponding true label $y_{\mathrm{ori}} \in \mathcal{Y}$. Adversarial attacks generally fall into two categories. The goal of an **untargeted adversarial attack** is to construct an adversarial example x_{adv} that simply causes the classifier to misclassify the image, meaning $C(x_{\mathrm{adv}}) \neq y_{\mathrm{ori}}$. Alternatively, given a specific target label $y_{\mathrm{tar}} \in \mathcal{Y}$ (where $y_{\mathrm{tar}} \neq y_{\mathrm{ori}}$), a **targeted adversarial attack** aims to force the classifier to predict that exact label, such that $C(x_{\mathrm{adv}}) = y_{\mathrm{tar}}$. In both scenarios, the distance between x_{adv} and x_{ori} must remain small. The corresponding optimization problem can be formulated as:

$$\min \mathcal{D}(x_{\mathrm{ori}}, x_{\mathrm{adv}}) \quad \text{subject to} \quad C(x_{\mathrm{adv}}) \in \mathcal{T} \quad \text{and} \quad x_{\mathrm{adv}} \in [0,1]^d,$$

where $\mathcal{T} = \{y_{\mathrm{tar}}\}$ for a targeted attack and $\mathcal{T} = \mathcal{Y} \setminus \{y_{\mathrm{ori}}\}$ for an untargeted attack, and $\mathcal{D}$ measures the distance (similarity) between x_{ori} and x_{adv}, typically via an L_1, L_2, or L_∞ norm. Directly solving this constrained optimization can be challenging. To address this, Szegedy et al. (2014) propose relaxing it into the following optimization problem:

$$\min_{x_{\mathrm{adv}} \in [0,1]^d} \; c_1 \, \mathcal{D}(x_{\mathrm{ori}}, x_{\mathrm{adv}}) \; + \; c_2 \, L(x_{\mathrm{adv}}) \tag{1}$$

where $c_1 > 0$ and $c_2 > 0$ are trade-off constants. For a targeted attack, we set $L(x_{\mathrm{adv}}) = f(x_{\mathrm{adv}}, y_{\mathrm{tar}})$ to minimize the loss with respect to the target label; for an untargeted attack, we set $L(x_{\mathrm{adv}}) = -f(x_{\mathrm{adv}}, y_{\mathrm{ori}})$ to maximize the loss with respect to the true label, thereby pushing the prediction away from the original class. In Szegedy et al. (2014)'s work, f is taken to be the cross-entropy loss[1]; Carlini & Wagner (2017) present additional choices for f.

2.2 Probabilistic Adversarial Attack

Zhang et al. (2024a) derive a probabilistic perspective on adversarial attacks by applying Langevin Dynamics as an optimizer for equation 1. Notably, their framework focuses exclusively on targeted attack. Under this specific setting (where $L(x_{\mathrm{adv}}) = f(x_{\mathrm{adv}}, y_{\mathrm{tar}})$), they introduce the adversarial distribution:

$$p_{\mathrm{adv}}(x_{\mathrm{adv}}; x_{\mathrm{ori}}, y_{\mathrm{tar}}) \propto p_{\mathrm{vic}}(x_{\mathrm{adv}}; y_{\mathrm{tar}}) \, p_{\mathrm{dis}}(x_{\mathrm{adv}}; x_{\mathrm{ori}}),$$

[1]When we say that $f(x, y)$ is the cross-entropy loss, we mean the cross-entropy between the Dirac categorical distribution concentrated at label y and the categorical distribution output by the neural network given input x.

where $p_{\text{vic}}(x_{\text{adv}}; y_{\text{tar}}) \propto \exp\big(-c_2\, f(x_{\text{adv}}, y_{\text{tar}})\big)$ is the "victim" distribution emphasizing misclassification toward the target label y_{tar}, and $p_{\text{dis}}(x_{\text{adv}}; x_{\text{ori}}) \propto \exp\big(-c_1\, \mathcal{D}(x_{\text{ori}}, x_{\text{adv}})\big)$ is the "distance" distribution around x_{ori}. This formulation leverages the fact that Langevin Dynamics converges to the corresponding Gibbs distribution (Lamperski, 2021), thereby providing a probabilistic interpretation of adversarial examples.

This probabilistic perspective aligns with traditional geometry-based adversarial attacks. For example, if $\mathcal{D}$ is the L_1 norm, then $p_{\text{dis}}(x_{\text{adv}}; x_{\text{ori}}) \propto \exp(-\|x_{\text{adv}} - x_{\text{ori}}\|_1)$ takes the form of a Laplace distribution; if $\mathcal{D}$ is the squared L_2 norm, then $p_{\text{dis}}(x_{\text{adv}}; x_{\text{ori}}) \propto \exp(-\|x_{\text{adv}} - x_{\text{ori}}\|_2^2)$ is a Gaussian distribution; if $\mathcal{D}$ is L_∞ norm, then p_{dis} is an L_∞-spherical distribution (Iglesias et al., 1998).

2.3 Probabilistic Robustness

Webb et al. (2018) introduced a statistical framework to assess neural network robustness by evaluating the probability of an attack's success. Let p denote the probability density function of perturbations around a clean input x_{ori}. To quantify misclassification, let $z(x) \in \mathbb{R}^{|\mathcal{Y}|}$ denote the pre-softmax logits output by the classifier. For an untargeted attack, the margin function $s(\cdot)$ is formally defined as the logit margin violation:

$$s(x) = \max_{y \neq y_{\text{ori}}} \left[z_y(x) - z_{y_{\text{ori}}}(x) \right].$$

The event $s(X) \geq 0$ strictly indicates that at least one incorrect class logit matches or exceeds the true class logit, signifying a successful adversarial attack (i.e., the classifier is fooled).

Consequently, the attack success probability, or **probabilistic vulnerability**, is defined as:

$$\mathcal{I}[p, s] = \mathbb{P}_{X \sim p}(s(X) \geq 0) = \int \mathbb{I}_{\{s(x) \geq 0\}} p(x) dx,$$

where $\mathbb{I}$ is the indicator function.

While directly minimizing $\mathcal{I}[p, s]$ reduces vulnerability, framing robustness as an objective to be maximized provides a more consistent narrative for our subsequent theoretical derivations. Therefore, throughout this paper, we formally define **Probabilistic Robustness** (PR) as the complementary probability of an attack failing:

$$\text{PR}(p, s) := 1 - \mathcal{I}[p, s] = \mathbb{P}_{X \sim p}(s(X) < 0).$$

Under this definition, enhancing the intrinsic security of a model corresponds directly to maximizing the value of PR.

2.4 Adversarial Training

Adversarial training (AT) is a widely adopted empirical method to defend against adversarial attack described above (Madry et al., 2018). The core idea of AT is to augment the training data with adversarial examples during the learning process. Formally, let θ denote the parameters of the classifier, and let p_{data} be the underlying data distribution. Adversarial training formulates the robustness objective as a min-max optimization problem:

$$\min_{\theta} \mathbb{E}_{(x_{\text{ori}}, y_{\text{ori}}) \sim p_{\text{data}}} \left[\max_{x_{\text{adv}} \in \mathcal{S}(x_{\text{ori}})} f(x_{\text{adv}}, y_{\text{ori}}; \theta) \right],$$

where $\mathcal{S}(x_{\text{ori}}) = \{x_{\text{adv}} \in [0,1]^d : \mathcal{D}(x_{\text{ori}}, x_{\text{adv}}) \leq \epsilon\}$ defines the set of allowed adversarial perturbations bounded by a small budget ϵ (e.g., an L_∞ or L_2 norm ball).

3 An Untargeted Version of Probabilistic Adversarial Attack

As introduced in Section 2.2, an untargeted adversarial attack can be formulated as the following optimization problem:

$$\min_{x_{\text{adv}} \in [0,1]^d} \; c_1\, \mathcal{D}(x_{\text{ori}}, x_{\text{adv}}) \; - \; c_2\, f(x_{\text{adv}}, y_{\text{ori}})$$

Similar to the work of Zhang et al. (2024a), applying Langevin dynamics to this optimization problem yields convergence to a Gibbs distribution concentrated around the solutions:

$$p_{\mathrm{adv}}(x_{\mathrm{adv}}; x_{\mathrm{ori}}, y_{\mathrm{ori}}) \propto p_{\mathrm{vic}}(x_{\mathrm{adv}}; y_{\mathrm{ori}})\, p_{\mathrm{dis}}(x_{\mathrm{adv}}; x_{\mathrm{ori}}),$$

where $p_{\mathrm{vic}}(x_{\mathrm{adv}}; y_{\mathrm{ori}}) \propto \exp\big(c_2\, f(x_{\mathrm{adv}}, y_{\mathrm{ori}})\big)$ is the "victim" distribution and $p_{\mathrm{dis}}(x_{\mathrm{adv}}; x_{\mathrm{ori}}) \propto \exp\big(-c_1\, \mathcal{D}(x_{\mathrm{ori}}, x_{\mathrm{adv}})\big)$ is the "distance" distribution around x_{ori}. For typical choices of $\mathcal{D}$, it is clear that p_{dis} is well-defined (Zhang et al., 2024a). The following proposition shows that p_{vic} is also well-defined in the untargeted setting:

Proposition 1 (Well-definedness of the untargeted victim distribution)**.** *Let $K = [0,1]^d$ and let $f(\cdot, y_{\mathrm{ori}}) : K \to \mathbb{R}$ be continuous. For any $c > 0$, define*

$$p_{\mathrm{vic}}(x; y_{\mathrm{ori}}) = \frac{\exp\big(cf(x, y_{\mathrm{ori}})\big)}{\int_K \exp\big(cf(u, y_{\mathrm{ori}})\big)\, du}, \qquad x \in K.$$

Then $p_{\mathrm{vic}}(\cdot; y_{\mathrm{ori}})$ is a well-defined probability density on K.

The proof is in Appendix B. In this paper, we take $\mathcal{D}$ to be the squared L_2 norm, so that p_{dis} is a Gaussian distribution, and choose f to be the cross-entropy loss. Under this setting, the Langevin-dynamics-based algorithm for sampling adversarial examples from p_{adv} is presented in Algorithm 1 (See Appendix A).

4 A KL-Based Lower Bound on Probabilistic Robustness

Recall that $\mathrm{PR}(p, s)$ (introduced in Section 2.3) evaluates the model's security against perturbations drawn from a distribution p around each clean (original) input x_{ori}. We observe that the distance distribution $p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})$ introduced in Section 2.2 naturally serves as the evaluation distribution p. Consequently, our objective formally becomes maximizing $\mathrm{PR}(p_{\mathrm{dis}}, s)$.

To establish a mathematical lower bound on PR, we must connect the attack-success event, characterized by the logit margin violation $s(x) \geq 0$ (Section 2.3), with our optimization objective, which relies on the continuous cross-entropy loss $f(x, y_{\mathrm{ori}})$. However, establishing a direct, tractable algebraic mapping between the margin $s(x)$ and the loss f is highly challenging due to softmax non-linearities and complex multi-class dynamics.

To circumvent this, we define the attack-success region $E_{\mathrm{adv}} = \{x \in \mathcal{X} : s(x) \geq 0\}$ as the vulnerable space where the classifier is fooled. We then introduce a connecting scalar γ_* representing the infimum of the scaled cross-entropy loss over this region:

$$\gamma_* := \inf_{x \in E_{\mathrm{adv}}} c_2 f(x, y_{\mathrm{ori}}; \theta).$$

Since the cross-entropy loss is inherently non-negative and the temperature scaling factor $c_2 > 0$, it strictly holds that $\gamma_* \geq 0$. By definition, any successful adversarial example $x \in E_{\mathrm{adv}}$ rigorously guarantees $c_2 f(x, y_{\mathrm{ori}}; \theta) \geq \gamma_*$. This property is pivotal, as it allows us to formally link the expected loss under p_{dis} to the probability mass of the vulnerable region, paving the way for the following theorem.

Theorem 2 (KL-Based Lower Bound on Probabilistic Robustness)**.** *Let $p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})$ be a probability density on $[0,1]^d$. Suppose the victim distribution parameterized by θ is defined as*

$$p_{\mathrm{vic}}(x; y_{\mathrm{ori}}, \theta) = \frac{\exp(c_2 f(x, y_{\mathrm{ori}}; \theta))}{Z_{\mathrm{vic}}},$$

where $Z_{\mathrm{vic}} = \int_{\mathcal{X}} \exp(c_2 f(u, y_{\mathrm{ori}}; \theta)) du$ is the normalizing constant. Let γ_* be defined as above and assume $\gamma_* > 0$. Then, the probabilistic robustness satisfies:

$$\mathrm{PR}(p_{\mathrm{dis}}, s) \geq 1 - \frac{\log Z_{\mathrm{vic}} - H(p_{\mathrm{dis}}) - \mathrm{KL}(p_{\mathrm{dis}} \| p_{\mathrm{vic}})}{\gamma_*},$$

where $H(p_{\mathrm{dis}}) = -\int_{\mathcal{X}} p_{\mathrm{dis}}(x) \log p_{\mathrm{dis}}(x) dx$ is the differential entropy of the distance distribution.

The proof is in Appendix B. Crucially, Theorem 2 formally unifies the probabilistic attack perspective (Zhang et al., 2024a) with the probabilistic robustness framework (Webb et al., 2018). By expressing the lower bound on PR using the statistical divergence between p_{dis} and p_{vic}, this theorem directly translates our intuition of "pushing the distributions apart" into a rigorous mathematical objective.

Notice that once the clean input x_{ori} and the perturbation budget are fixed, the distance distribution p_{dis} and its differential entropy $H(p_{\text{dis}})$ are constants. Furthermore, since γ_* acts as a positive scaling factor, maximizing the lower bound on PR requires maximizing the term $\text{KL}(p_{\text{dis}} \| p_{\text{vic}}) - \log Z_{\text{vic}}$. This fundamental insight serves as the theoretical bedrock for the tractable surrogate objective derived in the next section.

While intuitively one might assume that maximizing $KL(p_{dis}||p_{vic})$ alone is sufficient to separate the distributions, the $-\log Z_{vic}$ term holds crucial geometric and physical significance. We defer a detailed discussion on this to Appendix D.

5 Probabilistic Adversarial Training

Based on Section 4, maximizing the lower bound $KL(p_{dis}||p_{vic}) - \log Z_{vic}$ acts as a surrogate objective that indirectly improves probabilistic robustness. However, calculating the exact gradients of this theoretical bound is challenging due to intractable normalizing constants. In this section, we derive a computable empirical loss function through a structured, three-step derivation. We then explicitly reveal how this empirical formulation recovers a scaled, probabilistic counterpart to standard adversarial training, before formally proving its convergence. Algorithm 2 in Appendix A presents the pseudocode for the final practical implementation of probabilistic adversarial training.

5.1 Deriving the Empirical Loss Function

Step 1: The Idealized Objective. We first attempt to formulate an objective by directly applying importance sampling to the theoretical lower bound.

Proposition 3 (Idealized Objective)**.** *Let p_{dis}, p_{vic}, and p_{adv} be defined as in Section 3. Maximizing the probabilistic robustness lower bound with respect to the classifier parameters θ is mathematically equivalent to minimizing the following idealized expected loss:*

$$\mathcal{J}_{ideal}(\theta) = \mathbb{E}_{X \sim p_{adv}(\cdot; x_{ori}, y_{ori}, \theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} c_2 f(X, y_{ori}; \theta) \right]$$

See Appendix B for the proof. While Proposition 3 demonstrates that the intractable $\log Z_{vic}$ cancels out, directly optimizing $\mathcal{J}_{ideal}(\theta)$ is highly problematic. The sampling distribution $p_{adv}(X; x_{ori}, y_{ori}, \theta) \propto p_{dis}(X; x_{ori}) p_{vic}(X; y_{ori}, \theta)$ depends on the parameters θ. Differentiating through this expectation is notoriously unstable and requires complex score-function estimators.

Step 2: The Gradient-First Approach. To circumvent the challenge of differentiating through a parameterized distribution, we propose computing the gradient of the lower bound before applying importance sampling.

Proposition 4 (Gradient of the Objective)**.** *The exact gradient of the negative lower bound with respect to θ can be expressed as an expectation over the adversarial distribution p_{adv}:*

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{X \sim p_{adv}(\cdot; x_{ori}, y_{ori}, \theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} \nabla_\theta (c_2 f(X, y_{ori}; \theta)) \right]$$

See Appendix B for the proof. This formulation decouples the sampling distribution from the gradient computation. Because the gradient operator ∇_θ only applies to the loss term f, the importance weight $w(X; x_{ori}, y_{ori}, \theta) = \frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)}$ acts purely as a scaling constant for the gradients of individual samples.

Step 3: Self-Normalized Importance Sampling. Although Proposition 4 provides a relatively tractable gradient direction, evaluating the theoretical weight $w(X; x_{ori}, y_{ori}, \theta)$ is still impossible. By substituting $p_{vic}(X; y_{ori}, \theta) = \frac{1}{Z_{vic}} \exp(c_2 f(X, y_{ori}; \theta))$, we reveal the hidden constants:

$$w(X; x_{ori}, y_{ori}, \theta) = Z_{adv} Z_{vic} \exp(-c_2 f(X, y_{ori}; \theta))$$

Both Z_{adv} and Z_{vic} are intractable to compute. However, for a given clean image x_{ori} and a specific set of model parameters θ, their product is strictly a constant across different sampled perturbations.

To circumvent the intractability of these constants, we employ Self-Normalized Importance Sampling (SNIS). Given a mini-batch of adversarial examples $X^{(1)}, ..., X^{(B)}$ sampled from p_{adv}, we normalize the weights across the batch:

$$\hat{w}^{(i)} = \frac{w(X^{(i)}; x_{ori}, y_{ori}, \theta)}{\sum_{j=1}^{B} w(X^{(j)}; x_{ori}, y_{ori}, \theta)} = \frac{Z_{adv} Z_{vic} \exp(-c_2 f(X^{(i)}, y_{ori}; \theta))}{\sum_{j=1}^{B} Z_{adv} Z_{vic} \exp(-c_2 f(X^{(j)}, y_{ori}; \theta))}$$
$$= \frac{\exp(-c_2 f(X^{(i)}, y_{ori}; \theta))}{\sum_{j=1}^{B} \exp(-c_2 f(X^{(j)}, y_{ori}; \theta))}$$

The intractable terms Z_{adv} and Z_{vic} cancel out exactly. This self-normalized weighting scheme naturally corresponds to applying a Softmax function. Consequently, the true gradient derived in Proposition 4 can be estimated by backpropagating through the following empirical loss:

$$\mathcal{L}_{emp}(\theta) = \sum_{i=1}^{B} \hat{w}^{(i)} f(X^{(i)}, y_{ori}; \theta)$$

where $\hat{w}^{(i)}$ are treated as fixed constants (i.e., stop-gradient) during optimization, as explicitly implemented in Lines 16-18 of Algorithm 2 (Appendix A).

5.2 Connection to Standard Adversarial Training

Standard adversarial training (AT) methods, such as PGD with random restarts, can be viewed probabilistically as drawing samples from an implicit adversarial distribution p_{adv} to minimize the unweighted expected loss $\mathbb{E}_{X \sim p_{adv}(\cdot|x_{\text{ori}}, y_{\text{ori}})}[f(X, y_{ori}; \theta)]$. Empirically, this directly corresponds to optimizing a uniformly weighted batch loss $\frac{1}{B} \sum_{i=1}^{B} f(X^{(i)}, y_{ori}; \theta)$.

In this work, both our idealized objective $\mathcal{J}_{ideal}$ and empirical loss $\mathcal{L}_{emp}$ recover this exact AT paradigm, differing solely by a principled scaling factor. Instead of treating all samples equally, Probabilistic Adversarial Training (PAT) optimizes a weighted loss using the self-normalized importance weights $\hat{w}^{(i)} \propto \exp(-c_2 f(X^{(i)}, y_{ori}; \theta))$.

This negative exponent acts as a direct penalty on extreme adversarial examples. While standard AT strictly focuses on worst-case perturbations, PAT deliberately down-weights extreme outliers. This weighting mechanism prevents the model from over-optimizing for worst-case attacks, thereby improving overall probabilistic robustness.

5.3 Convergence of the Empirical Gradient Estimator

We present a convergence theorem demonstrating that our self-normalized importance sampling (SNIS) estimator is asymptotically consistent.

Theorem 5 (Asymptotic Consistency of the SNIS Gradient Estimator)**.** *Let $X^{(1)}, \ldots, X^{(B)}$ be i.i.d. samples drawn from the adversarial distribution $p_{adv}(\cdot; x_{ori}, y_{ori}, \theta)$. Define the unnormalized importance weights as $\tilde{w}(X; y_{ori}, \theta) = \exp(-c_2 f(X, y_{ori}; \theta))$. The SNIS empirical gradient estimator is given by:*

$$\hat{g}_B(\theta) = \sum_{i=1}^{B} \hat{w}^{(i)} \nabla_\theta f(X^{(i)}, y_{ori}; \theta) = \frac{\sum_{i=1}^{B} \tilde{w}(X^{(i)}; y_{ori}, \theta) \nabla_\theta f(X^{(i)}, y_{ori}; \theta)}{\sum_{i=1}^{B} \tilde{w}(X^{(i)}; y_{ori}, \theta)}$$

Assume the expected gradient under the distance distribution is finite, i.e., $\mathbb{E}_{X \sim p_{dis}(\cdot; x_{ori})}[|\nabla_\theta f(X, y_{ori}; \theta)|] < \infty$. As the batch size $B \to \infty$, the estimator $\hat{g}_B(\theta)$ converges almost surely to the true expected gradient under $p_{dis}(\cdot; x_{ori})$:

$$\hat{g}_B(\theta) \xrightarrow{a.s.} \mathbb{E}_{X \sim p_{dis}(\cdot; x_{ori})}[\nabla_\theta f(X, y_{ori}; \theta)]$$

The proof is provided in Appendix B.

5.4 Batch-Level Approximation and Implicit Regularization

While Theorem 5 assumes multiple perturbations per single image x_{ori}, this is computationally expensive. For efficiency, our practical implementation (Algorithm 2 in Appendix A) applies SNIS across a mini-batch of clean images $\{(x^{(i)}, y^{(i)})\} \sim p_{data}$, generating only one perturbation per image. This batch-level formulation alters the weighting scheme. A strict estimator over the joint distribution requires weighting each sample by an intractable local adversarial volume:

$$\overline{w}_{true}^{(i)} \propto Z_{adv}^{(i)} Z_{vic}^{(i)} \exp(-c_2 f(X^{(i)}, y^{(i)}; \theta))$$

By definition, $Z_{adv}^{(i)} Z_{vic}^{(i)} = \mathbb{E}_{U \sim p_{dis}(\cdot; x^{(i)})}[\exp(c_2 f(U, y^{(i)}; \theta))]$. For highly vulnerable samples, this large expected volume theoretically compensates for the sharp decay of $\exp(-c_2 f)$. Naively omitting the volume term while retaining a large attack temperature c_2 leads to **catastrophic weight collapse**: the weights of crucial hard samples are driven to near zero, forcing the Softmax to collapse onto the easiest samples and defeating the purpose of adversarial training.

To counteract this without calculating the intractable volume, we introduce a small smoothing factor β (e.g., $\beta = 0.001$) to replace c_2 in the final empirical weighting:

$$\tilde{w}_{emp}^{(i)} = \exp(-\beta f(X^{(i)}, y^{(i)}; \theta)) \tag{2}$$

This β acts as a conceptual temperature scaling that prevents weight collapse for valid hard samples. Concurrently, retaining the negative exponential structure preserves the algorithmic benefit of omitting $Z_{adv}^{(i)} Z_{vic}^{(i)}$: **implicit regularization against robust overfitting**. Outliers or mislabeled data, which theoretically possess exploding local volumes, are penalized. This extends our core philosophy of down-weighting extreme worst-case outliers to the global data manifold. The theoretical analysis of this batch-level estimator is provided in Appendix C.

6 Experiments

Due to space constraints, this section focuses on the main evaluation results. Comprehensive details regarding the experimental settings, datasets, and hyperparameters are deferred to Appendix E.

6.1 Probabilistic Robustness

Table 1 summarizes the performance of PAT and baselines (PGD (Madry et al., 2018), ALP (Kannan et al., 2018), CLP (Kannan et al., 2018), TRADES (Zhang et al., 2019), MART (Wang et al., 2019) and AT-PR (Zhang et al., 2025)) on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). PAT consistently achieves superior Probabilistic Robustness (PR) across all evaluated architectures (WRN-50-2, ResNet-18) and datasets, with the advantage becoming more pronounced at larger perturbation scales.

While PAT shows a nominal decrease in strictly worst-case robustness (PGD/CW Acc.) compared to unweighted AT, this directly aligns with our theory. Unlike standard AT which over-fits to anomalous worst-case boundaries, PAT's theoretically derived importance weight naturally penalizes extreme outliers. This optimizes overall distributional security, effectively pushing the distance and victim distributions apart.

6.2 Ablation Study

Table 2 isolates the impact of our theoretically derived importance weight through two variants: **PAT(WOS)** (Langevin sampling without the importance weight) and **PGD-COR/FGSM-COR** (standard attacks equipped with our theoretical scaling factor).

PAT consistently outperforms PAT(WOS). On CIFAR-100 (WRN-50-2), removing the importance weight drops PR(0.2) from 67.12% to 63.98%. This confirms that sampling from p_{adv} alone is insufficient; down-weighting extreme outliers via our theoretical formulation is an essential requirement for maximizing probabilistic robustness.

Our importance weight acts as a versatile plug-in regularizer. Applying it to standard PGD (PGD-COR) improves PR(0.2) on CIFAR-100 from 58.99% to 62.21%. Notably, the performance of

Table 1: Comprehensive evaluation of Probabilistic Robustness (PR) and test accuracies across CIFAR-10 and CIFAR-100 datasets. PR is evaluated under escalating perturbation scales

D.	M.	Attack	Acc. %	PGD %	CW %	PR(·, 0.1) %	PR(·, 0.12) %	PR(·, 0.15) %	PR(·, 0.2) %
CIFAR-10	ResNet-18	Clean	**93.75**	0.00	0.00	49.82	36.62	24.19	16.36
		FGSM	87.16	0.01	0.05	88.88	84.23	75.83	59.91
		PGD	82.00	45.80	44.94	95.26	92.55	87.72	77.03
		ALP	81.24	46.65	45.56	95.36	92.88	88.16	77.94
		CLP	83.23	44.11	43.34	95.37	93.03	88.47	78.77
		TRADES	79.62	48.78	45.91	94.33	91.97	87.33	77.33
		MART	78.56	**51.30**	**47.02**	94.26	91.86	87.36	78.29
		AT-PR	80.33	40.58	39.23	95.25	92.80	88.10	78.22
		PAT	82.44	32.53	30.51	**95.53**	**93.43**	**89.26**	**80.62**
	WRN-50-2	Clean	**91.89**	0.00	0.00	56.99	44.66	31.52	20.85
		FGSM	87.60	0.00	0.00	79.79	67.51	44.35	21.04
		PGD	76.49	46.54	**44.33**	95.59	93.56	89.44	79.98
		ALP	72.45	44.42	41.50	95.71	93.76	89.71	81.23
		CLP	79.61	37.71	36.41	95.86	93.69	88.94	78.70
		TRADES	74.03	44.38	40.82	94.72	92.62	88.48	79.89
		MART	69.87	**46.83**	42.29	94.93	92.83	89.26	81.93
		AT-PR	75.93	40.03	39.21	95.47	93.61	89.61	80.23
		PAT	78.26	31.94	29.91	**95.98**	**93.96**	**90.96**	**82.19**
CIFAR-100	ResNet-18	Clean	**74.61**	0.00	0.00	27.45	19.20	11.93	7.08
		FGSM	59.56	0.00	0.00	64.50	53.45	38.65	21.62
		PGD	54.50	22.48	21.18	89.27	83.31	71.01	46.90
		ALP	54.88	23.83	21.98	90.45	84.69	72.95	49.35
		CLP	55.71	22.08	19.28	90.05	84.80	74.50	54.53
		TRADES	53.56	26.74	22.66	89.07	84.23	74.74	56.24
		MART	52.13	**26.88**	**23.51**	90.02	84.86	74.66	51.76
		AT-PR	52.67	21.48	20.28	89.09	83.14	71.18	47.64
		PAT	53.54	17.50	14.87	**90.77**	**86.58**	**78.07**	**59.78**
	WRN-50-2	Clean	**70.41**	0.00	0.00	25.51	17.95	11.23	6.24
		FGSM	51.04	0.00	0.00	39.19	25.26	13.12	5.26
		PGD	49.55	24.04	**22.00**	92.06	87.94	78.85	58.99
		ALP	44.38	23.10	20.80	92.07	88.07	79.92	62.17
		CLP	52.95	19.15	16.94	90.91	86.36	76.20	56.99
		TRADES	49.85	23.01	19.18	90.64	86.46	78.32	62.03
		MART	42.61	**24.67**	21.15	92.03	88.52	81.41	66.51
		AT-PR	48.47	22.02	20.98	92.11	88.08	79.87	61.58
		PAT	49.95	17.26	15.98	**92.28**	**89.49**	**82.18**	**67.12**

FGSM-COR exhibits more irregularity. This discrepancy is theoretically expected: as a deterministic, single-step attack, FGSM lacks the iterative exploration and stochasticity required for the importance weight to effectively evaluate the local adversarial distribution. Nonetheless, the consistent gains in PGD-COR demonstrate the broader utility of our weighting scheme in improving the distributional robustness of traditional iterative defense methods.

7 Related Work

Adversarial Training for Worst-Case Robustness. Traditional adversarial training (AT) formulates a min-max optimization problem to defend against worst-case perturbations (Madry et al., 2018). A vast literature has developed around improving this worst-case robustness (WCR), including regularization techniques like TRADES (Zhang et al., 2019) and MART (Wang et al., 2019), logit pairing methods such as ALP and CLP (Kannan et al., 2018), and efficient training variants like Fast-AT (Wong et al., 2020) and Free-AT (Shafahi et al., 2019). Complementary to our work, Li & Li (2024) theoretically analyzed the feature learning dynamics of AT. However, these WCR-focused methods inherently optimize for extreme, deterministic adversarial outliers, which frequently compromises standard accuracy and broader distributional security (Zhao, 2026).

Probabilistic Robustness Assessment. To bridge the gap between extreme-case vulnerability and practical reliability, Webb et al. (2018) formally defined Probabilistic Robustness (PR) to evaluate the overall likelihood of adversarial examples within a local neighborhood. Following this, significant progress has been made in PR assessment, including scalable statistical estimators (Tit et al.,

Table 2: Ablation results for the importance weight across various architectures. PAT(WOS) removes the weighting factor; COR denotes the application of our scaling factor to standard baselines.

D.	M.	Attack	Acc. %	PGD %	CW %	PR(·, 0.1) %	PR(·, 0.12) %	PR(·, 0.15) %	PR(·, 0.2) %
CIFAR-10	ResNet-18	FGSM	87.16	0.01	**0.05**	88.88	84.23	**75.83**	**59.91**
		FGSM-COR	**89.04**	**0.02**	0.03	**89.03**	**84.53**	75.33	56.45
		PGD	**82.00**	45.80	**44.94**	**95.26**	**92.55**	**87.72**	**77.03**
		PGD-COR	81.75	**45.81**	44.72	95.03	92.38	87.33	75.97
		PAT(WOS)	**82.96**	32.15	30.34	94.86	92.56	88.36	79.81
		PAT	82.44	**32.53**	**30.51**	**95.53**	**93.43**	**89.26**	**80.62**
	WRN-50-2	FGSM	**87.60**	0.00	0.00	79.79	67.51	44.35	21.04
		FGSM-COR	76.46	**42.63**	**41.01**	**93.42**	**90.26**	**84.74**	**74.40**
		PGD	**76.49**	**46.54**	**44.33**	95.59	93.56	89.44	79.98
		PGD-COR	73.45	45.11	42.71	**95.79**	**93.85**	**89.79**	**80.62**
		PAT(WOS)	**79.06**	**31.99**	29.63	94.63	92.29	87.91	79.25
		PAT	78.26	31.94	**29.91**	**95.98**	**93.96**	**90.96**	**82.19**
CIFAR-100	ResNet-18	FGSM	**59.56**	**0.00**	0.00	**64.50**	**53.45**	**38.65**	**21.62**
		FGSM-COR	56.20	**0.00**	**0.01**	62.68	51.06	35.49	18.52
		PGD	54.50	22.48	21.18	89.27	83.31	71.01	46.90
		PGD-COR	**55.02**	**23.42**	**22.06**	**90.11**	**84.23**	**72.39**	**47.92**
		PAT(WOS)	53.21	16.84	14.53	90.47	86.24	**78.07**	**59.83**
		PAT	**53.54**	**17.50**	**14.87**	**90.77**	**86.58**	**78.07**	59.78
	WRN-50-2	FGSM	51.04	**0.00**	**0.00**	39.19	25.26	**13.12**	**5.26**
		FGSM-COR	**52.50**	**0.00**	**0.00**	**41.66**	**26.31**	11.75	3.88
		PGD	**49.55**	**24.04**	**22.00**	92.06	87.94	78.85	58.99
		PGD-COR	44.87	22.74	20.46	**92.66**	**88.79**	**80.55**	**62.21**
		PAT(WOS)	**50.17**	16.92	14.96	91.65	87.91	80.22	63.98
		PAT	49.95	**17.26**	**15.98**	**92.28**	**89.49**	**82.18**	**67.12**

2021; Karim et al., 2023), probabilistic verification approaches (Weng et al., 2019), and extensions to functional perturbations (Zhang et al., 2022). A comprehensive survey of these evaluation metrics is provided by Zhao (Zhao, 2026). While evaluating PR has become increasingly sophisticated, dedicated training frameworks designed to directly optimize it remain sparse.

Defenses Targeting Probabilistic Robustness. Only a few recent works have explicitly aimed at improving PR during training. Wang et al. (2021) demonstrated that adding random perturbations enhances PR, though it yields near-zero worst-case robustness. Other approaches (Robey et al., 2022; Zhang et al., 2024b) proposed training with risk measures such as Conditional Value-at-Risk (CVaR) or Entropic Value-at-Risk (EVaR), but also struggled to maintain competitive standard adversarial robustness. The most direct comparison to our work is the recent AT-PR method (Zhang et al., 2025), which explicitly alters the inner maximization of AT to target PR. However, their approach relies on a heuristic, gradient-based boundary search to empirically find the "widest peak" in the loss landscape. In contrast, our Probabilistic Adversarial Training (PAT) establishes a rigorous theoretical foundation: we formalize untargeted probabilistic attacks via Langevin Dynamics, prove a KL-based lower bound for PR, and employ Self-Normalized Importance Sampling (SNIS) to derive a fully tractable, principled optimization objective.

8 Conclusion

In this work, we provided a unified probabilistic perspective on adversarial defense by proving that $KL(p_{dis}||p_{vic}) - \log Z_{vic}$ serves as a lower bound for probabilistic robustness (PR). By maximizing this tractable surrogate objective via SNIS applied to gradient, we introduced Probabilistic Adversarial Training (PAT). Our theoretical framework not only offers a principled probabilistic interpretation of standard adversarial training but also yields a natural importance-weighting mechanism that penalizes extreme worst-case outliers. Empirically, PAT consistently enhances PR across various architectures. We hope this probabilistic perspective of adversarial vulnerabilities inspires more robust and theoretically grounded defense strategies in the future.

AI USE STATEMENT

We used generative AI tools to assist with translating, language polishing and identifying related work. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work aims to improve the robustness of neural networks against adversarial perturbations by providing a probabilistic interpretation of adversarial training and a tractable robustness-improving objective. Its potential positive impact lies in improving the reliability and security of machine learning systems, especially in safety-critical applications where small input perturbations may lead to incorrect predictions.

At the same time, adversarial robustness research is inherently dual-use. The probabilistic attack perspective studied in this work may also help adversaries better understand how to generate effective adversarial examples. However, the main contribution of this paper is defensive: it uses this perspective to derive Probabilistic Adversarial Training, a method designed to improve probabilistic robustness. We do not release tools intended for malicious attack deployment, and the experiments are conducted on standard benchmark datasets.

REPRODUCIBILITY STATEMENT

Our code is available at `https://anonymous.4open.science/r/PATICLR27/README.md`. Experimental settings, hyperparameters, and evaluation protocols are detailed in Appendix E.

REFERENCES

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.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6572`.

Pilar L Iglesias, Carlos AB Pereira, and Nelson I Tanaka. Characterizations of multivariate spherical distributions in l infinity norm. *Test*, 7(2):307–324, 1998.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

TIT Karim, Teddy Furon, and Mathias Rousset. Gradient-informed neural network statistical robustness estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 323–334. PMLR, 2023.

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

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=BJm4T4Kgx`.

Andrew Lamperski. Projected stochastic gradient langevin algorithms for constrained sampling and non-convex learning. In *Conference on Learning Theory*, pp. 2891–2937. PMLR, 2021.

Binghui Li and Yuanzhi Li. Adversarial training can provably improve robustness: Theoretical analysis of feature learning process under structured data. *arXiv preprint arXiv:2410.08503*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

Alexander Robey, Luiz Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average and worst-case performance. In *International Conference on Machine Learning*, pp. 18667–18686. PMLR, 2022.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Karim Tit, Teddy Furon, and Mathias Rousset. Efficient statistical assessment of neural network corruption robustness. *Advances in Neural Information Processing Systems*, 34:9253–9263, 2021.

Benjie Wang, Stefan Webb, and Tom Rainforth. Statistically robust neural network classification. In *Uncertainty in Artificial Intelligence*, pp. 1735–1745. PMLR, 2021.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019.

Stefan Webb, Tom Rainforth, Yee Whye Teh, and M Pawan Kumar. A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*, 2018.

Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. Proven: Verifying robustness of neural networks with a probabilistic approach. In *International Conference on Machine Learning*, pp. 6727–6736. PMLR, 2019.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Andi Zhang, Mingtian Zhang, and Damon Wischik. Constructing semantics-aware adversarial examples with a probabilistic perspective. *Advances in Neural Information Processing Systems*, 37: 136259–136285, 2024a.

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.

Tianle Zhang, Wenjie Ruan, and Jonathan E Fieldsend. Proa: A probabilistic robustness assessment against functional perturbations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 154–170. Springer, 2022.

Tianle Zhang, Yanghao Zhang, Ronghui Mu, Jiaxu Liu, Jonathan E Fieldsend, and Wenjie Ruan. Prass: Probabilistic risk-averse robust learning with stochastic search. In *IJCAI*, pp. 559–567, 2024b.

Yi Zhang, Yuhang Chen, Zhen Chen, Wenjie Ruan, Xiaowei Huang, Siddartha Khastgir, and Xingyu Zhao. Adversarial training for probabilistic robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1675–1685, 2025.

Xingyu Zhao. Probabilistic robustness in deep learning: A concise yet comprehensive guide. In *Adversarial Example Detection and Mitigation Using Machine Learning*, pp. 209–222. Springer, 2026.

APPENDIX

A PSEUDO CODE OF THE ALGORITHMS

Algorithm 1 Untargeted Probabilistic Adversarial Attack

Input: Original image $x_{\text{ori}} \in [0,1]^d$, original class y_{ori}, cross-entropy loss f corresponding to the victim classifier, parameters $c_1, c_2 > 0$, step size η, noise scale σ, gradient clipping threshold $\rho > 0$, total timesteps T.
Output: Adversarial example x_{adv}.
$x_{\text{adv}} \sim \text{Uniform}[0,1]^d$ ▷ Initialize
for $t = 1$ **to** T **do**
  $z_t \sim \mathcal{N}(0, I)$ ▷ Generate noise for Langevin Dynamics
  $x_{\text{adv}} \leftarrow \Pi_{[0,1]^d}\left(x_{\text{adv}} + \sigma z_t\right)$ ▷ Apply noise
  $\mathcal{E} \leftarrow c_1 \left\|x_{\text{adv}} - x_{\text{ori}}\right\|_2^2 - c_2\, f(x_{\text{adv}}, y_{\text{ori}})$ ▷ Calculate energy
  $v_t \leftarrow \text{clip}(\nabla_{x_{\text{adv}}} \mathcal{E}, -\rho, \rho)$ ▷ Compute and clip gradient
  $x_{\text{adv}} \leftarrow \Pi_{[0,1]^d}\left(x_{\text{adv}} - \eta v_t\right)$ ▷ update x_{adv}
end for
return x_{adv}.

Algorithm 2 Probabilistic Adversarial Training

Input: Training dataset $\mathcal{D}$, initial model parameters θ, cross-entropy loss f, learning rate α.
Hyperparameters: Total epochs E_{total}, batch size B, inner steps T, step size η, noise scale σ, gradient clip ρ, energy weights c_1, c_2, smoothing factor β.
Output: Robust model parameters θ.
for epoch $e = 0$ **to** $E_{\text{total}} - 1$ **do**
  for each minibatch $(\mathbf{X}, \mathbf{Y}) \sim \mathcal{D}$ with size B **do**
    ▷ **Phase 1: Probabilistic Adversarial Example Generation**
    $\mathbf{X}_{\text{adv}} \sim \text{Uniform}[0,1]^{B \times d}$ ▷ Initialize
    for $t = 1$ **to** T **do**
      $\mathbf{Z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})^{B \times d}$ ▷ Sample noise for Langevin Dynamics
      $\mathbf{X}_{\text{adv}} \leftarrow \Pi_{[0,1]^{B \times d}}\left(\mathbf{X}_{\text{adv}} + \sigma \mathbf{Z}_t\right)$ ▷ Apply noise
      $\mathcal{E} \leftarrow \sum_{i=1}^{B} \left(c_1 \left\|\mathbf{X}_{\text{adv}}^{(i)} - \mathbf{X}^{(i)}\right\|_2^2 - c_2\, f_\theta(\mathbf{X}_{\text{adv}}^{(i)}, \mathbf{Y}^{(i)})\right)$ ▷ Calculate batch energy
      $\mathbf{V}_t \leftarrow \text{clip}(\nabla_{\mathbf{X}_{\text{adv}}} \mathcal{E}, -\rho, \rho)$ ▷ Compute and clip gradient
      $\mathbf{X}_{\text{adv}} \leftarrow \Pi_{[0,1]^{B \times d}}\left(\mathbf{X}_{\text{adv}} - \eta \mathbf{V}_t\right)$ ▷ Update $\mathbf{X}_{\text{adv}}$
    end for
    ▷ **Phase 2: Model Parameter Update via Gradient Descent**
    $\boldsymbol{\ell}_i \leftarrow f_\theta(\mathbf{X}_{\text{adv}}^{(i)}, \mathbf{Y}^{(i)})$ ▷ Compute sample losses $\boldsymbol{\ell} \in \mathbb{R}^B$
    $\tilde{\mathbf{w}}_{\text{pvic}} \leftarrow \text{Detach}(-\beta\, \boldsymbol{\ell})$ ▷ Apply smoothing factor and stop gradient
    $\hat{\mathbf{w}}_{\text{pvic}} \leftarrow \text{Softmax}(\tilde{\mathbf{w}}_{\text{pvic}})$ ▷ Softmax over batch dimension
    $\mathcal{L} \leftarrow \boldsymbol{\ell}^\top \hat{\mathbf{w}}_{\text{pvic}}$ ▷ Final weighted loss
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$ ▷ Update model parameters
  end for
end for
return θ.

B PROOFS OF THEOREMS IN THE MAIN TEXT

Proposition 1 (Well-definedness of the untargeted victim distribution)**.** *Let $K = [0,1]^d$ and let $f(\cdot, y_{\text{ori}}) : K \to \mathbb{R}$ be continuous. For any $c > 0$, define*

$$p_{\text{vic}}(x; y_{\text{ori}}) = \frac{\exp\left(c f(x, y_{\text{ori}})\right)}{\int_K \exp\left(c f(u, y_{\text{ori}})\right) du}, \qquad x \in K.$$

Then $p_{\text{vic}}(\cdot; y_{\text{ori}})$ is a well-defined probability density on K.

Proof. Since $K = [0,1]^d$ is compact and $f(\cdot, y_{\text{ori}})$ is continuous, $\exp(cf(\cdot, y_{\text{ori}}))$ is also continuous on K. Hence it is bounded and measurable. Moreover, it is strictly positive on K, so the normalizing constant

$$Z_{\text{vic}} = \int_K \exp\left(cf(u, y_{\text{ori}})\right) du$$

satisfies $0 < Z_{\text{vic}} < \infty$. Therefore p_{vic} is non-negative, measurable, and integrates to one over K. Thus it is a well-defined probability density. □

Theorem 2 (KL-Based Lower Bound on Probabilistic Robustness)**.** *Let $p_{\text{dis}}(\cdot; x_{\text{ori}})$ be a probability density on $[0,1]^d$. Suppose the victim distribution parameterized by θ is defined as*

$$p_{\text{vic}}(x; y_{\text{ori}}, \theta) = \frac{\exp(c_2 f(x, y_{\text{ori}}; \theta))}{Z_{\text{vic}}},$$

where $Z_{\text{vic}} = \int_{\mathcal{X}} \exp(c_2 f(u, y_{\text{ori}}; \theta)) du$ is the normalizing constant. Let γ_* be defined as above and assume $\gamma_* > 0$. Then, the probabilistic robustness satisfies:

$$\text{PR}(p_{\text{dis}}, s) \geq 1 - \frac{\log Z_{\text{vic}} - H(p_{\text{dis}}) - \text{KL}(p_{\text{dis}} \| p_{\text{vic}})}{\gamma_*},$$

where $H(p_{\text{dis}}) = -\int_{\mathcal{X}} p_{\text{dis}}(x) \log p_{\text{dis}}(x) dx$ is the differential entropy of the distance distribution.

Proof. By the definition of γ_*, for all $x \in E_{adv}$, we have $c_2 f(x, y_{ori}) \geq \gamma_*$. Thus, we can lower bound the expected loss under p_{dis}:

$$\mathbb{E}_{X \sim p_{dis}}[c_2 f(X, y_{ori})] \geq \int_{E_{adv}} p_{dis}(x) c_2 f(x, y_{ori}) dx \geq \gamma_* \int_{E_{adv}} p_{dis}(x) dx = \gamma_* \mathcal{I}[p_{dis}, s]$$

On the other hand, expanding the KL divergence with $p_{vic}(x; y_{ori}) = \frac{\exp(c_2 f(x, y_{ori}))}{Z_{vic}}$ yields:

$$\begin{aligned} KL(p_{dis} || p_{vic}) &= \int p_{dis}(x) \log p_{dis}(x) dx - \int p_{dis}(x) \log p_{vic}(x; y_{ori}) dx \\ &= -H(p_{dis}) - \mathbb{E}_{X \sim p_{dis}}[c_2 f(X, y_{ori}) - \log Z_{vic}] \\ &= -H(p_{dis}) + \log Z_{vic} - \mathbb{E}_{X \sim p_{dis}}[c_2 f(X, y_{ori})] \\ &\leq -H(p_{dis}) + \log Z_{vic} - \gamma_* \mathcal{I}[p_{dis}, s] \end{aligned}$$

Rearranging the inequality gives:

$$\mathcal{I}[p_{dis}, s] \leq \frac{\log Z_{vic} - H(p_{dis}) - KL(p_{dis} || p_{vic})}{\gamma_*}$$

Finally, by the definition of probabilistic robustness $PR(p_{dis}, s) = 1 - \mathcal{I}[p_{dis}, s]$, we conclude:

$$PR(p_{dis}, s) \geq 1 - \frac{\log Z_{vic} - H(p_{dis}) - KL(p_{dis} || p_{vic})}{\gamma_*}$$

□

Proposition 3 (Idealized Objective)**.** *Let p_{dis}, p_{vic}, and p_{adv} be defined as in Section 3. Maximizing the probabilistic robustness lower bound with respect to the classifier parameters θ is mathematically equivalent to minimizing the following idealized expected loss:*

$$\mathcal{J}_{ideal}(\theta) = \mathbb{E}_{X \sim p_{adv}(\cdot; x_{ori}, y_{ori}, \theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} c_2 f(X, y_{ori}; \theta) \right]$$

Proof. By expanding the KL divergence and changing the measure from p_{dis} to p_{adv}, we have:

$$
\begin{aligned}
&KL(p_{dis}(\cdot; x_{ori})||p_{vic}(\cdot; y_{ori}, \theta)) - \log Z_{vic} \\
&= \mathbb{E}_{X\sim p_{dis}(\cdot;x_{ori})}[\log p_{dis}(X; x_{ori})] - \int p_{dis}(x; x_{ori}) \log p_{vic}(x; y_{ori}, \theta) dx - \log Z_{vic} \\
&= -H(p_{dis}) - \int p_{adv}(x; x_{ori}, y_{ori}, \theta) \frac{Z_{adv} p_{dis}(x; x_{ori})}{p_{dis}(x; x_{ori}) p_{vic}(x; y_{ori}, \theta)} \log p_{vic}(x; y_{ori}, \theta) dx - \log Z_{vic} \\
&= -H(p_{dis}) - \mathbb{E}_{X\sim p_{adv}(\cdot;x_{ori},y_{ori},\theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} (c_2 f(X, y_{ori}; \theta) - \log Z_{vic}) \right] - \log Z_{vic} \\
&= -H(p_{dis}) - \mathbb{E}_{X\sim p_{adv}(\cdot;x_{ori},y_{ori},\theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} c_2 f(X, y_{ori}; \theta) \right] \\
&\quad + \log Z_{vic} \underbrace{\mathbb{E}_{X\sim p_{adv}(\cdot;x_{ori},y_{ori},\theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} \right]}_{=1} - \log Z_{vic} \\
&= -H(p_{dis}) - \mathbb{E}_{X\sim p_{adv}(\cdot;x_{ori},y_{ori},\theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} c_2 f(X, y_{ori}; \theta) \right]
\end{aligned}
$$

Since $H(p_{dis})$ is constant with respect to θ, maximizing the lower bound is equivalent to minimizing $\mathcal{J}_{ideal}(\theta)$. □

Proposition 4 (Gradient of the Objective)**.** *The exact gradient of the negative lower bound with respect to θ can be expressed as an expectation over the adversarial distribution p_{adv}:*

$$
\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{X\sim p_{adv}(\cdot;x_{ori},y_{ori},\theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} \nabla_\theta (c_2 f(X, y_{ori}; \theta)) \right]
$$

Proof. We return to the negative lower bound before importance sampling:

$$
-KL(p_{dis}||p_{vic}) + \log Z_{vic} = H(p_{dis}) + \mathbb{E}_{X\sim p_{dis}(\cdot;x_{ori})}[c_2 f(X, y_{ori}; \theta)].
$$

Since $p_{dis}(\cdot; x_{ori})$ depends only on the clean image x_{ori} and is entirely independent of θ, the gradient operator can pass directly inside the expectation:

$$
\begin{aligned}
&\nabla_\theta \left(\mathbb{E}_{X\sim p_{dis}(\cdot;x_{ori})}[c_2 f(X, y_{ori}; \theta)] \right) \\
=&\mathbb{E}_{X\sim p_{dis}(\cdot;x_{ori})}[\nabla_\theta c_2 f(X, y_{ori}; \theta)] \\
=& \int p_{adv}(x; x_{ori}, y_{ori}, \theta) \frac{p_{dis}(x; x_{ori})}{p_{adv}(x; x_{ori}, y_{ori}, \theta)} \nabla_\theta c_2 f(x, y_{ori}; \theta) dx \\
=&\mathbb{E}_{X\sim p_{adv}(\cdot;x_{ori},y_{ori},\theta)} \left[\frac{Z_{adv}}{p_{vic}(X; y_{ori}, \theta)} \nabla_\theta c_2 f(X, y_{ori}; \theta) \right]
\end{aligned}
$$

□

Theorem 5 (Asymptotic Consistency of the SNIS Gradient Estimator)**.** *Let $X^{(1)}, \ldots, X^{(B)}$ be i.i.d. samples drawn from the adversarial distribution $p_{adv}(\cdot; x_{ori}, y_{ori}, \theta)$. Define the unnormalized importance weights as $\tilde{w}(X; y_{ori}, \theta) = \exp(-c_2 f(X, y_{ori}; \theta))$. The SNIS empirical gradient estimator is given by:*

$$
\hat{g}_B(\theta) = \sum_{i=1}^{B} \hat{w}^{(i)} \nabla_\theta f(X^{(i)}, y_{ori}; \theta) = \frac{\sum_{i=1}^{B} \tilde{w}(X^{(i)}; y_{ori}, \theta) \nabla_\theta f(X^{(i)}, y_{ori}; \theta)}{\sum_{i=1}^{B} \tilde{w}(X^{(i)}; y_{ori}, \theta)}
$$

Assume the expected gradient under the distance distribution is finite, i.e., $\mathbb{E}_{X\sim p_{dis}(\cdot;x_{ori})}[|\nabla_\theta f(X, y_{ori}; \theta)|] < \infty$. As the batch size $B \to \infty$, the estimator $\hat{g}_B(\theta)$ converges almost surely to the true expected gradient under $p_{dis}(\cdot; x_{ori})$:

$$
\hat{g}_B(\theta) \xrightarrow{a.s.} \mathbb{E}_{X\sim p_{dis}(\cdot;x_{ori})}[\nabla_\theta f(X, y_{ori}; \theta)]
$$

Proof. By the Strong Law of Large Numbers (SLLN), as $B \to \infty$, the sample averages in the numerator and denominator of $\hat{g}_B(\theta)$ converge almost surely to their respective expectations under $p_{adv}(\cdot; x_{ori}, y_{ori}, \theta)$:

$$\frac{1}{B}\sum_{i=1}^{B} \tilde{w}(X^{(i)}; y_{ori}, \theta)\nabla_\theta f(X^{(i)}, y_{ori}; \theta) \xrightarrow{a.s.} \mathbb{E}_{X\sim p_{adv}}\left[\tilde{w}(X; y_{ori}, \theta)\nabla_\theta f(X, y_{ori}; \theta)\right]$$

$$\frac{1}{B}\sum_{i=1}^{B} \tilde{w}(X^{(i)}; y_{ori}, \theta) \xrightarrow{a.s.} \mathbb{E}_{X\sim p_{adv}}\left[\tilde{w}(X; y_{ori}, \theta)\right]$$

Recall the definitions from Section 3. We can express the product of the sampling density and the unnormalized weight as:

$$\begin{aligned} p_{adv}(X; x_{ori}, y_{ori}, \theta)\tilde{w}(X; y_{ori}, \theta) &= \frac{p_{dis}(X; x_{ori})\exp(c_2 f(X, y_{ori}; \theta))}{Z_{adv}Z_{vic}}\exp(-c_2 f(X, y_{ori}; \theta)) \\ &= \frac{p_{dis}(X; x_{ori})}{Z_{adv}Z_{vic}} \end{aligned}$$

Substituting this identity into the expectation of the numerator yields:

$$\begin{aligned} \mathbb{E}_{X\sim p_{adv}}\left[\tilde{w}(X; y_{ori}, \theta)\nabla_\theta f(X, y_{ori}; \theta)\right] &= \int \frac{p_{dis}(x; x_{ori})}{Z_{adv}Z_{vic}}\nabla_\theta f(x, y_{ori}; \theta)dx \\ &= \frac{1}{Z_{adv}Z_{vic}}\mathbb{E}_{X\sim p_{dis}}\left[\nabla_\theta f(X, y_{ori}; \theta)\right] \end{aligned}$$

Applying the same substitution to the denominator:

$$\mathbb{E}_{X\sim p_{adv}}\left[\tilde{w}(X; y_{ori}, \theta)\right] = \int \frac{p_{dis}(x; x_{ori})}{Z_{adv}Z_{vic}}dx = \frac{1}{Z_{adv}Z_{vic}}$$

By the Continuous Mapping Theorem (or Slutsky's Theorem), the limit of the ratio is the ratio of their limits. Notice that the intractable constant term perfectly cancels out:

$$\hat{g}_B(\theta) \xrightarrow{a.s.} \frac{\frac{1}{Z_{adv}Z_{vic}}\mathbb{E}_{X\sim p_{dis}}\left[\nabla_\theta f(X, y_{ori}; \theta)\right]}{\frac{1}{Z_{adv}Z_{vic}}} = \mathbb{E}_{X\sim p_{dis}(\cdot; x_{ori})}\left[\nabla_\theta f(X, y_{ori}; \theta)\right]$$

This confirms that the self-normalized gradient is asymptotically unbiased and converges to the true gradient of our theoretical objective. □

C Theoretical Analysis of the Batch-Level Estimator

Theorem 5 establishes consistency of the SNIS gradient estimator in an idealized regime: a single clean image x_{ori} with label y_{ori}, a batch of B perturbations from its adversarial distribution $p_{\text{adv}}(\cdot; x_{\text{ori}}, y_{\text{ori}}, \theta)$. Algorithm 2 differs in two respects: it draws a mini-batch of distinct clean images with one perturbation each, and it uses the smoothing factor β in place of c_2 in the empirical weights (Eq. (2)). Sec. 5.4 observed that restoring exactness of the joint estimator would require the intractable volume weights $w^{(i)}_{\text{true}}$. This section analyses the estimator Algorithm 2 actually computes and proves:

- it is consistent, with an explicitly characterized limit (Theorem 7);
- the limit is the exact gradient of an entropic-risk objective under a vulnerability-based reweighting of the data (Proposition 8, Corollary 9);
- this objective still lower-bounds probabilistic robustness for every admissible β, with Theorem 2 recovered as the tightest member of the resulting family (Proposition 10, Corollary 11);

Setup. Let $\mathcal{X} = [0,1]^d$. For each $y \in \mathcal{Y}$, the loss $f(\cdot, y; \theta)$ is continuous on the compact set $\mathcal{X}$, as in Proposition 1, and

$$0 \le f(u, y; \theta) \le F_{\max}(\theta) := \max_{y \in \mathcal{Y}} \max_{u \in \mathcal{X}} f(u, y; \theta) < \infty.$$

For a clean pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $t \ge 0$, let

$$M_t(x, y; \theta) := \mathbb{E}_{U \sim p_{\mathrm{dis}}(\cdot; x)} \left[\exp\big(t f(U, y; \theta)\big) \right].$$

With $\beta \in [0, c_2]$ and $\lambda := c_2 - \beta$, define the tilted distribution, volume ratio, and reweighted data distribution:

$$p_\lambda(u; x, y, \theta) := \frac{p_{\mathrm{dis}}(u; x) \exp\big(\lambda f(u, y; \theta)\big)}{M_\lambda(x, y; \theta)},$$
$$r(x, y; \theta) := \frac{M_\lambda(x, y; \theta)}{M_{c_2}(x, y; \theta)},$$
$$\tilde{p}(x, y; \theta) \propto p_{\mathrm{data}}(x, y)\, r(x, y; \theta).$$

The last is well-defined since

$$\mathbb{E}_{(x,y) \sim p_{\mathrm{data}}} [r(x, y; \theta)] \in \big[\exp\big(-\beta F_{\max}(\theta)\big), 1\big]$$

by Lemma 6 (iii). Define the entropic risk

$$R_\lambda(x, y; \theta) := \frac{1}{\lambda} \log M_\lambda(x, y; \theta), \qquad \lambda > 0,$$

and

$$R_0(x, y; \theta) := \mathbb{E}_{X \sim p_{\mathrm{dis}}(\cdot; x)} [f(X, y; \theta)].$$

The map $\lambda \mapsto R_\lambda(x, y; \theta)$ is continuous at 0. As in Theorem 5, the Langevin sampler is treated as an exact sampler of $p_{\mathrm{adv}}(\cdot; x, y, \theta)$ (non-asymptotic guarantees for projected Langevin dynamics); all statements are at a fixed θ.

Lemma 6 (Tilting and volume identities)**.** *Fix $(x, y) \in \mathcal{X} \times \mathcal{Y}$, θ, and $\beta \in [0, c_2]$, and let $\lambda = c_2 - \beta$. Then:*

(i) For every $t \ge 0$,

$$1 \le M_t(x, y; \theta) \le \exp\big(t F_{\max}(\theta)\big),$$

and $p_t(\cdot; x, y, \theta)$ is a valid probability density. In particular,

$$p_0(\cdot; x, y, \theta) = p_{\mathrm{dis}}(\cdot; x), \qquad p_{c_2}(\cdot; x, y, \theta) = p_{\mathrm{adv}}(\cdot; x, y, \theta),$$

with

$$Z_{\mathrm{adv}} Z_{\mathrm{vic}} = M_{c_2}(x, y; \theta),$$

which is the identity used in Section 5.4.

(ii) Pointwise on $\mathcal{X}$,

$$p_{\mathrm{adv}}(u; x, y, \theta) \exp\big(-\beta f(u, y; \theta)\big) = r(x, y; \theta)\, p_\lambda(u; x, y, \theta).$$

(iii)

$$\exp\big(-\beta F_{\max}(\theta)\big) \le r(x, y; \theta) \le 1,$$

and

$$r(x, y; \theta)^{-1} = \mathbb{E}_{U \sim p_\lambda(\cdot; x, y, \theta)} \left[\exp\big(\beta f(U, y; \theta)\big) \right].$$

Proof. (i) The function $\exp\big(t f(\cdot, y; \theta)\big)$ is continuous on the compact set $\mathcal{X}$, with values in $[1, \exp(t F_{\max}(\theta))]$. Hence

$$1 \le M_t(x, y; \theta) \le \exp\big(t F_{\max}(\theta)\big),$$

and $p_t(\cdot; x, y, \theta)$ integrates to one. The case $t = 0$ gives

$$p_0(\cdot; x, y, \theta) = p_{\mathrm{dis}}(\cdot; x).$$

Since

$$p_{\mathrm{adv}}(u; x, y, \theta) = \frac{p_{\mathrm{dis}}(u; x)\exp\big(c_2 f(u, y; \theta)\big)}{Z_{\mathrm{adv}} Z_{\mathrm{vic}}},$$

integrating over $\mathcal{X}$ gives

$$Z_{\mathrm{adv}} Z_{\mathrm{vic}} = M_{c_2}(x, y; \theta),$$

and therefore

$$p_{\mathrm{adv}}(\cdot; x, y, \theta) = p_{c_2}(\cdot; x, y, \theta).$$

(ii) Using (i) and $\lambda = c_2 - \beta$,

$$\begin{aligned} &p_{\mathrm{adv}}(u; x, y, \theta)\exp\big(-\beta f(u, y; \theta)\big) \\ &\quad = \frac{p_{\mathrm{dis}}(u; x)\exp\big(\lambda f(u, y; \theta)\big)}{M_{c_2}(x, y; \theta)} \\ &\quad = \frac{M_\lambda(x, y; \theta)}{M_{c_2}(x, y; \theta)} \cdot \frac{p_{\mathrm{dis}}(u; x)\exp\big(\lambda f(u, y; \theta)\big)}{M_\lambda(x, y; \theta)} \\ &\quad = r(x, y; \theta)\, p_\lambda(u; x, y, \theta). \end{aligned}$$

(iii) Since $f(u, y; \theta) \geq 0$ and $\lambda \leq c_2$,

$$\exp\big(\lambda f(u, y; \theta)\big) \leq \exp\big(c_2 f(u, y; \theta)\big).$$

Thus

$$M_\lambda(x, y; \theta) \leq M_{c_2}(x, y; \theta), \qquad r(x, y; \theta) \leq 1.$$

Conversely,

$$\begin{aligned} M_{c_2}(x, y; \theta) &= \mathbb{E}_{U \sim p_{\mathrm{dis}}(\cdot; x)}\left[\exp\big(\lambda f(U, y; \theta)\big)\exp\big(\beta f(U, y; \theta)\big)\right] \\ &\leq \exp\big(\beta F_{\max}(\theta)\big) M_\lambda(x, y; \theta), \end{aligned}$$

so

$$r(x, y; \theta) \geq \exp\big(-\beta F_{\max}(\theta)\big).$$

Finally,

$$\mathbb{E}_{U \sim p_\lambda(\cdot; x, y, \theta)}\left[\exp\big(\beta f(U, y; \theta)\big)\right] = \frac{M_{c_2}(x, y; \theta)}{M_\lambda(x, y; \theta)} = r(x, y; \theta)^{-1}.$$

□

Part (iii) makes the mechanism transparent: the local adversarial volume $Z_{\mathrm{adv}} Z_{\mathrm{vic}} = M_{c_2}(x, y; \theta)$ omitted from the empirical weights resurfaces in the large-batch limit as the per-image factor

$$r(x, y; \theta) = \frac{M_\lambda(x, y; \theta)}{M_{c_2}(x, y; \theta)},$$

which down-weights a clean pair exactly according to the exponential moment of its loss over its perturbation neighborhood. This is the precise form of the implicit regularization described informally in Section 5.4.

Theorem 7 (Batch-level consistency of the empirical gradient estimator)**.** *Let*

$$(x^{(i)}, y^{(i)}) \overset{\text{i.i.d.}}{\sim} p_{\mathrm{data}}, \qquad i = 1, \ldots, B,$$

and, conditionally on these clean pairs, let

$$X^{(i)} \sim p_{\mathrm{adv}}\big(\cdot; x^{(i)}, y^{(i)}, \theta\big)$$

independently across i. Equivalently, the tuples $(x^{(i)}, y^{(i)}, X^{(i)})$ are i.i.d. from the joint density

$$\pi(x, y, u) := p_{\mathrm{data}}(x, y)\, p_{\mathrm{adv}}(u; x, y, \theta).$$

Let

$$\hat{g}_B^{\mathrm{emp}}(\theta) = \sum_{i=1}^{B} \hat{w}_{\mathrm{emp}}^{(i)} \nabla_\theta f(X^{(i)}, y^{(i)}; \theta)$$

be the update of Lines 16-19 of Algorithm 2, with self-normalized weights

$$\hat{w}_{\mathrm{emp}}^{(i)} = \frac{\tilde{w}_{\mathrm{emp}}^{(i)}}{\sum_{j=1}^{B} \tilde{w}_{\mathrm{emp}}^{(j)}}, \qquad \tilde{w}_{\mathrm{emp}}^{(i)} := \exp\bigl(-\beta f(X^{(i)}, y^{(i)}; \theta)\bigr),$$

treated as constants during backpropagation (stop-gradient, Line 17). If

$$\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[\mathbb{E}_{X\sim p_{\mathrm{adv}}(\cdot;x,y,\theta)}\left[\|\nabla_\theta f(X,y;\theta)\|\right]\right] < \infty,$$

then, as $B \to \infty$,

$$\hat{g}_B^{\mathrm{emp}}(\theta) \xrightarrow{\text{a.s.}} \mathbb{E}_{(x,y)\sim\tilde{p}(\cdot,\cdot;\theta)}\left[\mathbb{E}_{X\sim p_\lambda(\cdot;x,y,\theta)}\left[\nabla_\theta f(X,y;\theta)\right]\right].$$

Proof. Write $\hat{g}_B^{\mathrm{emp}}(\theta)$ as a ratio of sample means:

$$\hat{g}_B^{\mathrm{emp}}(\theta) = \frac{\frac{1}{B}\sum_{i=1}^{B} \tilde{w}_{\mathrm{emp}}^{(i)} \nabla_\theta f(X^{(i)}, y^{(i)}; \theta)}{\frac{1}{B}\sum_{j=1}^{B} \tilde{w}_{\mathrm{emp}}^{(j)}}.$$

Only independence across tuples is used. It holds because the clean pairs are i.i.d. and, conditionally on them, the B Langevin chains evolve independently: the batch energy in Line 11 is separable across i, so its gradient decouples blockwise, and the noise in Line 9 is independent across i. The within-tuple dependence is exactly what the factorization

$$\pi(x,y,u) = p_{\mathrm{data}}(x,y)\, p_{\mathrm{adv}}(u;x,y,\theta)$$

encodes.

Denominator. Since

$$\tilde{w}_{\mathrm{emp}}^{(j)} \in \left[\exp\bigl(-\beta F_{\max}(\theta)\bigr), 1\right],$$

the weights are integrable. By the Strong Law of Large Numbers (SLLN), the tower property, and Lemma 6 (ii),

$$\begin{aligned}
\frac{1}{B}\sum_{j=1}^{B} \tilde{w}_{\mathrm{emp}}^{(j)} &\xrightarrow{\text{a.s.}} \mathbb{E}_{(x,y,X)\sim\pi}\left[\exp\bigl(-\beta f(X,y;\theta)\bigr)\right] \\
&= \mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[\mathbb{E}_{X\sim p_{\mathrm{adv}}(\cdot;x,y,\theta)}\left[\exp\bigl(-\beta f(X,y;\theta)\bigr)\right]\right] \\
&= \mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[r(x,y;\theta)\right] \\
&\geq \exp\bigl(-\beta F_{\max}(\theta)\bigr) > 0,
\end{aligned}$$

where the lower bound follows from Lemma 6 (iii).

Numerator. Since

$$\left\|\tilde{w}_{\mathrm{emp}}^{(i)} \nabla_\theta f(X^{(i)}, y^{(i)}; \theta)\right\| \leq \left\|\nabla_\theta f(X^{(i)}, y^{(i)}; \theta)\right\|,$$

the integrability assumption allows the SLLN to be applied componentwise:

$$\begin{aligned}
&\frac{1}{B}\sum_{i=1}^{B} \tilde{w}_{\mathrm{emp}}^{(i)} \nabla_\theta f(X^{(i)}, y^{(i)}; \theta) \\
&\quad\xrightarrow{\text{a.s.}} \mathbb{E}_{(x,y,X)\sim\pi}\left[\exp\bigl(-\beta f(X,y;\theta)\bigr)\nabla_\theta f(X,y;\theta)\right].
\end{aligned}$$

By the tower property and Lemma 6 (ii),

$$\begin{aligned}
&\mathbb{E}_{X\sim p_{\mathrm{adv}}(\cdot;x,y,\theta)}\left[\exp\bigl(-\beta f(X,y;\theta)\bigr)\nabla_\theta f(X,y;\theta)\right] \\
&\quad = r(x,y;\theta)\, \mathbb{E}_{X\sim p_\lambda(\cdot;x,y,\theta)}\left[\nabla_\theta f(X,y;\theta)\right].
\end{aligned}$$

Thus the numerator converges almost surely to

$$\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[r(x,y;\theta)\, \mathbb{E}_{X\sim p_\lambda(\cdot;x,y,\theta)}\left[\nabla_\theta f(X,y;\theta)\right]\right].$$

Ratio. On the probability-one event where both limits hold, the denominator tends to a strictly positive constant by Lemma 6 (iii). The continuous mapping theorem therefore gives

$$\begin{aligned}\hat{g}_B^{\mathrm{emp}}(\theta) &\xrightarrow{\text{a.s.}} \frac{\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[r(x,y;\theta)\,\mathbb{E}_{X\sim p_\lambda(\cdot;x,y,\theta)}\left[\nabla_\theta f(X,y;\theta)\right]\right]}{\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[r(x,y;\theta)\right]} \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{p_{\mathrm{data}}(x,y)\,r(x,y;\theta)}{\mathbb{E}_{(x',y')\sim p_{\mathrm{data}}}\left[r(x',y';\theta)\right]} \\ &\qquad \times \mathbb{E}_{X\sim p_\lambda(\cdot;x,y,\theta)}\left[\nabla_\theta f(X,y;\theta)\right]\,\mathrm{d}(x,y) \\ &= \mathbb{E}_{(x,y)\sim\tilde{p}(\cdot,\cdot;\theta)}\left[\mathbb{E}_{X\sim p_\lambda(\cdot;x,y,\theta)}\left[\nabla_\theta f(X,y;\theta)\right]\right],\end{aligned}$$

where the last equality holds because the normalizing constant of

$$\tilde{p}(x,y;\theta) \propto p_{\mathrm{data}}(x,y)\,r(x,y;\theta)$$

is exactly

$$\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[r(x,y;\theta)\right].$$

□

Relation to Theorem 5.

(i) The integrability assumption is implied by the data-averaged assumption of Theorem 5. By Lemma 6 (i),

$$\frac{p_{\mathrm{adv}}(u;x,y,\theta)}{p_{\mathrm{dis}}(u;x)} = \frac{\exp\big(c_2 f(u,y;\theta)\big)}{M_{c_2}(x,y;\theta)} \le \exp\big(c_2 F_{\max}(\theta)\big),$$

since $M_{c_2}(x,y;\theta) \ge 1$. Hence

$$\begin{aligned}&\mathbb{E}_{(x,y,X)\sim\pi}\left[\|\nabla_\theta f(X,y;\theta)\|\right] \\ &\quad\le \exp\big(c_2 F_{\max}(\theta)\big)\,\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[\mathbb{E}_{X\sim p_{\mathrm{dis}}(\cdot;x)}\left[\|\nabla_\theta f(X,y;\theta)\|\right]\right].\end{aligned}$$

(ii) **Theorem 7 strictly generalizes Theorem 5:** with a single clean pair $(x_{\mathrm{ori}}, y_{\mathrm{ori}})$ and $\beta = c_2$, the factor $r(x_{\mathrm{ori}}, y_{\mathrm{ori}};\theta)$ is a common constant that cancels in the self-normalized ratio. Moreover, $\lambda = 0$ and

$$p_0(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta) = p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}).$$

The limit therefore reduces to

$$\mathbb{E}_{X\sim p_{\mathrm{dis}}(\cdot;x_{\mathrm{ori}})}\left[\nabla_\theta f(X,y_{\mathrm{ori}};\theta)\right],$$

which is the conclusion of Theorem 5.

(iii) At the other endpoint $\beta = 0$, the weights are uniform:

$$\hat{w}_{\mathrm{emp}}^{(i)} = \frac{1}{B}.$$

In this case,

$$r(x,y;\theta) = 1, \qquad \tilde{p}(x,y;\theta) = p_{\mathrm{data}}(x,y),$$

and

$$p_\lambda(\cdot;x,y,\theta) = p_{\mathrm{adv}}(\cdot;x,y,\theta).$$

This is unweighted adversarial training on Langevin samples, i.e., exactly the PAT(WOS) variant of Section 6.2.

(iv) Had the intractable weights

$$w_{\mathrm{true}}^{(i)} \propto Z_{\mathrm{adv}}^{(i)} Z_{\mathrm{vic}}^{(i)} \exp\big(-c_2 f(X^{(i)}, y^{(i)};\theta)\big)$$

of Section 5.4 been used, the same argument, using the identity

$$p_{\mathrm{adv}}(u;x,y,\theta)\,Z_{\mathrm{adv}} Z_{\mathrm{vic}}\,\exp\big(-c_2 f(u,y;\theta)\big) = p_{\mathrm{dis}}(u;x)$$

from Lemma 6, would give the limit

$$\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[\mathbb{E}_{X\sim p_{\mathrm{dis}}(\cdot;x)}\left[\nabla_\theta f(X,y;\theta)\right]\right],$$

the exact joint analogue of Theorem 5.

Proposition 8 (Entropic-risk gradient identity)**.** *Fix $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Assume, as in Proposition 4, that ∇_θ may be exchanged with integration over $\mathcal{X}$. Then, for $\lambda > 0$,*

$$\nabla_\theta R_\lambda(x, y; \theta) = \mathbb{E}_{X \sim p_\lambda(\cdot; x, y, \theta)} \left[\nabla_\theta f(X, y; \theta) \right].$$

The identity extends to $\lambda = 0$, with

$$R_0(x, y; \theta) = \mathbb{E}_{X \sim p_{\mathrm{dis}}(\cdot; x)} \left[f(X, y; \theta) \right], \qquad p_0(\cdot; x, y, \theta) = p_{\mathrm{dis}}(\cdot; x).$$

Proof. Since $p_{\mathrm{dis}}(\cdot; x)$ does not depend on θ,

$$\begin{aligned} \nabla_\theta \log M_\lambda(x, y; \theta) &= \frac{\lambda}{M_\lambda(x, y; \theta)} \int_{\mathcal{X}} p_{\mathrm{dis}}(u; x)\, \nabla_\theta f(u, y; \theta)\, \exp\big(\lambda f(u, y; \theta)\big)\, \mathrm{d}u \\ &= \lambda\, \mathbb{E}_{X \sim p_\lambda(\cdot; x, y, \theta)} \left[\nabla_\theta f(X, y; \theta) \right]. \end{aligned}$$

Dividing by λ gives the claim for $\lambda > 0$. □

Corollary 9 (What PAT optimizes in practice)**.** *Combining Theorem 7 and Proposition 8, as $B \to \infty$,*

$$\hat{g}_B^{\mathrm{emp}}(\theta) \xrightarrow{\text{a.s.}} \nabla_{\theta'} \mathbb{E}_{(x,y) \sim \tilde{p}(\cdot, \cdot; \theta)} \left[R_\lambda(x, y; \theta') \right]\big|_{\theta' = \theta},$$

where the gradient passes inside the outer expectation because $\tilde{p}(\cdot, \cdot; \theta)$ does not depend on θ'.

Proposition 10 (The smoothed objective preserves the lower bound on probabilistic robustness)**.** *Fix a clean pair $(x_{\mathrm{ori}}, y_{\mathrm{ori}})$, and let s, E_{adv}, and $\gamma^* > 0$ be as in Theorem 2. Then:*

(i) Monotone family. The map $\lambda \mapsto R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta)$ is continuous on $[0, \infty)$, and, for $\lambda > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) = \frac{1}{\lambda^2} \mathrm{KL}(p_\lambda(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta) \,\|\, p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})) \geq 0.$$

Hence, for every $\lambda \in [0, c_2]$,

$$R_0(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \leq R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \leq R_{c_2}(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta),$$

with the exact integral representation

$$R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) - R_0(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) = \int_0^\lambda \frac{1}{t^2} \mathrm{KL}(p_t(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta) \,\|\, p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}))\, \mathrm{d}t.$$

(ii) A family of lower bounds. For every $\beta \in [0, c_2]$, i.e., every $\lambda = c_2 - \beta \in [0, c_2]$,

$$\mathrm{PR}(p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}), s) \geq 1 - \frac{c_2 R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta)}{\gamma^*}.$$

At $\lambda = 0$ (i.e., $\beta = c_2$), this coincides with the bound of Theorem 2, since

$$\begin{aligned} \log Z_{\mathrm{vic}} - H(p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})) - \mathrm{KL}(p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}) \,\|\, p_{\mathrm{vic}}(\cdot; y_{\mathrm{ori}}, \theta)) \\ = \mathbb{E}_{X \sim p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})} \left[c_2 f(X, y_{\mathrm{ori}}; \theta) \right] \\ = c_2 R_0(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta). \end{aligned}$$

For $\lambda > 0$, the bound is looser by exactly

$$\frac{c_2}{\gamma^*} \left(R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) - R_0(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \right),$$

which is quantified by the integral representation in (i)*.*

Proof. (i) The boundedness of $f(\cdot, y_{\mathrm{ori}}; \theta)$ justifies differentiation under the integral:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \log M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) = \mathbb{E}_{X \sim p_\lambda(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta)} \left[f(X, y_{\mathrm{ori}}; \theta) \right].$$

Moreover,

$$\begin{aligned} \mathrm{KL}(p_\lambda(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta) \,\|\, p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})) &= \mathbb{E}_{X \sim p_\lambda(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta)} \left[\lambda f(X, y_{\mathrm{ori}}; \theta) - \log M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \right] \\ &= \lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} \log M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) - \log M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \geq 0. \end{aligned}$$

The quotient rule therefore gives

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\log M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta)}{\lambda}\right) = \frac{1}{\lambda^2}\,\mathrm{KL}(p_\lambda(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta) \,\|\, p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}))\,.$$

Continuity at $\lambda = 0$ follows from the expansion

$$\log M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) = \lambda\,\mathbb{E}_{X\sim p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})}\left[f(X, y_{\mathrm{ori}}; \theta)\right] + O(\lambda^2).$$

The integral representation follows from the fundamental theorem of calculus, with the integrand integrable at 0 since

$$\frac{1}{t^2}\,\mathrm{KL}(p_t(\cdot; x_{\mathrm{ori}}, y_{\mathrm{ori}}, \theta) \,\|\, p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})) \xrightarrow[t\downarrow 0]{} \frac{1}{2}\,\mathrm{Var}_{X\sim p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})}\left[f(X, y_{\mathrm{ori}}; \theta)\right].$$

(ii) Jensen's inequality gives

$$M_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \geq \exp\big(\lambda\,\mathbb{E}_{X\sim p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})}\left[f(X, y_{\mathrm{ori}}; \theta)\right]\big)\,,$$

and hence

$$R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta) \geq R_0(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta).$$

From the proof of Theorem 2,

$$\mathbb{E}_{X\sim p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}})}\left[c_2 f(X, y_{\mathrm{ori}}; \theta)\right] \geq \gamma^*\, I[p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}), s]\,.$$

Therefore,

$$I[p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}), s] \leq \frac{c_2 R_0(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta)}{\gamma^*} \leq \frac{c_2 R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta)}{\gamma^*},$$

and

$$\mathrm{PR}(p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}), s) = 1 - I[p_{\mathrm{dis}}(\cdot; x_{\mathrm{ori}}), s] \geq 1 - \frac{c_2 R_\lambda(x_{\mathrm{ori}}, y_{\mathrm{ori}}; \theta)}{\gamma^*}.$$

□

Corollary 11 (Aggregate control of average probabilistic robustness)**.** *For each $y \in \mathcal{Y}$, define the per-label quantities*

$$\begin{aligned} s_y(u) &:= \max_{y'\neq y}\big(z_{y'}(u) - z_y(u)\big), \\ E_{\mathrm{adv}}(y) &:= \{u \in \mathcal{X} : s_y(u) \geq 0\}\,, \\ \gamma^*(y; \theta) &:= \inf_{u\in E_{\mathrm{adv}}(y)} c_2 f(u, y; \theta), \end{aligned}$$

with the convention $\inf \varnothing = +\infty$*. Assume $\gamma^*(y; \theta) > 0$ for every $y \in \mathcal{Y}$. Since $\mathcal{Y}$ is finite,*

$$\gamma^*_{\min}(\theta) := \min_{y\in\mathcal{Y}} \gamma^*(y; \theta) > 0.$$

Define the average probabilistic robustness

$$\overline{\mathrm{PR}}(\theta) := \mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[\mathrm{PR}(p_{\mathrm{dis}}(\cdot; x), s_y)\right].$$

Then, for every $\beta \in [0, c_2]$,

$$\overline{\mathrm{PR}}(\theta) \geq 1 - \frac{c_2 \exp\big(\beta F_{\max}(\theta)\big)}{\gamma^*_{\min}(\theta)}\mathbb{E}_{(x,y)\sim\tilde{p}(\cdot,\cdot;\theta)}\left[R_\lambda(x, y; \theta)\right].$$

Proof. Apply Proposition 10(ii) to each clean pair. Using

$$\gamma^*(y; \theta) \geq \gamma^*_{\min}(\theta)$$

and $R_\lambda(x, y; \theta) \geq 0$ (since $M_\lambda(x, y; \theta) \geq 1$), and taking expectations over $(x, y) \sim p_{\mathrm{data}}$, gives

$$\overline{\mathrm{PR}}(\theta) \geq 1 - \frac{c_2}{\gamma^*_{\min}(\theta)}\mathbb{E}_{(x,y)\sim p_{\mathrm{data}}}\left[R_\lambda(x, y; \theta)\right].$$

By Lemma 6(iii),

$$\exp\big(-\beta F_{\max}(\theta)\big) \le r(x,y;\theta) \le 1.$$

Hence,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\tilde{p}(\cdot,\cdot;\theta)}\left[R_\lambda(x,y;\theta)\right] &= \frac{\mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[r(x,y;\theta)\, R_\lambda(x,y;\theta)\right]}{\mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[r(x,y;\theta)\right]} \\ &\ge \exp\big(-\beta F_{\max}(\theta)\big)\, \mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[R_\lambda(x,y;\theta)\right].\end{aligned}$$

Equivalently,

$$\mathbb{E}_{(x,y)\sim p_{\text{data}}}\left[R_\lambda(x,y;\theta)\right] \le \exp\big(\beta F_{\max}(\theta)\big)\, \mathbb{E}_{(x,y)\sim\tilde{p}(\cdot,\cdot;\theta)}\left[R_\lambda(x,y;\theta)\right].$$

Combining the two bounds completes the proof. □

D Further Discussion on the Global Vulnerability Capacity Z_{vic}

While our theoretical results establish $KL(p_{dis}||p_{vic}) - \log Z_{vic}$ as a surrogate for probabilistic robustness, a deeper understanding of the term $-\log Z_{vic}$ reveals a fundamental trade-off in robust optimization. From the perspective of distribution expansion, the normalization constant $Z_{vic} = \int_{\mathcal{X}} \exp(c_2 f(u, y_{ori}; \theta))du$ represents the total "volume" of the model's vulnerability across the entire input hypercube $\mathcal{X} = [0,1]^d$.

Geometrically, the probability density p_{vic} is a normalized representation of the classification loss. Because the domain $\mathcal{X}$ is compact and the integral of p_{vic} is constrained to 1, any attempt to minimize the density (i.e., increase robustness) in the local neighborhood of x_{ori} via $KL(p_{dis}||p_{vic})$ must be compensated by an increase in density elsewhere in the domain. This is analogous to the "water level" in a closed pool: pushing the water away from one area (the distance distribution) necessarily causes the overall water level (Z_{vic}) to rise if the container's volume is fixed.

This perspective implies that $-\log Z_{vic}$ is not merely a mathematical residue but a global regularizer that penalizes the "compression" of vulnerabilities. If the optimization focuses solely on the KL term, the model might achieve high local robustness by creating extremely sharp decision boundaries just outside the p_{dis} support, leading to a surge in Z_{vic} and potential global instability. Therefore, PAT's inherent structure forces a principled balance: it requires the model to not only push vulnerabilities away from the clean data but also to suppress the total growth of the global vulnerability volume. This justifies the use of the smoothing factor β in Equation equation 2 as a mechanism to manage this pressure and prevent catastrophic weight collapse during the mass redistribution process.

E Details of the Experiments

In this section, we provide the comprehensive experimental settings, hyperparameters, and evaluation protocols required to reproduce the results presented in Section 6.

E.1 Basic Training Setup

For all experiments conducted on the CIFAR-10 and CIFAR-100 datasets, we apply a standard data augmentation pipeline to the training set: a 4-pixel reflection padding (`Pad(4, reflect)`), followed by a random crop to 32×32 pixels, and a random horizontal flip with a probability of $p = 0.5$.

We train all models for a total of $E_{total} = 100$ epochs using a batch size of $B = 256$. The network parameters are optimized using Stochastic Gradient Descent (SGD) with a Nesterov momentum of 0.9 and a weight decay of 5×10^{-4}. The initial learning rate is set to $\alpha = 0.01$ and is scheduled using a MultiStepLR decay strategy, which multiplies the learning rate by a decay factor of $\gamma = 0.1$ at milestones of epoch 75 and epoch 90.

E.2 Hyperparameters of Probabilistic Adversarial Training

For our proposed Probabilistic Adversarial Training (PAT) detailed in Algorithm 2, the generation of the adversarial distribution via Langevin Dynamics requires careful calibration. The precise hyperparameter configurations used across our main experiments are summarized below:

- **Langevin Dynamics Steps (T):** 100 iterations.
- **Langevin Step Size (η):** 0.3.
- **Noise Scale (σ):** 0.001, which ensures sufficient stochastic exploration within the local neighborhood.
- **Gradient Clipping Threshold (ρ):** 1.0, applied to stabilize the energy landscape exploration.
- **Energy Weights:** We set the distance distribution weight to $c_1 = 0.3$ and the victim distribution (loss) weight to $c_2 = 0.42$.
- **Smoothing Factor (β):** As extensively discussed in Section 5.4 and Appendix D, we instantiate the smoothing factor as $\beta = 0.001$ to prevent global weight collapse during the Self-Normalized Importance Sampling (SNIS).

E.3 Baselines and Evaluation Protocol

Standard Robustness Evaluation: To assess standard worst-case robustness, we evaluate the models using the well-established Projected Gradient Descent (PGD) and Carlini-Wagner (CW) attacks. Both attacks are executed for 20 iterative steps (PGD-20 and CW-20). $\epsilon = 8/255, \alpha = 2/255$.

Probabilistic Robustness Evaluation: For the Probabilistic Robustness (PR) reported in Table 1, the test-time distance distributions are configured with escalating perturbation scales $\epsilon \in \{0.1, 0.12, 0.15, 0.2\}$. We sample $N = 100$ perturbations per image to empirically estimate the PR objective.

Baseline Training Settings: For standard adversarial training baselines such as PGD-AT and TRADES, we strictly follow their original implementations. The baselines generate training adversaries using a 10-step attack with a maximum perturbation bounded by $L_\infty \leq 8/255$.

E.4 Compute Resources

All experiments, including baseline training and PAT, were implemented in PyTorch. The models were trained and evaluated on a compute cluster equipped with NVIDIA RTX 5090 GPUs. A single complete training run for PAT on CIFAR-10 takes approximately 7 minutes per epoch on a single GPU.

F Limitations

Probabilistic adversarial attacks generate adversarial examples from noise and therefore typically require more generation steps than conventional adversarial attacks. This leads to a higher computational cost compared with standard attack methods.